ANNUAL REVIEWS

Annual Review of Marine Science

Historical Estimates of Surface Marine Temperatures

Elizabeth C. Kent[1] and John J. Kennedy[2]

[1]National Oceanography Centre, Southampton SO14 3ZH, United Kingdom; email: eck@noc.ac.uk

[2]Met Office Hadley Centre, Exeter EX1 3PB, United Kingdom

Annu. Rev. Mar. Sci. 2021. 13:283–311

The *Annual Review of Marine Science* is online at marine.annualreviews.org

https://doi.org/10.1146/annurev-marine-042120-111807

Keywords

sea-surface temperature, marine air temperature, climate records, observations, uncertainty

Abstract

Surface temperature documents our changing climate, and the marine record represents one of the longest widely distributed, observation-based estimates. Measurements of near-surface marine air temperature and sea-surface temperature have been recorded on platforms ranging from sailing ships to autonomous drifting buoys. The raw observations show an imprint of differing measurement methods and are sparse in certain periods and regions. This review describes how the real signal of global climate change can be determined from these sparse and noisy observations, including the quantification of measurement method–dependent biases and the reduction of spurious signals. Recent progress has come from analysis of the observations at increasing levels of granularity and from accounting for artifacts in the data that depend on platform types, measurement methods, and environmental conditions. Cutting across these effects are others caused by how the data were recorded, transcribed, and archived. These insights will be integrated into the next generation of global products quantified with validated estimates of uncertainty and the dependencies of its correlation structure. Further analysis of these records using improved data, metadata, and methods will certainly uncover more idiosyncrasies and new ways to improve the record.

1. INTRODUCTION

Uncertainty: a distribution or other measure of the dispersion around the true value

Gridded: produced on a regular spatiotemporal grid (here typically at monthly and 1–5° resolutions)

The Paris Agreement aims to avoid dangerous climate change through the implementation of policies that will limit the change of global mean surface temperature (GMST) to less than 2°C above the temperature of the preindustrial era. The increase in GMST in recent decades has been documented by authoritative assessments such as those of the Intergovernmental Panel on Climate Change (IPCC). Those assessments have, with increasing confidence, attributed the changes in temperature to the increase of greenhouse gases in the atmosphere, resulting primarily from the accumulation of CO_2 emissions from industry, transport, and changes in land use. The policies aim to reduce emissions of greenhouse gases and use GMST to measure progress. The sidebar titled What Will a Change in the Estimated Global Mean Surface Temperature Mean for the Paris Limits? discusses the implications of a change in the estimate of GMST on the assessments of our progress in keeping within the Paris limits.

The importance of high-quality observational data products to inform, support, and constrain model-based assessments of climate change is well established. Currently, it is difficult to correctly assess the uncertainties in both climate models and observation-based global gridded surface products that are based on temperature measurements dating from 1850 to the present (Hegerl et al. 2019). This difficulty means that when the observation-based products disagree with the temperatures predicted by the climate models, it can be difficult to understand precisely what the comparison tells us (Hegerl et al. 2018), which in turn leads to uncertainty in the relative importance of different processes controlling climate variability. The estimated uncertainty in GMST change is fairly small compared with the uncertainty associated with the process of attributing the change to its different components—human-induced factors, such as changing greenhouse gas concentrations, aerosols, and land use (anthropogenic variability); volcanoes and solar variations (natural variability); and the internal variability of the climate system (figure 10.5 in Stocker et al. 2013). Better constraining these contributions requires greater confidence not only in the global mean temperature change but also in the regional temperature patterns over periods long enough to represent long-period internal climate variability (Wang et al. 2017). Long observation-based records of marine surface temperature that accurately resolve regional changes and are characterized with validated estimates of uncertainty (Bulgin et al. 2016) are therefore needed to evaluate climate models and enable a better understanding of climate processes.

The marine observations used to construct these long gridded fields of surface temperature come from disparate sources that have been collated over centuries for a range of commercial,

WHAT WILL A CHANGE IN THE ESTIMATED GLOBAL MEAN SURFACE TEMPERATURE MEAN FOR THE PARIS LIMITS?

The research presented in this article underlines that estimates of surface marine temperature and GMST are uncertain and that further improvements can be expected in the data products representing GMST. These improvements will result in different estimates of global change and its uncertainty. So what do those changes in observation-based surface temperature estimates mean for assessing the state of the climate in the context of the Paris limits? The 1.5°C or 2°C values arise from specific combinations of climate models, their projected impacts, potential mitigation, and observations—combining the whole of climate science into a single number. It is not possible to mix and match these components and be confident that the result will be meaningful. In the future, we will need to reevaluate these limits using the best available GMST data products, projections from the new generation of climate models, and a better understanding of impacts and mitigation.

national, and international applications. The quality and quantity of observations now available are a direct result of the changing motivations for data collection, the technology available for measurement, and the limitations in data storage, transmission, and analysis capability. The importance of making careful measurements of environmental conditions at sea was systematically documented more than 150 years ago by Maury (1854), and even before that, seafarers recognized the importance of observations to improve their safety and efficiency (Franklin 1786).

The main driver for data collection presently is numerical weather prediction, and most of the observations used for climate applications are received through the near-real-time numerical weather prediction systems (e.g., the Global Telecommunication System; Viglione 2020). In situ measurements of sea-surface temperature (SST) are available from ships, moored buoys, and drifting buoys, but measurements of marine air temperature (MAT) rely heavily on ships, and its observational coverage is declining as weather forecasts prioritize observations from satellites, aircraft, and buoys (Kent et al. 2019).

As surface temperature has risen rapidly in recent decades, the assumption that anomalies of SST and MAT (i.e., differences from climatology) show similar variability and trends (Folland et al. 1984, Jones et al. 1988) has been questioned (Cowtan et al. 2015, Richardson et al. 2016). This is important because GMST data products (Morice et al. 2012, Vose et al. 2012, Lenssen et al. 2019) combine anomalies of near-surface temperature over land with anomalies of SST rather than MAT, a choice based on the belief that SST anomalies are more reliable because SST is less variable than MAT at the scales typically considered in these analyses. Furthermore, only nighttime MAT (NMAT) measurements are used in global products because of biases caused by solar ship heating during the day (Jones et al. 2016), reducing the available number of measurements. Tokarska et al. (2019) therefore argued that comparing climate model output with global surface temperature products requires extracting a similar hybrid temperature from the models. Presently, the evidence for this recommendation is based only on analysis of climate models, which typically shows that MAT increases faster than SST (Schurer et al. 2018). It is not possible to unambiguously evaluate the difference between SST and MAT using the current gridded observational products because the anticipated size of the difference (~0.01–0.02°C per decade; Cowtan et al. 2015) is smaller than the estimated uncertainties. Another approach would be to generate a homogeneous global observational record using air temperature over land, ocean, and sea ice. It seems plausible that a useful climate record could be constructed based on MAT: Gridded NMAT (Kent et al. 2013) is presently used to construct bias adjustments for SST because its long-term stability is expected to be better than that of SST (Huang et al. 2017), and a bias adjustment methodology for daytime MAT already exists (Berry et al. 2004).

One obvious way forward is to produce independent estimates of surface marine temperature based on SST and MAT and jointly analyze the records to produce best estimates of each measure. However, the number of reports containing MAT has declined (Berry & Kent 2017), and so the uncertainty in MAT products has increased over the past decade (Cornes et al. 2020). This increased uncertainty will make it difficult to confidently assess whether the differences between SST and MAT in models are also seen in the observations. Another related issue is the relatively poor sampling in high-latitude and polar regions, which are warming faster than lower-latitude regions (Cowtan & Way 2014). Care is required to ensure that the interactions between changing coverage and geographical variations in variability and secular changes are accounted for when comparing models and gridded products as well as when comparing different gridded products, lest changes in sampling be mistaken for changes in the climate itself.

The increase in GMST in recent decades is larger than the estimated observational uncertainty and has been attributed to human activities with high confidence. The likely range for human-induced warming in 2017 was between 0.8°C and 1.2°C relative to preindustrial values

Anomaly: a difference from climatology

Climatology: a representative or reference for a parameter, typically formed by a long period average for a particular grid cell and period (e.g., the representative value for January in a 5° grid cell)

Bias: the mean offset (error) of a group of observations relative to a reference thought to be more accurate

(Masson-Delmotte et al. 2018). However, recent attention has focused on changes relative to a poorly observed preindustrial period (Hawkins et al. 2017) and on small variations in recent trends (Medhaug et al. 2017). Both of these cases place severe demands on the data, in the latter case requiring trend stability on the order of 0.01°C per decade (Hausfather et al. 2017). For example, an update to the methodology used to generate the SST component of one GMST product, which resulted in trend changes within the estimated uncertainty range, was at the time controversial (Karl et al. 2015), highlighting the importance of understanding GMST in the context of the estimates of its uncertainty and the volatile nature of short-term temperature trends.

Error: a difference between the measured and true values, which is unknowable

Grid cell: a spatiotemporal division used for summary statistics or for gridded fields (e.g., 5° monthly)

The IPCC assessment reports are a great motivator for the construction of gridded fields of SST and NMAT and their blending with near-surface air temperature observations over land to produce global surface temperature estimates (see the **Supplemental Appendix**). The IPCC has made five assessments to date, with the next report due in 2021. Improvements over time to the data products used in IPCC assessments have been a result of an increase in the number of observations available (through both the development of observing networks and the recovery and digitization of old ship records), the application of more sophisticated statistical methods, and a deeper understanding of errors and variation in the data. It has become possible to more appropriately attribute the observed temperature variability in a grid cell (illustrated in **Figure 1**). Improvements in understanding of how some of the variability within a grid cell arises from real effects, such as spatial gradients across the grid cell or diurnal variability (Morak-Bozzo et al. 2016), or from biases that can be adjusted for have allowed the construction of more accurate data products with more robust uncertainty estimates.

Supplemental Material >

This article reviews the data and science behind the construction of historical marine surface temperature data products, which we have defined as those a century or more in length. This scope means that the majority of the record comes from ship observations, which have been supplemented with other types of data in recent decades. Satellite data records are considered only when they contribute to such a long historical data record or where they inform the understanding of the accuracy of these longer records. We look at some of the continuing challenges in data set construction and how they have been addressed, and also look to the future.

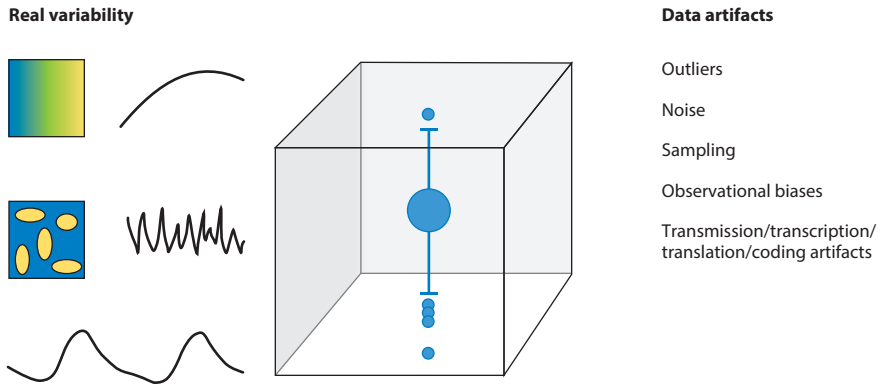


Figure 1

Partitioning of variability within a grid cell as a combination of the real effects of sub-grid-cell spatial and temporal variability (*left*), some of which can be quantified, and artifacts in the data (*right*). Artifacts can include outliers and mistakes, which can be reduced by quality control, and observational biases, which can be reduced by adjustment.

The review is structured to first consider the observing systems and data archives (Section 2) and then the factors that affect the quality of the temperature observations (Section 3). Methods to estimate biases and uncertainty in the observations are discussed next (Section 4), followed by the construction of gridded fields (Section 5). Finally, we anticipate future improvements (Section 6).

2. THE EVOLVING MARINE OBSERVING SYSTEM AND RESULTING DATA ARCHIVE

2.1. Marine Data and Archives

The observations that are available to us come from a variety of different sources, and the number of observations; their coverage, quality, and completeness; and the availability of observational metadata or other documentation can be different for each source.

2.1.1. The early observing system. The first observations were made on specialist voyages of exploration (e.g., Walker 2006) and are of limited use for the estimation of large-scale temperature fields. The first example of a substantial collection of observations is that from the English East India Company (Woodruff et al. 2005, Freeman et al. 2017), which made observations of air temperature in the late eighteenth and early nineteenth centuries in an effort to improve their business efficiency by producing charts to aid navigation (Brohan et al. 2012). The early charts and sailing directions focused mainly on winds and currents (Folger 1787, Bowditch 1802, Rennell 1832, Maury 1854), but SST provided important additional information (Strickland 1802, Franklin 1786). Most of the observations we have from this early period are from recent data rescue (Section 2.1.5).

