An observational record of global gridded near surface air temperature change over land and ocean from 1781

Colin P. Morice [1], David I. Berry[2], Richard C. Cornes[2], Kathryn Cowtan[3], Thomas Cropper[2], Ed Hawkins[4], John Kennedy[1], Timothy J. Osborn[5], Nick A. Rayner[1], Beatriz Recinos Rivas[6,2], Andrew P. Schurer[6], Michael Taylor[5], Praveen R. Teleti[4], Emily J. Wallis[5], Jonathan Winn[1], and Elizabeth C. Kent[2]

[1]Met Office, Exeter, EX1 3PB, United Kingdom
[2]National Oceanography Centre, Southampton, SO14 3ZH, United Kingdom
[3]University of York, York, YO10 5DD, United Kingdom
[4]National Centre for Atmospheric Science, Department of Meteorology, University of Reading, Reading. UK.
[5]Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, NR4 7TJ, United Kingdom
[6]University of Edinburgh, School of Geosciences, Edinburgh, EH9 3JW, United Kingdom

Correspondence: Colin P. Morice (colin.morice@metoffice.gov.uk)

Abstract. We present a new gridded data set of air temperature change across global land and ocean extending back to the 1780s. This data set, called the GloSAT reference analysis, has two novel features: it uses marine air temperature observations rather than the sea surface temperature measurements typically used by pre-existing data sets, and it extends further into the past than existing merged land and ocean instrumental temperature records which typically estimate temperature changes from the mid-to-late 19th century onwards. New estimates of diurnal heating biases in marine air temperatures have enabled the use of daytime observations, extending the dataset further into the past compared to nighttime-only marine air temperature data. The data set uses an extended version of the CRUTEM5 station database over land areas, incorporating newly available bias adjustments for non-standard thermometer enclosures used prior to the adoption of Stevenson screens and new climatological normal estimates for stations with limited data in the 1961-1990 baseline period. Land and marine temperature anomalies are combined to produce a gridded data set following the methods developed for HadCRUT5. The GloSAT global and hemispheric temperature anomaly

series show close agreement with those based on sea-surface temperature for much of the overlapping period of their records but with slightly less warming overall.

1 Introduction

Instrumental data sets recording changes and variations in near-surface temperature across the globe have been widely used to monitor changes in global and regional climate (World Meteorological Organisation, 2023). The Intergovernmental Panel on Climate Change 6th Assessment Report (IPCC AR6) assessment of global average temperature change from the mid-19th century (Gulev et al., 2021) is underpinned by instrumental data sets that combine information from Sea Surface Temperature (SST) observations with information from meteorological station observations of Land Surface Air Temperature (LSAT). The IPCC AR6 refers to change in mean near-surface temperature based on this combination of SST and LSAT as Global Mean Surface Temperature (GMST). Studies using climate model output commonly assess changes in near-surface temperature using air temperature changes over both land and ocean, using marine air temperature (MAT) rather than SST. The IPCC AR6 refers to this combination of LSAT and MAT as Global Surface Air Temperature (GSAT). Hence, the difference between GMST and GSAT is the use of SST or MAT to measure changes over the marine regions.

~~This article describes a new gridded GSAT data set. This data set has been constructed using a workflow in common with the existing HadCRUT5 GMST data set (Morice et al., 2021). We call this new data set the GloSAT reference analysis (GloSATref.1.0.0.0). It combines an extended and improved version of the CRUTEM5 LSAT data base, updated from Osborn et al. (2021), with a new marine MAT data set based on all-hours observations, building on Cornes et al. (2020). Section 1 summarises gridded datasets used in currently available observational GMST data sets. Section 2 describes the processing of input LSAT~~

~~and MAT data sets and their use to construct the GloSAT reference analysis. Section 3 presents global and regional climate diagnostics derived from the GloSAT reference analysis. Conclusions are presented in section 4.~~

Existing global instrumental data sets used to monitor changes in GMST combine measurements of near-surface air temperatures at meteorological stations (LSATs) with measurements of surface water temperatures obtained by ships, buoys, and moorings (SSTs). The currently available data sets of this type include HadCRUT5 (Morice et al., 2021), GISTEMP (Lenssen et al., 2019), NOAAGlobalTemp (Huang et al., 2022), Berkeley Earth (Rohde and Hausfather, 2020), Kadow et al. (Kadow et al., 2020), CMST 2.0 (Sun et al., 2022), DCENT (Chan et al., 2024), Calvert (Calvert, 2024), and COBE-STEMP3 (Ishii et al., 2025). Each of these data sets provides global grids that map GMST variability and change along with derived global and regional average time series. These data products have starting dates ranging from 1850 to 1880 as data availability, particularly of SST, and over land in the southern hemisphere, decreases rapidly prior to this.

Data sources for current observational surface temperature data sets are shown in Table 1. There is much common observational data underpinning the named LSAT and SST data sets because of global data sharing and consolidation of national records into global archives. For example, the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) (Freeman et al., 2017) is an underpinning source for each of the marine data sets in Table 1. Despite these overlaps in observation sources there are differences in selection criteria used in each data set and sometimes standard data holdings are extended through, e.g., addition of newly digitised historical sources. Each data set also differs in some or all of methods for data quality control (QC), methods used to account for systematic changes in observing practices, gridding methods and uncertainty modelling. Some data sets have independent processing workflows, such as NOAAGlobalTemp v6 and HadCRUT5. Others share aspects of processing, such as the reprocessing of HadCRUT5 data using alternative gridding methods by Kadow et al. (2020) and alternative gridding methods and bias adjustment in Calvert (2024). The set of eight currently updated GMST data sets each use a combination of one of five LSAT data sets and one of three SST data sets (Table 1).

Historically, SST rather than MAT has been used in global temperature products for three main reasons (Jones et al., 1999): 1) there are substantially fewer MAT observations available since ca. 1900; 2) diurnal heating biases in the MAT data that require adjustment; and 3) adjustment of MAT measurements to a common reference height to account for changes in measurement heights throughout the MAT record.

[1] Moved to end of section 1

[2] Merged old section 2 into introduction

[3] added for context as previous section title deleted

[4] R3 min1

Table 1. Global instrumental data sets reporting GMST and their land and marine data sources. LSAT data sets: CRUTEM5 (Osborn et al., 2021); GHCNm v4 (Menne et al., 2018); Berkeley Earth land record (Rohde et al., 2013a, b); C-LSAT 2.0 (Li et al., 2021); DCLSAT v1.0 (Chan et al., 2024). SST data sets: HadSST4 (Kennedy et al., 2019); ERSST v5 (Huang et al., 2017); DCSST v1.0 (Chan et al., 2024); COBE-STEMP3 (Ishii et al., 2025). All land data sources are LSAT data sets. All marine data sources are SST data sets except for GloSATMAT, used in GloSATref, which is an all-hours MAT data set.

Global data set	**Land data source**	**Marine data source**
HadCRUT5	CRUTEM5	HadSST4
GISTEMP v4	GHCNm v4	ERSST v5
NOAAGlobalTemp v6	GHCNm v4	ERSST v5
Berkeley Earth	Berkeley Earth land record	HadSST4
Kadow et al. (2020)	CRUTEM5 (via HadCRUT5)	HadSST4 (via HadCRUT5)
CMST 2.0	C-LSAT 2.0	ERSST v5
DCENT v1.0	DCLSAT v1.0	DCSST v1.0
COBE-STEMP3	COBE-LSAT3	COBE-SST3
Calvert (2024)	CRUTEM5 (via HadCRUT5)	HadSST4 (via HadCRUT5)
GloSATref	GloSATLAT	GloSATMAT (MAT)

SST observations are also affected by sampling limitations and systematic changes in measurement practices (Kennedy et al., 2019; Kent and Kennedy, 2021), but the adjustments required for MAT records were considered more complex and uncertain (Rayner et al., 2003). Furthermore, while a decline in ship coverage has resulted in fewer observations of both SST and MAT since the 1960s, in the case of SST this has been mitigated through the increasing coverage of drifting buoy and satellite data (Kent and Kennedy, 2021).