2.1.2. Maury: the start of the international observing system. The international marine observing system was established in the 1850s (Maury 1854) to standardize and coordinate marine observing and share the observations collected on merchant and naval ships. The potential for observations to contain significant errors was already recognized, and there were detailed discussions of what measurements should be made, how they should be made, and the documentation required. As Maury (1854, p. 52) wrote,

> An incorrect observation is not only useless of itself, but when it passes undetected among others that are correct, it becomes worse than useless; nay, it is mischievous there, for it vitiates results that are accurate, places before us wrong premises, and thus renders the good of no value.

2.1.3. National archives. Many collations of marine observations have been made over the years. They were used originally to construct navigational charts and instructions and later for scientific applications (e.g., Budyko et al. 1962, Bunker 1976). There were several national collections, including those of the United States, the United Kingdom, Germany, the Netherlands, Japan, and Norway, and these collections were exchanged as nations amassed global holdings. The duplication of reports that this merging of archives introduced remains an issue today (Section 2.2).

2.1.4. The World Meteorological Organization and Joint Technical Commission for Oceanography and Marine Meteorology. The collation and international exchange of ship observations was facilitated by the World Meteorological Organization (WMO 1968), which assigned responsibility for collating data and producing climatological charts to eight nations, each responsible for a particular ocean region or regions. This distributed system was later streamlined

by the World Meteorological Organization and Intergovernmental Oceanographic Commission's Joint Technical Commission for Oceanography and Marine Meteorology to provide a marine climate data stream through two Global Collecting Centres operated by Germany and the United Kingdom. The Global Collecting Centres provide more complete observations and metadata with quality control flags applied, but only a subset of observations are submitted to this climate data stream.

Duplicate: any data record deriving from the same original report that appears multiple times in a data bank (although it may not be an exact duplicate)

The data collected in support of numerical weather prediction are exchanged internationally in almost real time via the Global Telecommunication System (WMO 2018). Because numerical weather prediction needs rapid access to observations, data are shared across an integrated network. Each observation remains available for a specified length of time based on its utility for numerical weather prediction. Data are pulled from the Global Telecommunication System by numerical weather prediction centers; the exact subset of data retrieved will depend on how and when the interrogation occurs, and there are typically many duplicate observations.

2.1.5. Data rescue. Data rescue (Wilkinson et al. 2011, Allan et al. 2016) is the digitization of information into a machine-readable format. This digitization often involves the transcription of data from logbooks but can also include conversion between different media (e.g., paper, punch cards, microfiche, tapes, or disks) and sometimes recovery from archaic formats or data systems. Each conversion needs to be done with care, retaining as much information as possible, including metadata and environmental parameters beyond those of immediate interest (Kent et al. 2019). Transcribing ship logbooks and other historical information is time-consuming and has so far proved resistant to automatic techniques to convert paper records to digital form. Citizen science, where volunteers transcribe the observations, can be very effective (Brohan et al. 2009) but requires large volumes of records in similar formats.

2.1.6. Other observing networks. In addition to ship observations, other types of marine platforms provide temperature measurements, as described by Kent et al. (2019). Research vessels provide measurements of SST and MAT made underway, which can be of high quality (Smith et al. 2018), and also near-surface measurements from seawater temperature profiles (Atkinson et al. 2014, Boyer et al. 2016). Since about 2000, ship-derived profiles have been supplemented by data from autonomous profiling floats in the Argo program (Argo 2020). Arrays of moored buoys provide time series measurements of SST and MAT at limited coastal or tropical locations (Centurioni et al. 2019) but may become more widespread (Davis et al. 2018). Particularly valuable for SST coverage are drifting buoys, although their typically high data quality (Kennedy 2014) can deteriorate over time (e.g., Atkinson et al. 2013).

Other sources of surface temperature data are fixed platforms such as oil rigs, coastal stations, tide gauges, ice buoys, and light vessels (Freeman et al. 2017). **Figure 2** shows the coverage of SST and MAT as counts of 5° monthly grid cells sampled for each platform type and the total. Ships dominated the observing system until the 1970s, when observations from the other platform types became important, especially for SST. Overall coverage of ocean grid cells by ships has declined in recent years but has been compensated for SST by increased coverage from drifting and moored buoys. MAT sampling has declined since the 1980s, with the other platform types contributing sampling only in a small number of additional grid cells.

2.2. The Archived Record of Marine Surface Temperature

The observations available for use in climate products are a subset of those that have been recorded. Many potentially valuable observations have been lost entirely—perhaps

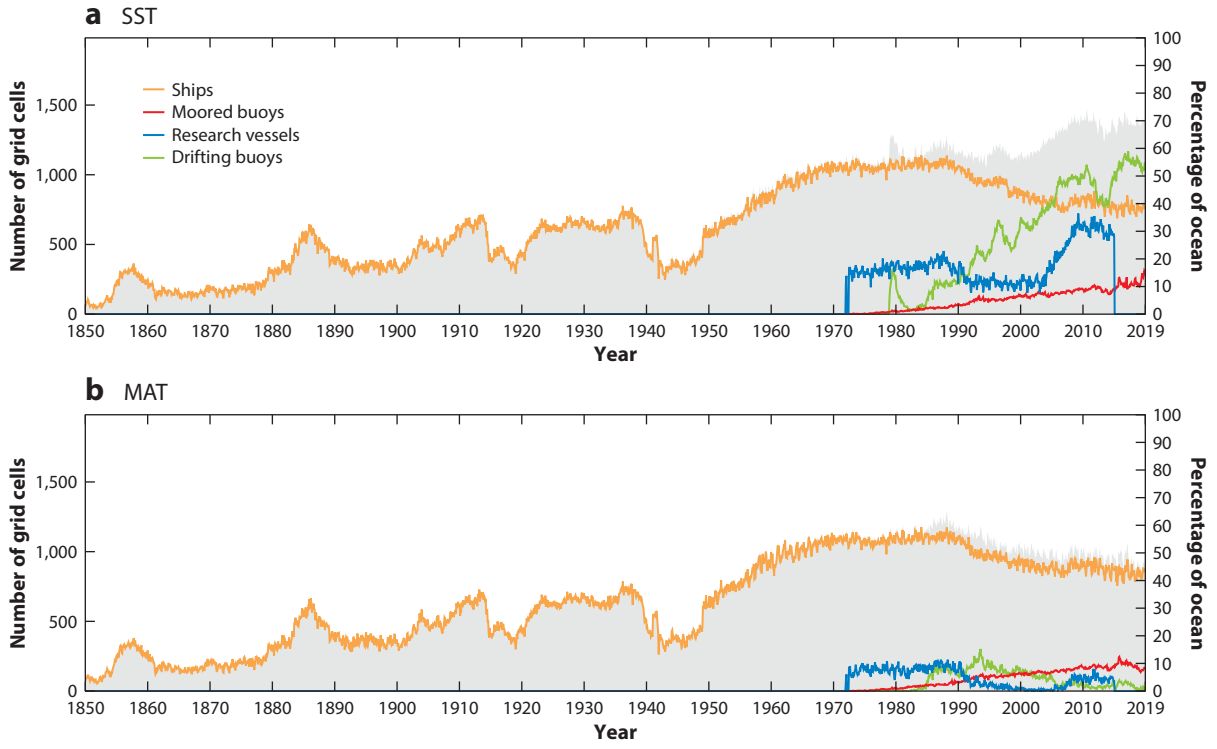


Figure 2

Number of 5° monthly grid cells with at least 10 observations in ICOADS release 3.0 (Freeman et al. 2017) for (*a*) SST and (*b*) MAT, over the period 1850–2019. Gray shading shows the number (*left axis*) and percentage (*right axis*) of ocean grid cells sampled for all observation types. The lines show grid cells sampled by ships (*orange*), moored buoys (*red*), research vessels (*blue*), and drifting buoys (*green*). Note that research vessel observations are not available from the near-real-time data sources and so do not contribute to ICOADS after 2014. Abbreviations: ICOADS, International Comprehensive Ocean–Atmosphere Data Set; MAT, marine air temperature; SST, sea-surface temperature.

destroyed because they were not recognized as useful, or sometimes lost because of accidents or through deterioration in inadequate storage. Even when we have observations, important information may not be available, either because it was not recorded or because it was lost in subsequent translation, keying, or reformatting. The data available today represent a somewhat haphazard selection of the observations that have been made, and the content of each report may not contain all the information that was originally available.

The US national marine surface archive (Section 2.1.3) was the first collection to be made publicly available and is the most complete record of near-surface marine temperatures. Now called the International Comprehensive Ocean–Atmosphere Data Set (ICOADS) (Freeman et al. 2017), this archive is used in all the marine surface temperature products and has become the focus for international efforts to extend and improve the marine surface data archive (Woodruff et al. 2005). This approach has obvious benefits in terms of data availability and traceability, but ICOADS is now in need of modernization (Freeman et al. 2019), which will be necessary to support progress in the development of surface temperature records (Kent et al. 2017, 2019). In particular, the ICOADS approach to identification of duplicate records among the many contributing data sources (Slutz et al. 1985) has led to fragmentation of data from individual ships and buoys

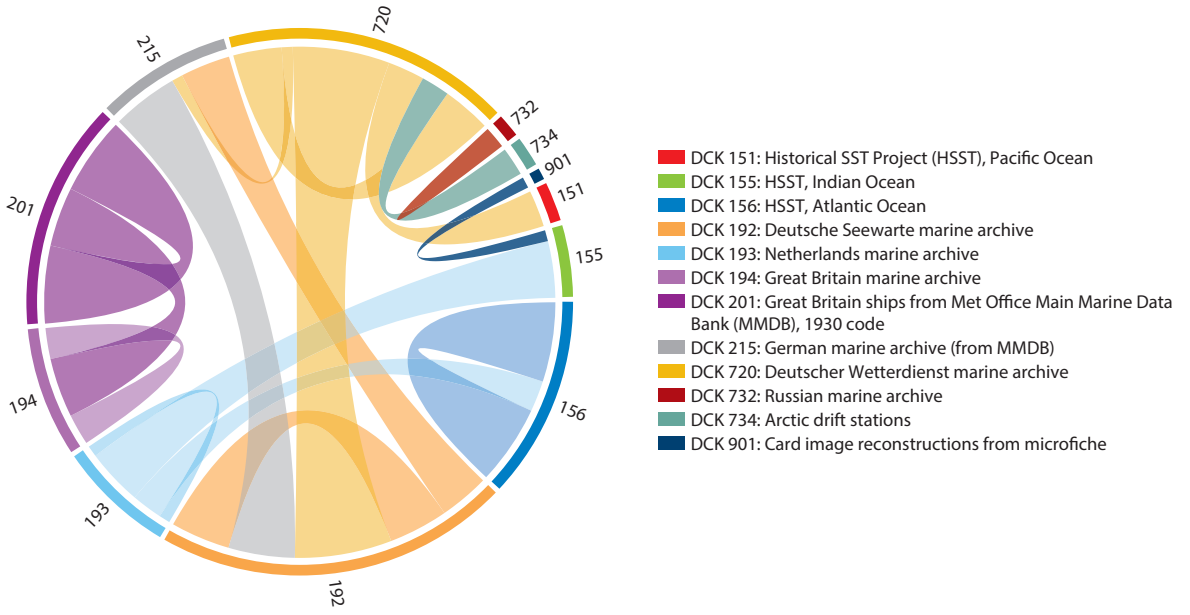


Figure 3

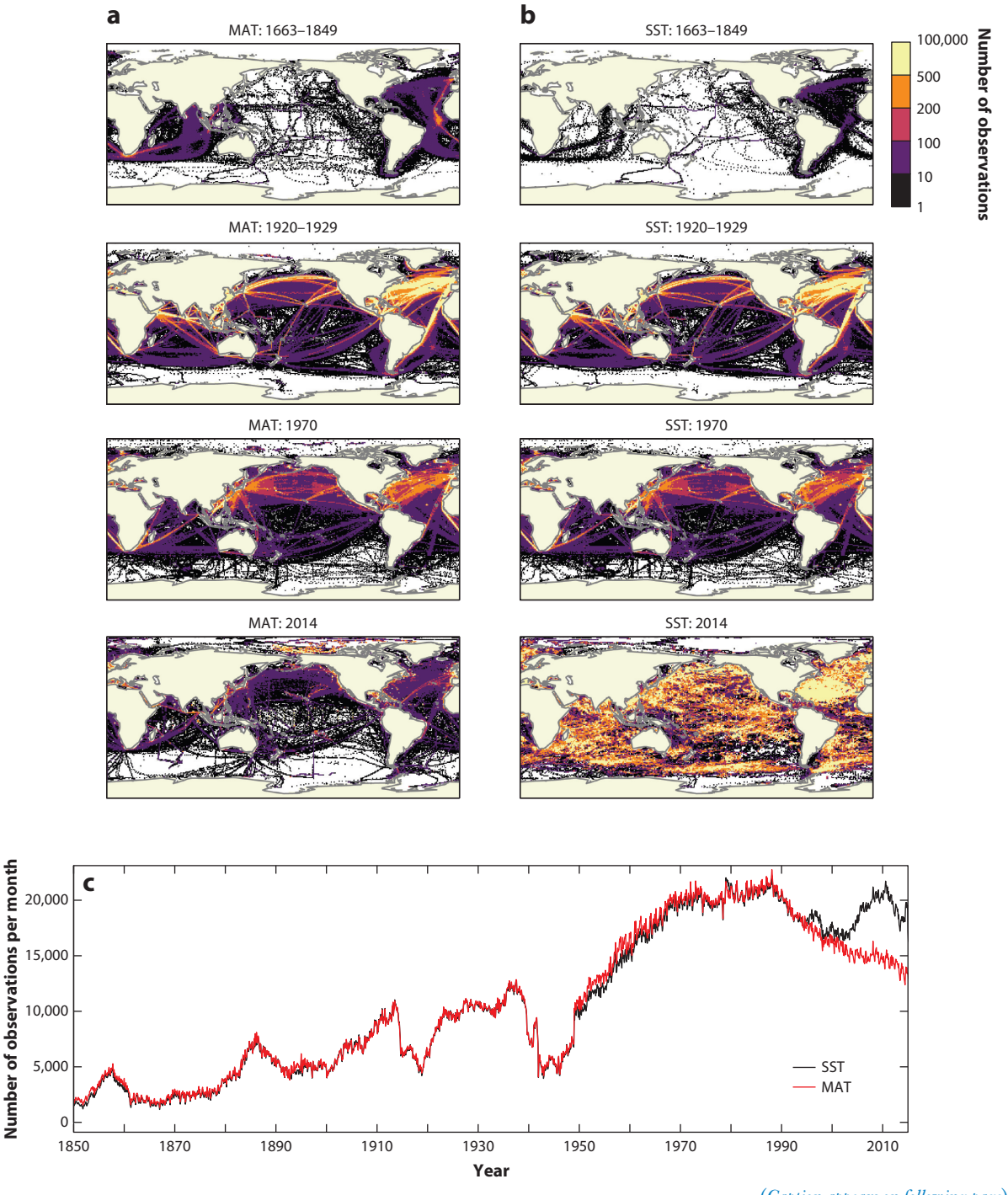
Chord diagram (Gu et al. 2014) showing relationships between report duplicates from 1900. Each segment on the diagram represents a data source (ICOADS deck number) containing duplicates and is linked to the sources containing alternative versions of the duplicate reports. Segment sizes are scaled by the log of the number of duplicates, and links are colored according to the source of the preferred version of the report. Freeman et al. (2017) provide more information on data sources. Abbreviation: ICOADS, International Comprehensive Ocean–Atmosphere Data Set.

between different data sources (Carella et al. 2017a), making it harder to diagnose source- or platform-specific biases (Carella 2017, Chan & Huybers 2019). **Figure 3** illustrates this fragmentation, showing the duplicates identified across different sources.

3. MARINE TEMPERATURE OBSERVATIONS

3.1. Observations of Sea-Surface Temperature

The evolution of preferred shipboard measurement methods for SST has been summarized elsewhere (Folland & Parker 1995; Kent & Taylor 2006; Kent et al. 2007, 2017; Kennedy et al. 2011b). The first measurements predominantly used mercury thermometers to measure the temperature of samples of near-surface water collected using buckets; over time, there was increasing use of measurements of engine room cooling water, and then an increase in measurements from dedicated installations, including sensors fixed to the ship's hull. Prior to 1850, MAT was observed more frequently than SST, but by around 1920 the sampling had become more comparable (**Figure 4**). Sampling and coverage subsequently increased for both variables, with the exception of periods around the two world wars, until the late 1980s. Coverage then declined for MAT, but since the early 1990s, SST measurements from ships have been hugely augmented by measurements from other types of platforms, particularly moored and drifting buoys (Freeman et al. 2017, Davis et al. 2018, Centurioni et al. 2019) (**Figure 2**) and satellites (Minnett et al. 2019). The changing sampling causes a coverage error in estimates of the global mean SST that has been estimated to have a maximum value of approximately 0.2°C (Kennedy 2014).



(*Caption appears on following page*)

Figure 4 (*Figure appears on preceding page*)

Sampling density from ICOADS release 3.0 (Freeman et al. 2017). (*a*,*b*) Numbers of observations of MAT (panel *a*) and SST (panel *b*) per 1° grid cell over the periods shown above each map. Note that aggregations over different numbers of years are shown. (*c*) Number of observations per month over the period 1850–2014. Abbreviations: ICOADS, International Comprehensive Ocean–Atmosphere Data Set; MAT, marine air temperature; SST, sea-surface temperature.

Although the general narrative of changing observing methods for SST is well known, it remains difficult to unambiguously associate measurement methods with archived observations (Section 3.2). Even when this can be done with high confidence, some details are less well known. For bucket-derived measurements, examples include the exact type of bucket used, the measurement protocols, and environmental conditions (Kent et al. 2017). Some details are almost never known, such as the duration of the measurement and whether the bucket and sample came into full equilibrium with the water temperature.

We do have valuable information on observing protocols from the instructions that have been issued to observers. For example, the Maury instructions recommended using a wooden bucket to make the near-surface seawater temperature measurement, placing the bucket in the shade, allowing two to three minutes for the thermometer to equilibrate, and ensuring that the thermometer bulb stayed in the water during the reading. Throughout the record, differences between the observations made by different vessels and nations have been identified (Saur 1963, Kent et al. 1993, Kent & Taylor 2006, Chan & Huybers 2019, Chan et al. 2019), illustrating the importance of identifying and archiving observing instructions and other metadata describing observational practice in a way that can be traced back to the associated observations.

The possible errors in particular types of measurement have been quantified in several comparative studies. Brooks (1928) compared observations on US steamships derived from buckets of different types with engine room intake (ERI) and thermograph observations. He concluded that ERI measurements were to be preferred except in quiet weather (when vertical gradients are expected) and that ERI measurements should be supplemented with careful bucket measurements. The WMO (1957) echoed this conclusion, recommending improvements to ERI installations to include remote reading and resistance thermometers. They also recommended bucket measurements in conditions where shallow warm layers were expected (quiet, sunny weather at lower latitudes) and improved-quality ERI at latitudes above 45°.

Despite these recommendations, an evaluation by Saur (1963) of ERI SSTs from US military vessels showed mean biases and scatter across the fleet that varied by ship (from −0.5°F to 3.0°F) and by trip for each single ship. Saur recommended evaluating biases in ERI SSTs for each individual ship.

Probably the most extensive collection of colocated SST observations made using different methods was gathered during 1968–1970 and analyzed by James & Fox (1972). They concluded that biases in ERI SSTs could be reduced by using better intake thermometers located close to the water inlet and that observers should take great care when measuring bucket SSTs under strong winds, extreme air–sea temperature differences, and heavy precipitation, suggesting that the recommendations of the WMO (1957) were not widely adopted. Although James & Fox (1972) tabulated differences between colocated ERI and bucket SSTs for a wide range of potential covariates, including environmental conditions, ship particulars, and bucket types, it is hard to draw further conclusions due to the large number of possible confounding factors.

Different bucket types were evaluated by measuring the temperature change of the water sample over time in controlled or known conditions. Ashford (1948) measured the heat exchange characteristics of several buckets in a wind tunnel using a single airflow speed, finding that a standard canvas bucket showed a rate of change of temperature approximately three times greater

than newer, better-insulated buckets. Roll (1951) assessed the bucket issued to German observers in a wind tunnel at a range of airflow speeds. Carella et al. (2017b) found it hard to reconcile their results for the German bucket and concluded that different flow characteristics in the two wind tunnels may mean that it is not possible to uniquely define heat exchange coefficients for buckets in this way. The functional relationship between the temperature change of the bucket water sample and external forcing is typically represented by the difference between the water temperature and the wet-bulb temperature of the ambient air (Ashford 1948, Farmer et al. 1989, Folland & Parker 1995) and is expected to also be a function of the airflow around the bucket (Roll 1951). The effects of solar radiation cannot be determined in such wind tunnel and laboratory experiments.

Matthews (2013) made an extensive comparison of coincident measurements from three different bucket types (wooden, canvas, and insulated) on a research cruise in the tropical Pacific. Unlike other studies, he concluded that there was little difference between the bucket types, but total hauling and measurement times were relatively short (less than one minute), and no effort was made to shade the buckets from the sun.

Kent et al. (1993) showed that microbiases for then-recent ERI SSTs were typically larger than those for bucket SSTs. They also found that there were strong variations of both microbiases and random measurement uncertainty among the fleets of different nations.

Kennedy (2014) tabulated the results of studies where uncertainty was estimated separately for random and microbias components. In general terms, the contribution of each type of error to the total uncertainty is comparable, with most studies finding that the total uncertainty for ship observations was in the range 1.0–1.5°C. Similar estimates have not been made for observations before about 1970, but they are not thought to be markedly different in earlier periods.

These studies, and others reviewed by Folland & Parker (1995), Kent & Taylor (2006), Kennedy et al. (2011b), Matthews (2013), Kennedy (2014), and Kent et al. (2017), make it clear that a one-size-fits-all approach to SST bias adjustment can be successful only at the broadest spatial and temporal scales. Bias estimation and adjustment must be at the appropriate level of granularity—by ship, by nation, or by method—and account for the prevailing conditions and measurement protocols (Section 4.4). This requires extensive platform and observational metadata and ancillary environmental information (Kent et al. 2019), but the availability of this information is patchy (Section 2.2), making such detailed bias adjustments challenging.

Microbias: bias for a particular ship or other type of platform

Random measurement uncertainty: the distribution of residual uncorrelated errors after any adjustment for bias

3.2. Attributing Methods for Measuring Sea-Surface Temperature

The different error characteristics of different measurement types make the identification of measurement methods in the archive of critical importance (Kennedy 2014, Kent et al. 2017). Bias adjustments initially focused on reconciling the differences in pre–World War II biases shown by Wright (1986), in particular those arising from the use of canvas buckets. The prevailing view was that it was not necessary to explicitly consider biases in the postwar period, and that it would be sufficient to adjust early data to be consistent with the mix of observations in the period used for the climatology (e.g., Folland & Parker 1995). However, the availability of large volumes of more accurate data from drifters in the past decade has made this approach untenable: The referencing of adjustments to a period containing greater uncertainty than the modern period gave the perverse impression that modern SST was more uncertain than SST in the recent past (Kent et al. 2017, Kennedy et al. 2019). SST gridded products now apply adjustments using the most accurate data as a reference (Hirahara et al. 2014, Huang et al. 2017, Kennedy et al. 2019).