Nighttime Marine Air Temperature (NMAT) observations have previously been used to avoid the need to make diurnal heating bias corrections and are available from CLASSnmat (Cornes et al., 2020) and UAH-NMATv1 (Junod and Christy, 2020). Both ERSST v5 and HadSST4 use NMAT observations within their

SST bias adjustment methods. ERSST v5 uses NMAT-SST differences to derive a bias adjustment applied to ship SST observations prior to 2010. The HadSST4 bias adjustment model uses NMAT-SST differences to constrain estimates of contributions of different measurement types for observations made from 1850 to 1920. These NMAT data sets have not previously been combined with LSAT to create GSAT data sets.

The IPCC AR6 (Gulev et al., 2021) reviewed whether differences between GMST and GSAT reconstructions should be expected. This included a review of a range of model-based studies and process understanding and differences between instrumental NMAT and SST data sets. Differences between the two diagnostics have been the subject of recent debate due to the common (but not exclusive) use of GSAT to assess near surface temperature changes in climate models while observational records have been based on GMST. Studies investigating differences between GMST and GSAT diagnostics have included physical reasoning (Richardson, 2023) and have assessed the impact of choices of methods used to merge SST and LSAT data (Cowtan et al., 2015; Jones, 2020; Richardson et al., 2018), particularly in regions of changing sea ice cover (Cowtan et al., 2015; Richardson et al., 2018). The IPCC AR6 concluded that "There is high confidence that long-term changes in GMST and GSAT differ by at most 10% in either direction" with "low confidence in the sign of any difference in long term trends" (IPCC WG1 Cross-Chapter Box 2.3) (Gulev et al., 2021). Based on this understanding, and with no available instrumental GSAT data set, the IPCC AR6 derived estimates of GSAT change as equal to the change in observed GMST, with expanded uncertainty ranges to account for differences between the two diagnostics.

This article describes a new gridded surface air temperature (SAT) data set, the GloSAT reference analysis (GloSATref.1.0.0.0, hereafter referred to as GloSATref). The data set adds to the available set of estimates ~~The GloSAT reference analysis will extend the available ensemble of estimates~~ of global surface temperature in several ways. Firstly, the use of MAT enables an extension of the instrumental record prior to the start of systematic SST observing in the 1850s. Secondly, this earlier extension requires adjustments~~means that data adjustments are required~~ for the effects of solar heating of the measurements, with adjustments having ~~these have~~ been developed for both marine and land observations and applied in GloSAT. Thirdly, because climate models show different evolution of GMST and GSAT, improving our instrumental records of both measures will help inform ongoing discussions about resolving the disagreements between methods and datasets. Finally~~And finally~~, SST and MAT observations contain different biases so producing analyses based on both measures samples another dimension of the structural uncertainty inherent in estimates of

[5] Adapted from old end of section 1

global surface temperature. The name "reference analysis" is adopted because the data set is intended to provide a reference for comparison of observed GMST and GSAT, with processing based on the processing workflow of the HadCRUT5 GMST data set (Morice et al., 2021).

The new data set combines an extended and improved version of the CRUTEM5 LSAT data base, updated from Osborn et al. (2021), with a new MAT data set based on all-hours observations, building on Cornes et al. (2020). Section 2 describes the processing of input LSAT and MAT data sets and their use to construct the combined GloSATref analysis, following the processing workflow of Morice et al. (2021). Section 3 presents global and regional climate diagnostics derived from the merged SAT analysis. Conclusions are presented in section 4.

[6] Adapted from old end of section 1

2 Methods for producing the GloSAT dataset

2.1 Land surface air temperature data processing

The GloSAT reference analysis (GloSATref) uses an updated database of monthly average air temperature at surface meteorological stations. This is an extended version of the CRUTEM5 station database (Osborn et al., 2021), including additional station data, updated methods for climatological normal calculation (Taylor et al., 2025) and new bias adjustments for early instrumental measurements prior to the adoption of louvred Stevenson-style screens (Wallis et al., 2024). The following subsections describe the land station data contributing to the GloSATref analysis.