Uncertainty remains in the attribution of measurement methods to SST observations. Measurement method flags are often missing, and when present they can be contradictory (Kent et al. 2010, Kennedy et al. 2011b). In addition, documented measurement methods are being

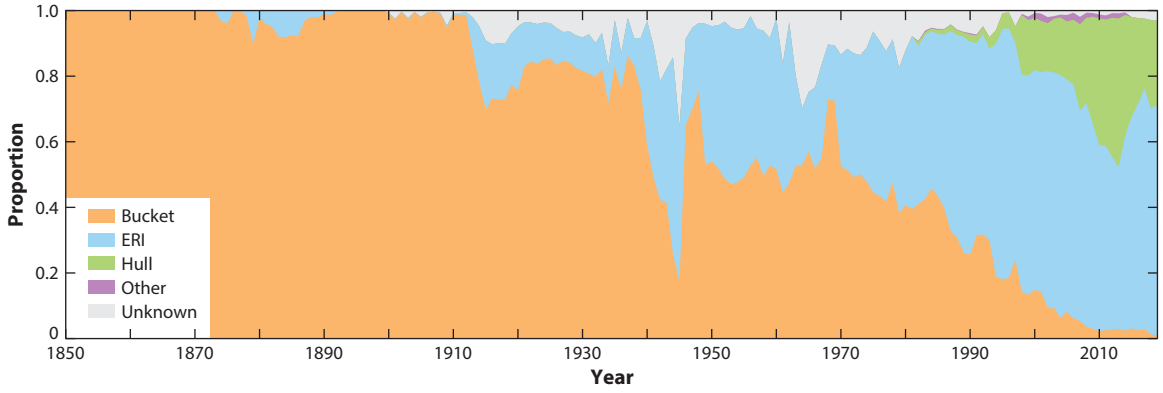


Figure 5

Annual SST measurement methods as a proportion of ICOADS release 3.0 ship observations. The proportions were estimated using available metadata from (in order of preference) ICOADS, WMO Publication No. 47 (Kent et al. 2007), the prevalent method for the recruiting country of the ship, and the data characteristics, following Carella et al. (2018). Abbreviations: ERI, engine room intake; ICOADS, International Comprehensive Ocean–Atmosphere Data Set; SST, sea-surface temperature.

increasingly tested or supplemented by techniques that estimate observing methods based on the characteristics of the observations (Hirahara et al. 2014, Carella et al. 2018, Chan & Huybers 2020) and by modeling the uncertainty due to any ambiguity in the measurement method (Kennedy et al. 2019). **Figure 5** shows estimated changes in measurement methods over time. The difference between global mean SST estimated from buckets and that from ERI is in the range 0–0.5°C, which is reduced to less than 0.2°C after adjustment and typically within the estimated uncertainty (Kennedy et al. 2019). Global mean SST bias estimates are typically within the range ±0.3°C (Huang et al. 2017, Kennedy et al. 2019), with measurements too cold on average before 1940 due to a mix of partly insulated and uninsulated buckets and too warm afterward due to a mix of mainly insulated buckets and ERI (Kent et al. 2017). The change from cold to warm bias near the start of World War II is abrupt (Thompson et al. 2008) and estimated to be between 0.25°C (Huang et al. 2017) and 0.4°C (Kennedy et al. 2019). Recent bias estimates are close to zero in the global mean due to the availability of relatively accurate SST measurements from drifting buoys (**Figure 2**) and a general reduction in ship biases.

3.3. Observations of Marine Air Temperature

Measurement methods for MAT have received less attention than those for SST (Bottomley et al. 1990; Kent et al. 2007, 2013). Ships remain the predominant source of MAT observations and have declined in coverage in recent decades (Berry & Kent 2017) (**Figure 2**).

Prior to approximately 1900, MAT was more frequently observed than SST, but early in the record it was common to report temperatures only at local noon. Because of the known biases in MAT due to daytime heating of the ship and sensor environment, typically only NMAT observations have been used to produce climate records (Houghton et al. 1990), effectively reducing the earliest start date of gridded global data sets based on MAT observations to approximately 1880 or later (Bottomley et al. 1990, Rayner et al. 2003, Kent et al. 2013, Cornes et al. 2020, Junod & Christy 2020).

Chenoweth (1996) reviewed nineteenth-century marine temperature observing practices and concluded, based on an assessment of mean anomalies and variability, that some measurements

of MAT had been misclassified as SST. In this period, some temperature observations were likely made in cabins (Chenoweth 2000), and some thermometers were not shaded. However, after adjustment, the measurements were thought to be of sufficient quality to detect the cooling associated with volcanic eruptions (Chenoweth 2001).

After the standardization of observing practice (Maury 1854), it is more likely that thermometers were either housed in radiation shelters or read in the shade, but several sources of inhomogeneity remained. Temperature observations need to be made in well-exposed locations with sufficient airflow to minimize the effects of local heating. Until the advent of remote-reading electrical thermometers, the sensors had to be conveniently accessible. In practice, sensors were typically housed in radiation shelters fixed to the ship's rails near the bridge, ideally with one on each side. The windward, better-exposed sensor should be read, and this constrains the positioning of sensors, which are installed at varying heights above sea level.

The varying sensor height must be accounted for, requiring an estimate of the vertical profile of temperature near the surface, usually achieved using Monin–Obukhov similarity theory (e.g., Businger et al. 1971). This theory gives estimates of gradients in the near-surface atmospheric boundary layer, dependent on wind speed, air and sea temperatures, and humidity. Junod & Christy (2020) chose instead to use the temperature gradient calculated from the European Centre for Medium-Range Weather Forecasts Interim (ERA-Interim) 2-m and 950-mbar temperatures (Dee et al. 2011). Observing heights either are estimated from the literature or source documentation (e.g., Wallbrink et al. 2009) or are available in metadata (Kent et al. 2007). A reference height above sea level needs to be selected: 10 m is a common choice (Josey et al. 1999, Berry & Kent 2009, Kent et al. 2013, Junod & Christy 2020), but others have been made. Bottomley et al. (1990) and Rayner et al. (2003) used the mean spatially varying observation height during their climatology reference periods (1951–1980 and 1961–1990, respectively), which was approximately 15 m. Cornes et al. (2020) provided NMAT anomalies referenced to each of 2, 10, and 20 m. The height adjustment on average decreases temperatures measured below the reference height and increases those measured above the reference height. The overall effect is to increase the NMAT change over the period from 1880 to the present by approximately 0.2°C, resulting from an estimated change in global average observing height from approximately 6 to 23 m.

The accurate measurement of air temperature requires the thermometer to be shaded from the sun and well exposed to the prevailing airflow (CIMO 2017). Even if the thermometer is screened from the direct effects of solar radiation, the ambient temperature of structures near the sensor can be elevated by solar heating, and if the ventilation of the sensor is inadequate, the measurement will be biased high. The global mean difference between ship measurements of daytime and nighttime MAT is approximately 1°C; the true value is uncertain, but modern research vessels show about half the diurnal air temperature range of the main observing fleet. Because the sensor arrangement on each ship will be different (Kent et al. 1993, Berry & Kent 2005) and the details are almost always unknown, most climate records based on MAT (Rayner et al. 2003, Kent et al. 2013, Junod & Christy 2020) have used anomalies based on measurements between one hour after sunset and one hour after sunrise as the most consistent representation of MAT change (Parker et al. 1995). If measurements made outside of these hours are to be used, they must be adjusted to account for the effects of daytime heating (Berry et al. 2004), which will vary from ship to ship due to the differences in sensor exposure (e.g., Berry & Kent 2005) and ventilation, both of which may also vary with the relative wind direction over the ship (Kent et al. 1993), the type of radiation shelter, and observing practice (CIMO 2017). Under conditions with high solar radiation, Kent et al. (1993) documented temperature biases greater than 3°C for ships with poorly exposed sensors, compared with a typical value of approximately 1°C for ships with better-exposed sensors.

There are some periods and regions where the archived observations (Section 2.2) show unexpected offsets or variations that are thought to be due to nonstandard observing practices (Bottomley et al. 1990). Most obvious is a warm bias during World War II, which is due to abrupt changes in data sources and measurement methods and to a change in timings of observations to prefer sampling during daylight hours (Folland et al. 1984). There is also some evidence for a warm bias in temperatures in the period 1876–1893 in the Mediterranean and North Indian Ocean, which Bottomley et al. (1990) attributed to the tariff regime for the Suez Canal.

4. APPROACHES TO BIAS AND UNCERTAINTY ESTIMATION

4.1. Introduction to Bias Estimation

In the 1960s and 1970s, gridded analyses of marine observations tended to focus on air–sea interaction and estimating heat budgets (e.g., Bunker 1976) and to produce monthly climatologies rather than values for actual months. This changed during the 1980s, when attempts were made to produce long-term SST data sets based on a variety of input data sources (Paltridge & Woodruff 1981, Barnett 1984, Wright 1986). On global scales, all of the data sets showed divergence from near-surface land air temperature, and Wright (1986) showed that there was a divergence of SST and MAT anomalies. These differences were attributed to the effect of the increasing proportion of ERI measurements over time in the source data (Section 3.2). It was clear that adjustments would be needed for the biases documented for subsets of SST data (Section 3.1) before the SST observations could be used in climate studies. Developing bias adjustments is challenging because we lack a consistent multicentury reference standard (Kent et al. 2017), so exploration of the large-scale biases began with comparisons of gridded fields, with estimates expected to be largely independent, such as MAT or air temperature over land.

At this stage, computing limitations meant that it was not possible to analyze the individual observations, so initial bias adjustment methodologies adjusted gridded SST fields constructed from all ship observations, applying time-varying weights to spatially varying adjustment fields to improve consistency with a reference thought to be more stable over time (Section 4.2). Analysis of the resulting products showed that, although these methods were able to reduce biases at the global scale, they performed badly when changes were rapid (e.g., Thompson et al. 2008). The next generation of SST bias adjustments was developed based on single-method subsets (Section 4.3), again maximizing consistency with independent observations but also internal consistency between data subsets. It is now possible to analyze large volumes of data concurrently, opening up the potential for estimating bias adjustments based on individual observations, taking into account the ambient environmental conditions, and producing bias estimates for individual ships (or groups of ships) based on, for example, a particular recruiting country or data source (Carella et al. 2017b, Chan & Huybers 2019, Kennedy et al. 2019). There remain considerable challenges to a comprehensive analysis of temperature biases, and improvements to the data archive are needed to fully exploit these new methods. Alongside improvements to bias estimation, there has been a better quantification of uncertainties, both independent random errors and correlated ones (Section 4.5).

4.2. Homogenization of All-Method Gridded Fields

The first SST bias adjustments (Farmer et al. 1989, Folland & Parker 1995—hereafter FP95) developed physics-based models for the expected heat exchange experienced by SSTs sampled in wooden and canvas buckets. The FP95 model was used to develop a set of four climatological monthly fields of SST bias adjustments (for wooden and canvas buckets and fast and slow ships),

which have been weighted based on the characteristics of the gridded observations to produce time-varying monthly fields of estimated biases for the Met Office Hadley Centre SST data sets (Bottomley et al. 1990; Parker et al. 1995; Rayner et al. 2003, 2006; Kennedy et al. 2011a, 2019) and Centennial In Situ Observation-Based Estimates Sea Surface Temperature 2 (COBE-SST2) (Hirahara et al. 2014). The FP95 model includes most of the effects thought to be important (see, e.g., Kent et al. 2017) and performs reasonably well based on limited evaluation in the field (FP95) and the laboratory (Carella et al. 2017b). The adjustments have also been evaluated regionally (Folland & Salinger 1995, Hanawa et al. 2000) and globally (Folland et al. 2001, Kent et al. 2017).

Kennedy et al. (2019) found that modern buckets exhibited a warm bias that was strongest during summer months and was not anticipated by the FP95 model. Carella et al. (2018) found that bucket measurements had a diurnal cycle that was on average 0.3°C greater than that seen in drifting buoys, with a peak closer to noon local time. These lines of evidence suggest that, for modern buckets, solar heating is a significant source of bias.

A different approach, used in the Extended Reconstructed Sea Surface Temperature (ERSST) data products (Huang et al. 2017 and predecessors), made the assumption that the large-scale biases in NMAT could be adequately removed to enable homogenization based on air–sea temperature difference using the approximation that unexpected changes in air–sea temperature difference could be attributed to biases in SST (Smith & Reynolds 2002). They noted evidence for uncertainty in the adjusted NMAT fields (Bottomley et al. 1990, Christy 2001) but argued that this would be small relative to biases in SST. In each climatological month, the SST biases were assumed to have the same spatial pattern as the air–sea temperature difference, and a time-varying scaling factor was estimated for every month.

Smith & Reynolds (2002) noted that their adjustment fields have spatial patterns similar to those for sensible heat transfer, as expected from their reliance on air–sea temperature differences, whereas those of FP95 include a greater contribution of evaporation, as their adjustment is approximately based on differences between SST and wet-bulb temperature (Ashford 1948, Carella et al. 2017b). This results in a marked difference in the latitude dependence of the adjustment fields (Smith & Reynolds 2002), which leads to substantial regions and periods where the adjustments differ by more than their joint uncertainty range even at times when the differences in the global mean bias adjustments are adequately explained by the uncertainties (Kent et al. 2017). Kennedy et al. (2019) combined FP95- and Smith & Reynolds (2002)–type bucket adjustments using a linear mix of the two.

Cowtan et al. (2018) used coastal air temperatures to assess the homogeneity of gridded SSTs from unadjusted Hadley Centre Sea Surface Temperature 3 (HadSST3) data. Recognizing the many assumptions needed in their analysis, they concluded that the overall bias adjustments applied in HadSST3 and ERSSTv5 were confirmed but raised questions about the adjustments in some periods, especially around World War II and in the early twentieth century.

The broad-scale approach of using fixed maps with globally invariant weights that vary smoothly in time means that, although this type of adjustment is able to capture the biases in the global mean averaged over several years (Folland et al. 2001), it cannot adjust for variations in the bias that are regional (Chan et al. 2019, Davis et al. 2019) or abrupt (Thompson et al. 2008, Cowtan et al. 2015).

4.3. Homogenization of Data Subsets

Reynolds (1988) performed a homogenization of satellite and in situ gridded SSTs. He used the expected difference in spatiotemporal scales of biases in the different data sources to minimize the biases in each source without the use of an explicit physics-based error model.

In the COBE-SST2 data product (Hirahara et al. 2014), the error model for each type of observation was assumed to be known—a constant value for ERI and hull sensors and fields based on FP95 for buckets. Unknown measurement methods were then estimated to ensure that, after the adjustments had been applied, the global mean SST from the observations of unknown method agreed with that from adjusted observations from a known method. Coefficients were estimated annually and smoothed over five years. Further constraints were applied for certain periods, based either on NMAT or on expected measurement method proportions.

For HadSST4, Kennedy et al. (2019) took gridded subsets of SST data and near-surface measurements from oceanographic profiles with their estimated error covariances and assimilated them into a simple statistical model. By assuming that the oceanographic profiles and buoy measurements were unbiased, they estimated temporally and spatially varying biases for a variety of data subsets. The subsets included observations identified as bucket measurements, ERI measurements, and hull sensor measurements using a variety of metadata from ICOADS (Freeman et al. 2017). The method was tested by estimating errors and their uncertainties in individual ship biases and using synthetic data.

4.4. Estimating Biases in Individual Observations and Data Subsets

Kent & Kaplan (2006) used individual observations to estimate the coefficients in a simple physically based model of SST biases. They assumed that the errors in bucket-derived SSTs were dependent on the air–sea temperature difference (Δt) and that ERI errors could be modeled as a constant offset. Analyzing paired nighttime bucket and ERI SST observations, they estimated coefficients for variations of bucket SST error with Δt in the range of 10–20% and ERI offsets that were warm on average by approximately 0.15°C in the period 1975–1989 and cold by approximately 0.13°C in 1990–1994. The analysis aggregated all observations in five-year periods for a limited subset meeting their selection criteria—a necessity because their analysis required pairs of observations from different methods, and data were limited by the availability of ERI SSTs at the start and bucket SSTs at the end of their analysis period (1975–1994). They therefore could not account for ship-to-ship differences or calculate coefficients for individual ships or other data subsets. This task was eventually tackled by Carella (2017) using a Bayesian hierarchical analysis of ship SSTs, with individual ships and observations grouped by country or data source.

A physically based model was used for errors in nighttime bucket SST (Carella et al. 2017b), following Ashford (1948):

$$\Delta T_{\mathrm{b}} = \beta_1 \Delta t_{\mathrm{wbt}|_{\mathrm{loc,cm,ws,wd}}} + \beta_0, \qquad 1.$$

where the subscript b denotes bucket measurement; ΔT is the estimate of the difference between the measured SST and the true SST (approximated by the reference field); Δt_{wbt} is the difference between the climatological SST and wet-bulb temperature (wbt), here conditional on location (loc), climatological month (cm), wind speed (ws), and wind direction (wd), all derived from ERA-Interim (Dee et al. 2011); and β_1 and β_0 are empirically derived coefficients. The analysis required a reference field; results were shown based on satellite data (Merchant et al. 2014) and a gridded data product (Rayner et al. 2003). A spatial model down-weighted observations clustered regionally, which made the analysis less sensitive to differences between the observations and reference that were location dependent, such as near coasts or in areas where the reference was itself biased. The coefficients were fitted ship by ship, giving an effective heat exchange coefficient (β_1) and offset (β_0) for each ship making bucket observations.

One unexpected result of the analysis was an error model for ERI SST. At first, the ERI biases were modeled as a constant offset per ship, but the residuals for some ships showed large regional

and seasonal biases. The ERI error model was therefore extended to include a term (β_1) that varied with climatological SST, which allowed the ERI error to vary with the expected ambient temperature:

$$\Delta T_e = \beta_1 t_{\text{clim}|_{\text{loc,cm}}} + \beta_0, \qquad 2.$$

where the subscript e denotes ERI measurement and t_{clim} is the climatological SST, here conditional on location and climatological month. This model allows for heat exchange in the pipes within the ship between the inlet and where the temperature is measured; the water in the pipe will be warmed if the ship's interior is warmer than outside and vice versa. Not all ships showed this effect, but on average, ERI-derived temperatures in high latitudes or cold seasons are likely to have warm biases relative to low latitudes or summer seasons.

Chan & Huybers (2019) analyzed a subset of ICOADS (Freeman et al. 2017) selected to be bucket only. They used a linear-mixed-effect model to simultaneously analyze paired observations from many different combinations of the ICOADS constituent data sources. It can be hard to interpret the large number of relative offsets such an analysis produces, but offsets for sources representing a large proportion of bucket measurements were shown to be biased relatively cold, and a smaller number were biased relatively warm. The relative offsets were typically less than 0.2°C but reached 1°C in a small number of cases. The method works well to reduce inhomogeneity in the bucket SSTs but needs to be combined with some additional information to robustly diagnose true biases. This is possible in some cases (Chan et al. 2019), but the technique works best when combined with other methods to estimate pervasive biases.

4.5. Estimating Observational Uncertainties

Estimating uncertainties in surface temperature observations is an important step in the production of gridded products (Section 5). Typically, this estimation is achieved using semivariogram analysis or by comparison with a reference field (see review in Kennedy 2014). The uncertainty estimates made in this way have several contributions, including the effects of any biases in the data, mismatch between the observations being compared (e.g., due to inexact collocation or to representativeness), and random noise. Because any bias adjustment will be imperfect (due to approximations or incorrectness of the bias model, and depending on environmental conditions whose precise state is unknown and on uncertain measurement methods and protocols), there is inevitably a residual uncertainty. The errors that arise from the difference between the true and assumed conditions of the measurement will have a variety of different correlation structures. Where practice is peculiar to a particular ship, errors will correlate along the whole course of the ship's voyage. Where undocumented practices are common to a country's fleet, the errors will correlate at that level. Where climatological weather conditions are assumed, errors will correlate at scales typically associated with weather systems. This poses particular difficulties because the spatial and temporal scales of actual temperature variations will be similar.

While more detailed models (see Section 4.4) may not be enough on their own to fully correct a particular data point, they can be used to estimate the magnitude and correlation structures of errors or, where other variables are available (e.g., from reanalyses), to provide an improved correction.

5. GRIDDED TEMPERATURE PRODUCTS

5.1. Unfilled Grids

Many applications require gridded data sets where individual reports have been aggregated onto a regular grid that summarizes temperatures within a set of grid cells. Aggregation needs to deal

with several difficulties. The first is that reports are subject to measurement errors (where the measured value is related to the true value) as well as mistakes (where the reported and true values are unrelated). Quality control is applied to reject mistakes and other low-quality observations. Met Office data sets have used an outlier-resistant averaging technique known as winsorization. ERSST weights ship and buoy data in favor of the more reliable buoy data.

Superobservations: summary statistics (e.g., means or medians plus uncertainties) that are based on groupings of nearby observations and used in gridding as if they were measurements

The second difficulty in aggregation is uneven and imperfect sampling. Ships of different kinds, drifting buoys, and moored buoys will each sample a grid box differently, potentially biasing a simple average of the measurements within a grid cell toward the more densely sampled areas or times. Several approaches can reduce this bias. For example, measurements can be converted to anomalies before aggregation by subtracting a climatological average, which reduces (but does not eliminate) the effect of strong gradients in temperature across a grid cell. Observations might also be aggregated into superobservations before those superobservations are aggregated onto the final grid.

Other methods use optimal interpolation to deal with the observational clumping (e.g., Williams & Berry 2020). Optimal interpolation can also deal with combining observations of different quality based on their assessed uncertainty. Regardless of the method used, there is a residual uncertainty, sometimes referred to as sampling uncertainty.

The simplest gridded data sets provide grid-cell averages only where there are enough data to calculate one. Gridded fields without infilling have the virtue of simplicity: It can be easier to communicate and understand what exactly a grid-cell average and its uncertainty represent. However, even in well-observed periods, this leaves a significant fraction of grid cells with no estimated value. Simple gridded data sets also tend to have unequal variance: Estimates based on a smaller number of observations will have a larger uncertainty, which may be undesirable for many applications. The techniques employed to create gridded data sets of this kind and estimate the grid-cell uncertainties are described in the **Supplemental Appendix**.

Supplemental Material >

5.2. Infilling Missing Grid Cells

For many applications, it is desirable to have globally complete gridded fields. Because of the often poor observational coverage, this is achieved using statistical or dynamical analyses to infer reasonable values in the data voids. Different approaches have been used (see the **Supplemental Appendix**), but the main infilled historical SST data products—ERSST, HadISST, and COBE-SST2—all use pattern-based methods of reconstruction based on empirical orthogonal functions (EOFs). These techniques are powerful, projecting the available observations onto the large-scale modes of variability estimated from the analyzed observations, sometimes supplemented with satellite SSTs to better define variability and covariance. However, they also must be used with caution, and can be especially problematic when the observations are sparse (providing too weak a constraint and leading to damped variability) or noisy (which can project spurious values over the large areas covered by the EOFs).

Each product uses a different approach to avoid these problems. In ERSST, the data are smoothed at large spatial and temporal scales, and the residual differences from this smooth field are then analyzed using empirical orthogonal teleconnections, which are EOFs restricted in their spatial range. Individual empirical orthogonal teleconnections are used only where available observations provide a good constraint. In HadISST (Rayner et al. 2003), a single global pattern is first removed, which preserves the large-scale trend. Residuals from this are analyzed using a relatively small number of leading EOFs (Kaplan et al. 1997) with a prior constraint on their variability. The trend pattern is reinstated and values from well-sampled grid cells are blended with the EOF-reconstructed fields to minimize the loss of variance that would otherwise occur

from using a reduced set of EOFs. COBE-SST2 (Hirahara et al. 2014) uses a set of EOF patterns estimated from satellite and in situ observations to reconstruct gridded fields of in situ data following a similar two-step process to the one used by HadISST to reconstruct the monthly variability. In addition, a daily optimal interpolation analysis is performed to reconstruct high-frequency variability.

Because of the lack of observation in marginal sea ice areas, SSTs in areas covered by sea ice have been estimated based on the measured sea-ice concentration. As the concentration approaches 100%, the SST tends toward the freezing point of water (Reynolds et al. 2002, Rayner et al. 2003). A recent comparison study by Banzon et al. (2020) found that the errors in current schemes were largest during the warm season and lowest when sea-ice concentrations were highest in the cold season. Banzon et al. (2020) showed differences between the approaches they tested of more than 1°C locally but more typically approximately 0.2°C, but the effect on the global mean will be smaller because the marginal ice zone covers approximately 10% of the ocean area at 5° grid-cell resolution and a smaller percentage for smaller grid cells. However, SST anomalies in areas uncovered by the long-term retreat of sea ice, where the climatological average is at the freezing point of sea water, can be large, reaching several degrees in some cases in the Arctic.

All of the methods produce estimates of the analysis error, although this does not represent the full uncertainty in the resulting data product. Ensembles are increasingly being used to provide uncertainty information (Karspeck et al. 2012, Huang et al. 2017, Kennedy et al. 2019).