[7] Moved from below subsection

2.1.1 Land station data acquisitions and quality control

~~The GloSAT reference analysis (GloSATref) uses an updated database of monthly average air temperature at surface meteorological stations. This is an extended version of the CRUTEM5 station database (Osborn et al., 2021), including additional station data, updated methods for climatological normal calculation (Taylor et al., 2025) and new bias adjustments for early instrumental measurements prior to the adoption of louvred Stevenson-style screens (Wallis et al., 2024).~~

[8] Moved from below subsection

The new acquisitions and improved/updated series are summarised in Table 2. Quality control is applied to these data at source by the national meteorological services together with further quality control following the methods described in Osborn et al. (2021). These include station neighbour and station extreme threshold

checks with threshold modifications based on neighbour stations, and identification and removal of physically implausible values (taking into account month of year, station elevation and station latitude).

2.1.2 Early record exposure biases

Measurements in the early record were made in thermometer shelters of widely varying designs. Disparities between measurements made using these different shelters were already being noted in the mid-19th century (Naylor, 2018) and the process of understanding those differences, and the more general difficulties of making reliable and practical measurements across a global network of observers, led eventually to standards for the siting and design of instrument shelters such as those provided by the World Meteorological Organisation (2018, WMO). The transition from non-standard shelters to Stevenson-style screens is a known source of systematic bias – called exposure bias – because earlier shelters were subject to differences in ventilation and direct or indirect radiative heating of the thermometer. Resulting biases exhibit seasonal structure with an identifiable impact on regional trends (Parker, 1994).

Current global data sets do not explicitly account for exposure bias across their networks of station data. Some (e.g. Berkeley Earth land record or those using GHCNm v4; see Table 1) rely on their general statistical homogenisation algorithms to detect and adjust for breakpoints arising from exposure changes. These are not specifically designed for identifying exposure biases, but where those biases are detected then they may yield a similar outcome. However, the power of breakpoint detection is reduced if changes in exposure were introduced within the same timeframe across many stations within a country or region (Menne and Williams, 2009), which was often the case for the introduction of Stevenson screens (Wallis et al., 2024). Other datasets leave exposure biases mostly uncorrected and instead represent their effects via a component in their uncertainty model, developed from previous exposure bias assessments (Brohan et al., 2006; Morice et al., 2012).

~~Global data sets have partially accounted for exposure biases through their general statistical homogenisation processes and/or by inclusion of a component in their uncertainty models, developed from previous exposure bias assessments (Brohan et al., 2006; Folland et al., 2001; Menne et al., 2018; Morice et al., 2012).~~ Only a limited number of data sources have explicitly adjusted for exposure biases in early temperature records; e.g. in CRUTEM5 only some data sources for Spain (Brunet et al., 2006, 2011), the Greater Alpine Region (Böhm et al., 2009), and Australia (Ashcroft et al., 2014) have had adjustments applied. Specific exposure

[9] R1Min3

bias adjustments have not been developed and applied more widely because quantitative estimates of the bias and comprehensive metadata describing screen types in use over time at each meteorological station have been lacking.

Wallis et al. (2024) made a significant step forward in addressing these gaps by compiling temperature observations from 54 parallel measurement series at sites where multiple thermometer shelters were simultaneously in operation and using them to develop regression-based predictive models of the bias between measurements taken in Stevenson-style screens versus those in earlier shelters. These regression models were selected and calibrated using parallel measurements for four classes of shelter:

- Open exposures that, barring shielding to the top and one side, fully expose the thermometer to the air, including Glaisher and Montsouris stands. Bias predictors are annual temperature and a climatology of surface solar radiation.
- Intermediate exposures that provide additional lateral protection to the thermometer, including thermometer sheds and summer houses. This bias prediction model was not applied because its skill arose mostly from a single parallel measurement site.
- Closed exposures in which the thermometer is fully shielded on all sides, including Wild huts. Bias predictor is a climatology of surface solar radiation.
- Wall-mounted exposures, including wall-, fence-, or window-mounted measurements, both screened and unscreened. Bias predictor is top of atmosphere downward solar radiation.

Wallis et al. (2024) then compiled a database of shelter types for the majority of mid-latitude stations that had some pre-1961 data in the GloSAT station database. This metadata provides an estimate of the shelter types in use up to the time that a Stevenson-style screen was introduced and allows a prediction of the exposure bias arising from the transition from these earlier shelters to the Stevenson-style screen using the aforementioned regression models.

The bias estimates from Wallis et al. (2024) have been applied here to adjust the GloSAT station database for exposure bias. Of the 5,031 stations between 30° and 60° latitude (in either hemisphere) with some data prior to 1961, exposure bias was considered to be absent in 1,898 stations (either because they had already been adjusted or because the metadata suggests a Stevenson screen had been in place from the start of the

record). Of the remaining 3,133 stations, new exposure bias adjustments were applied to 1,960 stations. No adjustments were made to 1,173 stations either because shelter metadata were absent or because Wallis et al. (2024) had not found an acceptable bias prediction model (e.g. for intermediate exposures).

Even though we apply these new adjustments for exposure bias, we retain the exposure bias error term in the Morice et al. (2012) uncertainty model without any reduction. This provides conservative bounds on remaining exposure-related uncertainty in regional and global averages and is appropriate because residual exposure biases are still present in the GloSAT station database (from those tropical and high-latitude stations not assessed by Wallis et al. (2024) or lacking necessary metadata, or due to transitions from very early exposures that have not been considered.)~~While Wallis et al. (2024) provides a useful step forward, residual exposure biases are still present in the GloSAT station database. A sizable fraction of the station records either reside in tropical or high latitude regions that are not assessed by Wallis et al. (2024) or do not have the metadata required to apply adjustments. Additionally, transitions from very early exposures (such as thermometers placed in unheated poleward-facing rooms) have not been considered. Hence, the ensemble uncertainty model for exposure biases from Morice et al. (2012) is used here, without any reduction, to provide conservative bounds on remaining exposure-related uncertainty in regional and global averages.~~

[10] R2Com2

2.1.3 Improved and additional station normals

The GloSATref gridding process requires the monthly temperatures to be expressed as anomalies by subtracting an estimate of each station's average (hereafter "station normal") during the 1961-1990 baseline. In CRUTEM5 (Osborn et al., 2021), station normals were computed for each calendar month where at least 15 out of 30 years of data were available in the 1961-1990 period. For stations where this criterion was not met, some station normals were obtained from the WMO (World Meteorological Organisation, 1996) or were estimated by computing 1951-1970 normals which were then adjusted to represent the 1961–1990 mean based on the difference between the grid-box averages for 1961–90 and 1951–70 at nearby locations (Jones et al., 2012; Jones and Moberg, 2003). Applying this approach for the GloSAT reference analysis would have led to 2,699~~2,656~~ stations being unused due to the absence of an estimated normal and some estimated normals would have greater uncertainty and small biases (Calvert, 2024; Taylor et al., 2025) due to being estimated from incomplete data.

[11] Corrected minor error

For the GloSAT reference analysis, therefore, the CRUTEM5 approach is revised to augment the available station observations with individual monthly values estimated by the Local Expectation Kriging (LEK) method described by Taylor et al. (2025). The local expectation at a given station location is estimated as a linear weighted average of the temperature values recorded at other stations in the neighbourhood, with the weightings determined by an approximation to Kriging with hold out and taking into account the covariance between stations. The monthly values estimated by LEK are used to fill in missing values within each station time series, but only for the purpose of calculating station normals from complete 1961-1990 values. Taylor et al. (2025) evaluate these LEK-estimated normals against normals calculated directly from observed values (for those stations with adequate data during 1961-1990) and find a root-mean-squared difference of approximately 0.2 °C with no systematic dependence on latitude. This measure of similarity is representative only of stations with data close to the baseline period. For station data fragments further away (in time) from 1961-1990, the normals based on LEK are more uncertain though this is difficult to quantify; comparison with simple neighbour averages suggests that the uncertainty on those normals may be twice as large as for stations with data close to 1961-1990. The new data set applies the pre-existing HadCRUT uncertainty model for extrapolated normals to those based on LEK, assigning them a larger error than for those calculated from complete 1961-1990 data.~~evaluate these calculated normals against normals calculated directly from observed values and find a root-mean-squared difference of approximately 0.2 °C with no systematic dependence on latitude.~~

[12] RR2

For GloSATref, the normals are therefore calculated either from complete 1961-1990 observations (5,806 stations), or from a mixed set of observed and LEK-estimated values during 1961-1990 (3,568 stations, with reduced bias for cases where the observed values were either at the start or end of the baseline period), or from only LEK-estimate values (1,742 stations, typically short-segment stations during earlier or very recent periods). This permits use of station series for which few or no observations are available in the 1961-1990 period and reduces the use of less reliable WMO normals.