5.3. Current Gridded Products

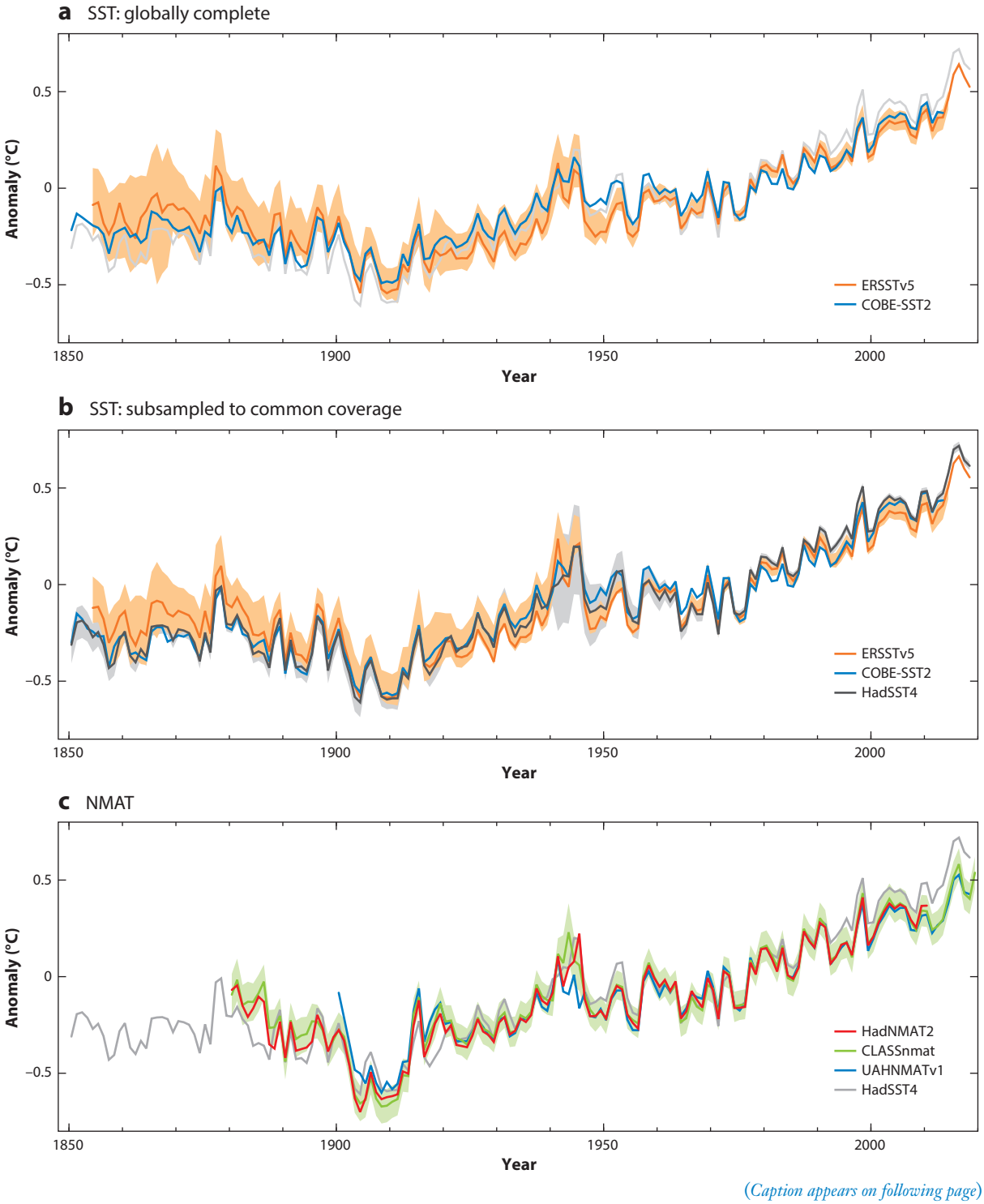
Gridded data products are periodically updated to incorporate recent data additions and the latest analysis methods. The most recent products are summarized here; the **Supplemental Appendix** provides more information about these data products and shows their evolution through preceding versions.

Supplemental Material >

5.3.1. Sea-surface temperature. The following data sets describe the latest versions of the three main lines of historical SST data set development. These products are available for analysis in the IPCC's Sixth Assessment Report, and time series of their anomalies are shown in **Figure 6**. A 1961–1990 climatology is used to provide a common baseline for all the data sets shown; this was the period used in chapter 2 of the IPCC's Fifth Assessment Report.

ERSSTv5 (Huang et al. 2017) is based on the latest release of ICOADS (Freeman et al. 2017) supplemented using near-surface measurement from Argo floats. Bias adjustments, based on air–sea temperature differences, ship–buoy differences, and Argo–buoy differences, are applied to the whole data set. Gaps are filled using a combination of large-scale smoothing and empirical orthogonal teleconnections. In marginal sea ice areas, SST is inferred from the sea-ice concentration in HadISST2 (Titchner & Rayner 2014). Uncertainty is expressed using an ensemble method, where the ensemble is generated by varying parameters and choices in the analysis method and includes uncertainty in the bias adjustments and infilling. ERSSTv5 is provided as an operational member, which represents the preferred choices and parameters, and an ensemble.

HadSST4 (Kennedy et al. 2019) is also based on the latest version of ICOADS. Bias adjustments, based on physical models of buckets and comparisons with buoys and oceanographic profiles, have been applied to the whole data set. Gaps in the data set are not filled. Uncertainties arising from measurement and sampling errors are estimated at a grid box level, and error covariances summarizing error correlations between grid boxes are also calculated. Additional uncertainty in the bias adjustments is represented using a 200-member ensemble in which uncertain parameters and underdetermined choices are varied.



(*Caption appears on following page*)

Figure 6 (*Figure appears on preceding page*)

(*a*) Globally complete annual average SST anomalies (relative to 1961–1990) for ERSSTv5 (*orange line*, with a *shaded orange area* showing the 95% uncertainty range from a 1,000-member ensemble) and COBE-SST2 (*blue line*). (*b*) Annual average SST anomalies (relative to 1961–1990) for ERSSTv5 (*orange line*, with a *shaded orange area* showing the 95% uncertainty range), COBE-SST2 (*blue line*), and HadSST4 (*dark gray line*, with a *shaded gray area* showing the 95% uncertainty range from a 200-member ensemble, with additional uncertainties due to measurement and sampling), where all three are reduced to the grid and coverage of HadSST4. (*c*) Annual average NMAT anomalies (relative to 1961–1990) for HadNMAT2 (*red line*), UAHNMATv1 (*blue line*), and CLASSnmat (*green line*, with a *shaded green area* showing the two-standard-deviation uncertainty range), along with HadSST4 data (*light gray line*) for comparison. Abbreviations: CLASSnmat, Climate Linked Atlantic Sector Science Night Marine Air Temperature; COBE-SST2, Centennial In Situ Observation-Based Estimates Sea Surface Temperature 2; ERSSTv5, Extended Reconstructed Sea Surface Temperature version 5; HadNMAT2, Hadley Centre Night Marine Air Temperature 2; HadSST4, Hadley Centre Sea Surface Temperature 4; NMAT, nighttime marine air temperature; UAHNMATv1, University of Alabama in Huntsville Nighttime Marine Air Temperature version 1.

COBE-SST2 (Hirahara et al. 2014) is based on ICOADS release 2.5 (Woodruff et al. 2011) supplemented by observations archived by the Japanese Fisheries Agency. Bias adjustments based on physical models of buckets and comparisons with buoys and MAT have been applied to the whole data set. Gaps in the data set are filled using an EOF analysis performed at a 30-day timescale and an optimal interpolation analysis performed daily. Uncertainty in the fields is estimated based on the uncertainty estimates from the statistical infilling.

5.3.2. Nighttime marine air temperature. There are two current NMAT data products, and time series of their anomalies are shown in **Figure 6**, along with data from the older product Hadley Centre Night Marine Air Temperature 2 (HadNMAT2) (Kent et al. 2013). The University of Alabama in Huntsville Nighttime Marine Air Temperature version 1 (UAHNMATv1) data product (Junod & Christy 2020) is based on ICOADS release 3.0 (Freeman et al. 2017) NMAT data since 1900. Adjustments for height are applied using atmospheric temperature profiles from reanalysis, and the World War II bias is removed by assuming a linear change over the period 1936–1950; other adjustments follow Kent et al. (2013). Anomalies are calculated using a daily 1.25° climatology, then averaged monthly and interpolated to the final 5° grid. Filtering is applied as part of this process to remove unlikely values. Uncertainties are estimated for grid cells following Rayner et al. (2006), acknowledging that this will underestimate the uncertainties (Kent et al. 2013).

Climate Linked Atlantic Sector Science Night Marine Air Temperature (CLASSnmat) (Cornes et al. 2020) is also based on ICOADS release 3.0 (Freeman et al. 2017) and is an update to the HadNMAT2 data set (Kent et al. 2013) that makes improvements to the methods to estimate measurement heights and to the gridding and uncertainty methodologies. The World War II bias was tackled by using high-quality naval observations as a reference to generate per-ship adjustments for lower-quality data sources. Gridding uses a weighted mean that accounts for random, spatially correlated, and platform-correlated effects.

5.3.3. Global mean surface temperature. The main GMST gridded data products are based on only two SST data products. Morice et al. (2012), Rohde et al. (2013), Cowtan & Way (2014), and Rohde & Hausfather (2020) used HadSST3 (Kennedy et al. 2011a), and Vose et al. (2012) and Lenssen et al. (2019) used ERSSTv5 (Huang et al. 2017). No operational GMST products currently use NMAT. Blending of the land air temperatures and SST is done simply by weighted averages of grid-cell values spanning land and ocean. In areas that are covered by sea ice, statistically interpolated land air temperatures are used in the Goddard Institute of Space Studies Surface Temperature Analysis (GISTEMP) and Berkeley Earth.

5.4. How Well Do Observational Products Represent Historical Surface Marine Temperatures?

The global mean temperature anomaly time series shown in **Figure 6** illustrate our knowledge from the best products currently available, with the caveat that the most recent insights have not yet been incorporated into the global products. The SST products agree better when subsampled to common coverage (compare **Figure 6*a*** and **6*b***). The subsampling cannot work perfectly, however; the different products have different resolutions, so sub-grid-cell variations may contribute to the differences. Unsurprisingly, the record in the nineteenth century is the most uncertain, and improvements in this period will help to constrain our estimates of early industrial temperatures and hence the full extent of anthropogenically forced temperature change. These improvements may come from new approaches to bias estimation (Kent et al. 2017), additional observations from data rescue, and better methods to analyze the sparse data. Ultimately, understanding of the changes in this early period may be limited by the lack of reference data.

Temperature changes in the early twentieth century are inconsistent with climate model predictions (Hegerl et al. 2018), and there are unexplained differences between SST and NMAT at both global and regional scales. Observations from this period derive from many different sources, and Chan et al. (2019) showed that reconciling differences between sources could resolve regional inconsistencies. This type of analysis has not yet been implemented in any of the products shown here. Estimating temperature change in the period around World War II remains problematic, but Cornes et al. (2020) managed to substantially reduce the spurious warmth in NMAT during this period by adjusting data from each ship using data sources showing warm biases, based on nearby measurements from a higher-quality source as a reference.

There has been an intense focus on the most recent decades, when the temperature has increased rapidly, and differences among the products can been seen right up to the present. Care is needed, however, when interpreting these plots. Plotted anomalies are referenced to a specific climatological period—in this case, 1961–1990—so the averages of temperatures for all Januaries, and for each of the 11 other months, will be zero over this period. Some of the differences in recent anomalies therefore arise from global or regional differences between the products during the climatological period, when ship-based measurement methods were changing rapidly (**Figure 5**) and new platform types were entering the record (**Figure 2**). In **Figure 6*c***, for example, the NMAT anomalies increase more slowly than the SST anomalies, as represented by HadSST4. However, the NMAT and SST anomalies are referenced to a climatology where SST biases are known to be large and the adjustments uncertain. It could be that the differential in trends is an artifact of the choice of climatological period and that using a more recent climatology, along with a detailed examination of regional trends, might help to diagnose the source of the differences.

This article has focused on the many problems associated with the creation of gridded global surface temperature products from the available observations, and the challenges to be overcome can seem daunting. So where do we stand? There is a small but growing range of historical SST and NMAT products, and all show that the world is warming. All of these products now come with uncertainty estimates, which enable users to understand the likely limitations of each product. Regions and periods where differences between products cannot be explained by the uncertainty estimates expose deficiencies in our understanding that can spur new developments. The uncertainty of observation-based estimates, especially regionally, potentially limits our assessment of climate models, and to keep pace with model improvements will require exploitation of new data and methods—including in ways outlined in the next section.

6. THE WAY FORWARD

The direction of future research is likely to continue to increasingly analyze individual observations to more appropriately account for the varying sources of error. Recent papers have continued to improve our understanding of the sources of error and uncertainty in marine data, but the ideas are not yet incorporated into gridded data sets. Carella et al. (2017a) used a probabilistic algorithm to assign marine reports to coherent ship tracks. Such information could be used to improve quality control of data, estimates of uncertainty, and bias adjustments. Chan & Huybers (2019), Chan et al. (2019), and Davis et al. (2019) identified biases at a basin level that remained after the then-current generation of bias adjustments (Kennedy et al. 2011b) had been applied to the data. Biases were identified with errors that are correlated at the level of individual decks in ICOADS (Chan & Huybers 2019). Errors at this intermediate level between ships and the global fleet are not explicitly dealt with in current gridded data sets, though Kennedy et al. (2019) allowed for spatial variation in biases associated with ERI measurements, and Huang et al. (2018) included ensemble members with a greater degree of spatial flexibility in the bias adjustment fields. Methods applied to SST should also be applied to MAT and independent climate records derived for each parameter.

As the focus moves from global biases to individual observations, statistical models will be needed that are capable of efficiently analyzing temperature at scales from a single observation to the global mean as well as the computing capacity to run them. The European Union Surface Temperature for All Corners of Earth (EUSTACE) project (Rayner et al. 2020) produced a gridded data set of daily air temperatures at 0.25° spatial resolution for the period 1850–2015 based on daily measurements of temperature at land stations and individual NMAT reports as well as satellite-based estimates of air temperature over land, oceans, and ice. The data were brought together in a statistical space-time model that also incorporated and jointly estimated data errors with varying correlation structures. Getting the best out of such models requires a sound understanding of the errors and error structures in the observations. Increased collaboration with statisticians to develop the tools needed to do these analyses and with metrologists to fully understand and quantify the uncertainties will speed progress.