In addition to extending the record back to 1781, the total number of stations in the database is increased from 10,632~~10,639~~ in CRUTEM.5.0.1.0 to 11,865~~11,873~~ in the GloSAT land station database. The number of stations with normals, and thus able to be used to create the gridded data set, is increased from 7,983 stations in CRUTEM.5.0.1.0 to 11,134~~11,135~~ stations for GloSATref.1.0.0.0 (Table 3). Together with new data acquisitions, the use of LEK-derived normals increases the number of usable stations. Many of the

stations with new normal information are, however, situated in relatively well observed locations, and cover relatively short periods of time (mean station length is 35 years for those with normals fully or partially calculated from LEK estimates, versus 94 years for those with full 1961-1990 data), hence the spatial coverage of the globe is not increased commensurately with the increased number of stations with climatological normal estimates.

2.2 Marine air temperature data processing

This section describes the processing used to construct a 200-member ensemble of gridded MAT fields that forms the input to the GloSATref spatial analysis. We refer to this MAT dataset as GloSATMAT. The production of stable GSAT climate records requires assessment of the systematic biases and uncertainty in MAT observations. The most important of these are diurnal biases and measurement height changes.

Spurious d~~D~~iurnal effects are caused by solar heating of the ship which in turn warms the air around the sensor and have been addressed previously through two approaches. The first, and most common, approach is to discard all observations affected by daytime heating to produce a nighttime marine air temperature (NMAT) data set (Bottomley et al., 1990; Cornes et al., 2020; Junod and Christy, 2020; Kent et al., 2013; Rayner et al., 2003), with the definition of nighttime typically being from one hour after sunset to one hour after sunrise. Although simple to apply, this restricts the starting date of NMAT datasets to the late nineteenth century as before that time an increasing proportion of observations were taken during the daytime and commonly at local noon (Kent and Kennedy, 2021). The second approach is to model and adjust for daytime biases. This has the advantage of increasing the sample size of observations and was the approach taken by Berry and Kent (2011) who produced a global gridded all-hours air temperature product starting in the 1970s. This approach is also taken here but is used to construct a MAT dataset back to 1784. As noted by Berry et al. (2004) and Cropper et al. (2023), this adjustment approach acts on average to remove any real diurnal cycle from the MAT observations. Whilst this is not an ideal solution this is a pragmatic approach that can be assessed by the similarity between datasets based on MAT and NMAT. Adjustments from the measurement height to a standard reference height are required as the temperature typically decreases with increasing height above the ocean as the sea is usually warmer than the air above. Height adjustments have been made by estimating the lapse rate of the lower atmospheric boundary and using this information with

[13] RR4

[14] RR4

known values (or estimates) of the temperature recording height to adjust the temperature to a common reference height.

Processing of GloSATMAT largely follows Cornes et al. (2020) with the following description of the method focusing on additions and variations from that approach. The most notable changes are:

- A refined Quality Control (QC) procedure
- Development of an ensemble version of the Cornes et al. (2020) error model to produce a 200-member ensemble data set
- The inclusion of observations of marine air temperature made during the daytime, following adjustment for daytime heating bias (Cropper et al., 2023) using the method proposed by Berry et al. (2004)

A new version of CLASSnmat (version 2.1.0.2) has also been produced (Cornes et al., 2024), which uses the same input data and methodology as GloSATMAT but is constructed from nighttime-only values and hence omits the diurnal adjustments. CLASSnmat v2 only extends back to 1880 due to the reduced sampling of nighttime observations before that time.

2.2.1 Marine observations and quality control

The principal source of ship data is the International Comprehensive Ocean-Atmosphere Data Set (ICOADS, Freeman et al., 2017; Li et al., 2021) and data are processed from 1784 to present using the methods described in Cornes et al. (2020). In addition, ship data from the following sources are also included:

- Citizen science data digitisation undertaken under the Old Weather initiative (Spencer et al., 2019)
- Research Vessel data from ICOADS holdings of the Ship-based Automated Meteorological and Oceanographic System (SAMOS) archive (Smith et al., 2018)
- Voluntary Observing Ship Global Data Assembly Centre (VOS-GDAC), obtained on 19th January 2021 https://www.dwd.de/EN/ourservices/gcc/gcc.html. In cases where an observation in ICOADS has the same ID, date, and position as a VOS-GDAC observation, the information from the VOS-GDAC observation is used.