Fresh perspectives will also come from joint evaluation of in situ–based historical marine surface temperature and other estimates. Examples include paleoreconstructions, temperatures over land, ocean climate model output, and centennial-scale reanalyses. Collaborative work between those developing data products and those evaluating or analyzing them will over time lead to more robust and useful data products.

FUTURE ISSUES

1. Improvements to the historical archive of surface marine data are needed to capitalize on new approaches to estimate bias and uncertainty in marine surface temperatures. This will require reprocessing some existing data sources from their earliest available versions.
2. More data and metadata are always helpful; data rescue must continue. Following best practices will ensure that documentation, metadata, and information on ambient environmental conditions are available to improve bias adjustment and uncertainty estimation.
3. Future observing systems must take account of the requirements for climate data records in addition to those for operational activities, such as numerical weather prediction (see

recommendations in Kent et al. 2019). Specifically, quantifying errors and uncertainty associated with observations used to construct climate records typically requires documentation about how the observations were made along with information on ambient environmental conditions, making multivariate observations particularly valuable.

4. Many different approaches are required to bias and uncertainty estimation in order to enable the development of a wider range of gridded data products and ensembles to help quantify structural uncertainties.
5. The challenge of seamlessly incorporating satellite data into centennial-scale analyses needs to be met.
6. Open and interoperable data along with open-access analysis code will allow a wider range of researchers to contribute and increase the number of data sources available for improved analysis and evaluation, as well as speeding up the incorporation of new understanding into gridded data sets.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Supplemental Material >

Sources of data used in this article are included in the **Supplemental Appendix**. Some of the graphics were produced using the R software package (**https://www.r-project.org**). E.C.K. received funding from the Natural Environment Research Council projects Climate Linked Atlantic Sector Science Nighttime Marine Air Temperature (CLASS, NE/R015953/1); Historical Ocean Surface Temperatures: Accuracy, Characterisation, and Evaluation (HOSTACE, NE/J020788/1); and Global Surface Air Temperature (GloSAT, NE/S015647/2). J.J.K. was supported by the Met Office Hadley Centre Climate Program, funded by BEIS and Defra.

LITERATURE CITED

Allan R, Endfield G, Damodaran V, Adamson G, Hannaford M, et al. 2016. Toward integrated historical climate research: the example of Atmospheric Circulation Reconstructions over the Earth. *Wiley Interdiscip. Rev. Clim. Change* 7:164–74

Argo. 2020. Argo float data and metadata from Global Data Assembly Centre (Argo GDAC). *SEANOE*. **https://doi.org/10.17882/42182**

Ashford OM. 1948. A new bucket for measurement of sea surface temperature. *Q. J. R. Meteorol. Soc.* 74:99–104

Atkinson CP, Rayner NA, Kennedy JJ, Good SA. 2014. An integrated database of ocean temperature and salinity observations. *J. Geophys. Res. Oceans* 119:7139–63

Atkinson CP, Rayner NA, Roberts-Jones J, Smith RO. 2013. Assessing the quality of sea surface temperature observations from drifting buoys and ships on a platform-by-platform basis: assessing buoy and ship SST observations. *J. Geophys. Res. Oceans* 118:3507–29

Banzon V, Smith TM, Steele M, Huang B, Zhang HM. 2020. Improved estimation of proxy sea surface temperature in the arctic. *J. Atmos. Ocean. Technol.* 37:341–49

Barnett TP. 1984. Long-term trends in surface temperature over the oceans. *Mon. Weather Rev.* 112:303–12

Berry DI, Kent EC. 2005. The effect of instrument exposure on marine air temperatures: an assessment using VOSClim data. *Int. J. Climatol.* 25:1007–22

Berry DI, Kent EC. 2009. A new air-sea interaction gridded dataset from ICOADS with uncertainty estimates. *Bull. Am. Meteorol. Soc.* 90:645–56

Berry DI, Kent EC. 2017. Assessing the health of the in situ global surface marine climate observing system. *Int. J. Climatol.* 37:2248–59

Berry DI, Kent EC, Taylor PK. 2004. An analytical model of heating errors in marine air temperatures from ships. *J. Atmos. Ocean. Technol.* 21:1198–215

Bottomley M, Folland C, Hsiung J, Newell R, Parker D. 1990. *Global Ocean Surface Temperature Atlas "GOSTA."* Bracknell, UK/Cambridge, MA: Meteorol. Off./Mass. Inst. Technol.

Bowditch N. 1802. *The American Practical Navigator*. Washington, DC: US Navy Hydrogr. Off.

Boyer TP, Antonov JI, Baranova OK, Coleman C, Garcia HE, et al. 2016. *World Ocean Database 2013 (NCEI accession 0117075)*. Data Set, Natl. Cent. Environ. Inf., Natl. Ocean. Atmos. Adm., Silver Spring, MD. **https://doi.org/10.7289/v54q7s16**

Brohan P, Allan R, Freeman JE, Waple AM, Wheeler D, et al. 2009. Marine observations of old weather. *Bull. Am. Meteorol. Soc.* 90:219–30

Brohan P, Allan R, Freeman JE, Wheeler D, Wilkinson C, Williamson F. 2012. Constraining the temperature history of the past millennium using early instrumental observations. *Clim. Past* 8:1551–63

Brooks C. 1928. Problems related to surface-water temperature: reliability of different methods of measuring sea-surface temperatures. *J. Wash. Acad. Sci.* 18:525–45

Budyko MI, Yefimova NA, Aubenok LI, Strokina LA. 1962. The heat balance of the surface of the earth. *Soviet Geogr.* 3:3–16

Bulgin C, Embury O, Corlett GK, Merchant C. 2016. Independent uncertainty estimates for coefficient based sea surface temperature retrieval from the Along-Track Scanning Radiometer instruments. *Remote Sens. Environ.* 178:213–22

Bunker AF. 1976. Computations of surface energy flux and annual air-sea interaction cycles of the North Atlantic Ocean. *Mon. Weather Rev.* 104:1122–40

Businger JA, Wyngaard JC, Izumi Y, Bradley EF. 1971. Flux-profile relationships in the atmospheric surface layer. *J. Atmos. Sci.* 28:181–89

Carella G. 2017. *New estimates of uncertainty in the marine surface temperature record*. PhD Thesis, Univ. Southampton, Southampton, UK

Carella G, Kennedy JJ, Berry DI, Hirahara S, Merchant CJ, et al. 2018. Estimating sea surface temperature measurement methods using characteristic differences in the diurnal cycle. *Geophys. Res. Lett.* 45:363–71

Carella G, Kent EC, Berry DI. 2017a. A probabilistic approach to ship voyage reconstruction in ICOADS. *Int. J. Climatol.* 37:2233–47

Carella G, Morris AKR, Pascal RW, Yelland MJ, Berry DI, et al. 2017b. Measurements and models of the temperature change of water samples in sea-surface temperature buckets. *Q. J. R. Meteorol. Soc.* 143:2198–209

Centurioni LR, Turton J, Lumpkin R, Braasch L, Brassington G, et al. 2019. Global in situ observations of essential climate and ocean variables at the air–sea interface. *Front. Mar. Sci.* 6:419

Chan D, Huybers P. 2019. Systematic differences in bucket sea surface temperature measurements among nations identified using a linear-mixed-effect method. *J. Clim.* 32:2569–89

Chan D, Huybers P. 2020. Systematic differences in bucket sea surface temperature measurements caused by misclassification of engine room intake measurements. *J. Clim.* 33:7735–53

Chan D, Kent EC, Berry DI, Huybers P. 2019. Correcting datasets leads to more homogeneous early-twentieth-century sea surface warming. *Nature* 571:393–97

Chenoweth M. 1996. Nineteenth-century marine temperature data: comments on observing practices and potential biases in marine datasets. *Weather* 51:280–85

Chenoweth M. 2000. A new methodology for homogenization of 19th century marine air temperature data. *J. Geophys. Res. Atmos.* 105:29145–54

Chenoweth M. 2001. Two major volcanic cooling episodes derived from global marine air temperature, AD 1807–1827. *Geophys. Res. Lett.* 28:2963–66

Christy JR. 2001. Differential trends in tropical sea surface and atmospheric temperatures since 1979. *Geophys. Res. Lett.* 28:183–86

CIMO (Comm. Instr. Methods Obs.). 2017. *Guide to Meteorological Instruments and Methods of Observation*. Geneva: World Meteorol. Organ.
Cornes RC, Kent EC, Berry DI, Kennedy JJ. 2020. CLASSnmat: a global night marine air temperature data set, 1880–2018. *Geosci. Data J.* **https://doi.org/10.1002/gdj3.100**
Cowtan K, Hausfather Z, Hawkins E, Jacobs P, Mann ME, et al. 2015. Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures. *Geophys. Res. Lett.* 42:6526–34
Cowtan K, Rohde R, Hausfather Z. 2018. Evaluating biases in sea surface temperature records using coastal weather stations. *Q. J. R. Meteorol. Soc.* 144:670–81
Cowtan K, Way RG. 2014. Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Q. J. R. Meteorol. Soc.* 140:1935–44
Davis LLB, Thompson DWJ, Kennedy JJ, Kent EC. 2019. The importance of unresolved biases in twentieth-century sea surface temperature observations. *Bull. Am. Meteorol. Soc.* 100:621–29
Davis RE, Talley LD, Roemmich D, Owens WB, Rudnick DL, et al. 2018. 100 years of progress in ocean observing systems. *Meteorol. Monogr.* 59:3.1–3.46
Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, et al. 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* 137:553–97
Farmer G, Wigley TML, Jones PD, Salmon M. 1989. *Documenting and explaining recent global-mean temperature changes*. Rep., Clim. Res. Unit, Norwich, UK
Folger T. 1787. Chart of the Gulf Stream. In *Philosophical and Miscellaneous Papers*, ed. B Franklin, facing p. 122. London: C. Dilly
Folland CK, Parker DE. 1995. Correction of instrumental biases in historical sea surface temperature data. *Q. J. R. Meteorol. Soc.* 121:319–67
Folland CK, Parker DE, Kates FE. 1984. Worldwide marine temperature fluctuations 1856–1981. *Nature* 310:670–73
Folland CK, Rayner NA, Brown SJ, Smith TM, Shen SSP, et al. 2001. Global temperature change and its uncertainties since 1861. *Geophys. Res. Lett.* 28:2621–24
Folland CK, Salinger MJ. 1995. Surface temperature trends and variations in New Zealand and the surrounding ocean, 1871–1993. *Int. J. Climatol.* 15:1195–218
Franklin B. 1786. A Letter from Dr. Benjamin Franklin, to Mr. Alphonsus le Roy, member of several academies, at Paris. Containing sundry maritime observations. *Trans. Am. Philos. Soc.* 2:294–329
Freeman E, Kent EC, Brohan P, Cram T, Gates L, et al. 2019. The International Comprehensive Ocean-Atmosphere Data Set – meeting users needs and future priorities. *Front. Mar. Sci.* 6:435
Freeman E, Woodruff SD, Worley SJ, Lubker SJ, Kent EC, et al. 2017. ICOADS release 3.0: a major update to the historical marine climate record. *Int. J. Climatol.* 37:2211–32
Gu Z, Gu L, Eils R, Schlesner M, Brors B. 2014. *circlize* implements and enhances circular visualization in R. *Bioinformatics* 30:2811–12
Hanawa K, Yasunaka S, Manabe T, Iwasaka N. 2000. Examination of correction to historical SST data using long-term coastal SST data taken around Japan. *J. Meteorol. Soc. Jpn. II* 78:187–95
Hausfather Z, Cowtan K, Clarke DC, Jacobs P, Richardson M, Rohde R. 2017. Assessing recent warming using instrumentally homogeneous sea surface temperature records. *Sci. Adv.* 3:e1601207
Hawkins E, Ortega P, Suckling E, Schurer A, Hegerl G, et al. 2017. Estimating changes in global temperature since the preindustrial period. *Bull. Am. Meteorol. Soc.* 98:1841–56
Hegerl GC, Brönnimann S, Cowan T, Friedman AR, Hawkins E, et al. 2019. Causes of climate change over the historical record. *Environ. Res. Lett.* 14:123006
Hegerl GC, Brönnimann S, Schurer A, Cowan T. 2018. The early 20th century warming: anomalies, causes, and consequences. *Wiley Interdiscip. Rev. Clim. Change* 9:e522
Hirahara S, Ishii M, Fukuda Y. 2014. Centennial-scale sea surface temperature analysis and its uncertainty. *J. Clim.* 27:57–75
Houghton JT, Jenkins GJ, Ephraums J, eds. 1990. *Climate Change: the IPCC Scientific Assessment*. Cambridge, UK: Cambridge Univ. Press
Huang B, Angel W, Boyer T, Cheng L, Chepurin G, et al. 2018. Evaluating SST analyses with independent ocean profile observations. *J. Clim.* 31:5015–30

Huang B, Thorne PW, Banzon VF, Boyer T, Chepurin G, et al. 2017. Extended Reconstructed Sea Surface Temperature, version 5 (ERSSTv5): upgrades, validations, and intercomparisons. *J. Clim.* 30:8179–205
James RW, Fox PT. 1972. *Comparative sea surface temperature measurements*. Rep. Mar. Sci. Aff. 5, WMO Rep. 336, World Meteorol. Organ., Geneva
Jones GS, Stott PA, Mitchell JFB. 2016. Uncertainties in the attribution of greenhouse gas warming and implications for climate prediction. *J. Geophys. Res. Atmos.* 121:6969–92
Jones P, Wigley T, Folland C, Parker D, Angell J, et al. 1988. Evidence for global warming in the past decade. *Nature* 332:790
Josey SA, Kent EC, Taylor PK. 1999. New insights into the ocean heat budget closure problem from analysis of the SOC air-sea flux climatology. *J. Clim.* 12:2856–80
Junod RA, Christy JR. 2020. A new compilation of globally gridded night-time marine air temperatures: the UAHNMATv1 dataset. *Int. J. Climatol.* 40:2609–23
Kaplan A, Kushnir Y, Cane MA, Blumenthal MB. 1997. Reduced space optimal analysis for historical data sets: 136 years of Atlantic sea surface temperatures. *J. Geophys. Res. Oceans* 102:27835–60
Karl TR, Arguez A, Huang B, Lawrimore JH, McMahon JR, et al. 2015. Possible artifacts of data biases in the recent global surface warming hiatus. *Science* 348:1469–72
Karspeck AR, Kaplan A, Sain SR. 2012. Bayesian modelling and ensemble reconstruction of mid-scale spatial variability in North Atlantic sea-surface temperatures for 1850–2008. *Q. J. R. Meteorol. Soc.* 138:234–48
Kennedy JJ. 2014. A review of uncertainty in in situ measurements and data sets of sea surface temperature. *Rev. Geophys.* 52:1–32
Kennedy JJ, Rayner NA, Atkinson CP, Killick RE. 2019. An ensemble data set of sea surface temperature change from 1850: the Met Office Hadley Centre HadSST.4.0.0.0 data set. *J. Geophys. Res. Atmos.* 124:7719–63
Kennedy JJ, Rayner NA, Smith RO, Parker DE, Saunby M. 2011a. Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties. *J. Geophys. Res. Atmos.* 116:D14103
Kennedy JJ, Rayner NA, Smith RO, Parker DE, Saunby M. 2011b. Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *J. Geophys. Res. Atmos.* 116:D14104
Kent EC, Kaplan A. 2006. Toward estimating climatic trends in SST. Part III: systematic biases. *J. Atmos. Ocean. Technol.* 23:487–500
Kent EC, Kennedy JJ, Berry DI, Smith RO. 2010. Effects of instrumentation changes on sea surface temperature measured in situ. *Wiley Interdiscip. Rev. Clim. Change* 1:718–28
Kent EC, Kennedy JJ, Smith TM, Hirahara S, Huang B, et al. 2017. A call for new approaches to quantifying biases in observations of sea surface temperature. *Bull. Am. Meteorol. Soc.* 98:1601–16
Kent EC, Rayner NA, Berry DI, Eastman R, Grigorieva VG, et al. 2019. Observing requirements for long-term climate records at the ocean surface. *Front. Mar. Sci.* 6:441
Kent EC, Rayner NA, Berry DI, Saunby M, Moat BI, et al. 2013. Global analysis of night marine air temperature and its uncertainty since 1880: the HadNMAT2 data set. *J. Geophys. Res. Atmos.* 118:1281–98
Kent EC, Taylor PK. 2006. Toward estimating climatic trends in SST. Part I: methods of measurement. *J. Atmos. Ocean. Technol.* 23:464–75
Kent EC, Taylor PK, Truscott BS, Hopkins JS. 1993. The accuracy of voluntary observing ships meteorological observations—results of the VSOP-NA. *J. Atmos. Ocean. Technol.* 10:591–608
Kent EC, Woodruff SD, Berry DI. 2007. Metadata from WMO Publication No. 47 and an assessment of voluntary observing ship observation heights in ICOADS. *J. Atmos. Ocean. Technol.* 24:214–34
Lenssen NJL, Schmidt GA, Hansen JE, Menne MJ, Persin A, et al. 2019. Improvements in the GISTEMP uncertainty model. *J. Geophys. Res. Atmos.* 124:6307–26
Masson-Delmotte V, Zhai P, Pörtner H-O, Roberts D, Skea J, et al., eds. 2018. *Global warming of 1.5°C: an IPCC special report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty*. Rep., World Meteorol. Organ., Geneva
Matthews JBR. 2013. Comparing historical and modern methods of sea surface temperature measurement – part 1: review of methods, field comparisons and dataset adjustments. *Ocean Sci.* 9:683–94

Maury MF. 1854. *Maritime Conference Held at Brussels for Devising an Uniform System of Meteorological Observations at Sea, August and September 1853*. Philadelphia: E.C. & J. Biddle

Medhaug I, Stolpe MB, Fischer EM, Knutti R. 2017. Reconciling controversies about the 'global warming hiatus'. *Nature* 545:41–47

Merchant CJ, Embury O, Roberts-Jones J, Fiedler E, Bulgin CE, et al. 2014. Sea surface temperature datasets for climate applications from phase 1 of the European Space Agency Climate Change Initiative (SST CCI). *Geosci. Data J.* 1:179–91

Minnett PJ, Alvera-Azcárate A, Chin T, Corlett G, Gentemann C, et al. 2019. Half a century of satellite remote sensing of sea-surface temperature. *Remote Sens. Environ.* 233:111366

Morak-Bozzo S, Merchant CJ, Kent EC, Berry DI, Carella G. 2016. Climatological diurnal variability in sea surface temperature characterized from drifting buoy data. *Geosci. Data J.* 3:20–28

Morice CP, Kennedy JJ, Rayner NA, Jones PD. 2012. Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: the HadCRUT4 data set. *J. Geophys. Res. Atmos.* 117:D08101

Paltridge G, Woodruff S. 1981. Changes in global surface temperature from 1880 to 1977 derived from historical records of sea surface temperature. *Mon. Weather Rev.* 109:2427–34

Parker DE, Folland CK, Jackson M. 1995. Marine surface temperature: observed variations and data requirements. *Clim. Change* 31:559–600

Rayner NA, Auchmann R, Bessembinder J, Brönnimann S, Brugnara Y, et al. 2020. The EUSTACE project: delivering global daily information on surface air temperature. *Bull. Am. Meteorol. Soc.* In press. **https://doi.org/10.1175/BAMS-D-19-0095.1**

Rayner NA, Brohan P, Parker DE, Folland CK, Kennedy JJ, et al. 2006. Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: the HadSST2 dataset. *J. Clim.* 19:446–69

Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, et al. 2003. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res. Atmos.* 108:4407

Rennell J. 1832. *An Investigation of the Currents of the Atlantic Ocean, and of Those Which Prevail Between the Indian Ocean and the Atlantic*. London: J.G. & F. Rivington

Reynolds RW. 1988. A real-time global sea surface temperature analysis. *J. Clim.* 1:75–87

Reynolds RW, Rayner NA, Smith TM, Stokes DC, Wang W. 2002. An improved in situ and satellite SST analysis for climate. *J. Clim.* 15:1609–25

Richardson M, Cowtan K, Hawkins E, Stolpe MB. 2016. Reconciled climate response estimates from climate models and the energy budget of earth. *Nat. Clim. Change* 6:931–35

Rohde RA, Hausfather Z. 2020. The Berkeley Earth land/ocean temperature record. *Earth Syst. Sci. Data Discuss.* In review. **https://doi.org/10.5194/essd-2019-259**

Rohde RA, Muller R, Jacobsen R, Perlmutter S, Rosenfeld A, et al. 2013. Berkeley Earth temperature averaging process. *Geoinform. Geostat. Overview* 1. **https://doi.org/10.4172/2327-4581.1000103**

Roll H. 1951. The accuracy of measuring water temperature with the water scoop thermometer (Marinepütz–German scoop thermometer). *Ann. Meteorol.* 10–12:480–82

Saur JFT. 1963. A study of the quality of sea water temperatures reported in logs of ships' weather observations. *J. Appl. Meteorol.* 2:417–25

Schurer AP, Cowtan K, Hawkins E, Mann ME, Scott V, Tett SFB. 2018. Interpretations of the Paris climate target. *Nat. Geosci.* 11:220–21

Slutz RJ, Lubker SJ, Hiscox JD, Woodruff SD, Jenne RL, et al. 1985. *Comprehensive Ocean-Atmosphere Data Set: Release 1*. Boulder, CO: Clim. Res. Program, Environ. Res. Lab., Natl. Ocean. Atmos. Adm.

Smith SR, Briggs K, Bourassa MA, Elya J, Paver CR. 2018. Shipboard automated meteorological and oceanographic system data archive: 2005–2017. *Geosci. Data J.* 5:73–86

Smith TM, Reynolds RW. 2002. Bias corrections for historical sea surface temperatures based on marine air temperatures. *J. Clim.* 15:73–87

Stocker TF, Qin D, Plattner G, Tignor M, Allen S, et al., eds. 2013. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK: Cambridge Univ. Press

Strickland W. 1802. On the use of the thermometer in navigation. *Trans. Am. Philos. Soc.* 5:90–103
Thompson DW, Kennedy JJ, Wallace JM, Jones PD. 2008. A large discontinuity in the mid-twentieth century in observed global-mean surface temperature. *Nature* 453:646–49
Titchner HA, Rayner NA. 2014. The Met Office Hadley Centre sea ice and sea surface temperature data set, version 2. *J. Geophys. Res. Atmos.* 119:2864–89
Tokarska KB, Schleussner CF, Rogelj J, Stolpe MB, Matthews HD, et al. 2019. Recommended temperature metrics for carbon budget estimates, model evaluation and climate policy. *Nat. Geosci.* 12:964–71
Viglione G. 2020. How COVID-19 could ruin weather forecasts and climate records. *Nature*, Apr. 13. **https://www.nature.com/articles/d41586-020-00924-6**
Vose RS, Arndt D, Banzon VF, Easterling DR, Gleason B, et al. 2012. NOAA merged land-ocean surface temperature analysis. *Bull. Am. Meteorol. Soc.* 93:1677–85
Walker M. 2006. The weather observations of Surgeon Menzies. *Weather* 61:315–19
Wallbrink H, Koek F, Brandsma T. 2009. *The US Maury collection metadata 1796–1861*. Rep. 225, HISKLIM-11, K. Ned. Meteorol. Inst., De Bilt, Neth.
Wang J, Yang B, Ljungqvist FC, Luterbacher J, Osborn TJ, et al. 2017. Internal and external forcing of multidecadal Atlantic climate variability over the past 1,200 years. *Nat. Geosci.* 10:512–17
Wilkinson C, Woodruff SD, Brohan P, Claesson S, Freeman E, et al. 2011. Recovery of logbooks and international marine data: the RECLAIM project. *Int. J. Climatol.* 31:968–79
Williams SPD, Berry DI. 2020. ACSIS Atlantic Ocean medium resolution SST dataset: reconstructed 5-day, half degree, Atlantic Ocean SST (1950–2014). *Geosci. Data J.* **https://doi.org/10.1002/gdj3.94**
WMO (World Meteorol. Organ.). 1957. *Abridged Final Report of the Second Session: Paris, 18th June - 6th July 1957*. Geneva: World Meteorol. Organ.
WMO (World Meteorol. Organ.). 1968. *Abridged Final Report of the Fourth Session: Geneva, 23 November - 8 December 1964*. Geneva: World Meteorol. Organ.
WMO (World Meteorol. Organ.). 2018. *Manual on the Global Telecommunication System: Annex III to the WMO Technical Regulations*. Geneva: World Meteorol. Organ.
Woodruff SD, Diaz HF, Worley SJ, Reynolds RW, Lubker SJ. 2005. Early ship observational data and ICOADS. *Clim. Change* 73:169–94
Woodruff SD, Worley SJ, Lubker SJ, Ji Z, Freeman JE, et al. 2011. ICOADS release 2.5: extensions and enhancements to the surface marine meteorological archive. *Int. J. Climatol.* 31:951–67
Wright PB. 1986. Problems in the use of ship observations for the study of interdecadal climate changes. *Mon. Weather Rev.* 114:1028–34