As in Kent et al. (2013) and Cornes et al. (2020), data for the period 1876–1893 for ships passing through the Suez canal were excluded due to the warm bias in those observations.

The QC checks applied to the MAT data follow those described in Cornes et al. (2020). An updated climatology-based quality control check has been developed for the MAT data, described below. For GloSAT-MAT additional checks have been made in relation to the diurnal heating bias adjustments (Cropper et al., 2023).

Climatology QC procedure

At the start of marine data processing climatological outliers are rejected based on the climatology generation step. Under the new procedure, a 1961-1990 pentad (5-day) climatology is generated on a 1° latitude by 1° longitude resolution grid. Due to the sampling frequency of ship-based observations, if a grid cell does not initially contain at least one value in each pentad throughout the year or one value in each month, the size of the grid cell, centred on the target 1° by 1° grid cell, is expanded in 1° latitude and longitude increments. This occurs until either the criterion of 500 values across at least 48 pentads and 10 months is met or there is one complete set of pentads. These criteria prevent climatologies being formed from data with preferential sampling in one part of the year. Harmonics are fitted to the pentad averages per grid-cell and an optimum number of functions is chosen using the Akaike Information Criterion up to a maximum of three harmonics. The fitted harmonic coefficients are then used to generate a daily climatology for each grid cell.

[15] R3 min2

Using the grid-cell climatology values, outliers are identified from the distribution of anomaly values per year. Values that exceed a lower or upper bound from the distribution are flagged for exclusion. These bounds, b_{lower} and b_{upper}, are defined by the 5th and 95th percentiles (p_5 and p_{95}), the 5th to 95th percentile range and an expansion factor (f_{exp}).

$$b_{\text{lower}} = p_5 - 0.5 f_{\text{exp}}(p_{95} - p_5) \quad (1)$$

$$b_{\text{upper}} = p_{95} + 0.5 f_{\text{exp}}(p_{95} - p_5) \quad (2)$$

To exclude gross outliers, an initial check is performed across each latitude band per year with $f_{\text{exp}} = 2$. Following exclusion of values that fail that check, the outlier flag is applied to observations with $f_{\text{exp}} = 0.5$ on the highest grid cell spatial resolution available (starting at 1° resolution) that contains 50 observations per month, progressively increasing the grid box size in 1° increments. If, despite that increasing box size,

the number of observations is still less than 50 then the distribution is formed from anomalies across the 10° latitude band per month.

The generation of the climatology and the quantile-based removal of values are applied as follows. A climatology is generated using data that passes the Met Office QC checks (after Cornes et al., 2020). The above quantile-based QC procedure is then applied and then a new climatology is generated using data that passed the first quantile-based QC check and a second iteration of quantile-based QC is applied.

Diurnal bias-related QC procedure

Additional QC is applied as part of the diurnal bias adjustment process (see section 2.2.4) as documented in the appendix of Cropper et al. (2023). Four modifications to the QC method described in Cropper et al. (2023) are made here:

- precipitation-flagged observations were retained for use in the analysis (but excluded from the heating bias fitting process)
- all ships with unrealistic diurnal heating were excluded, not just those in the 1854-1894 period described in Cropper et al. (2023)
- observations with an estimated heating bias $\geq$ 15 °C were also flagged and excluded from the analysis.
- for ships lacking ID information only nighttime observations are retained as IDs are required for diurnal bias adjustment

2.2.2 MAT measurement height adjustments

In HadNMAT2 (Kent et al., 2013) and CLASSnmat v1 (Cornes et al., 2020) the measurement height uncertainty is represented as a standard deviation around the estimated or known measurement height. This uncertainty is propagated through the height adjustment calculation (see following section) using a Monte Carlo approach sampling a distribution of heights and joint distribution of wind speed and air-sea temperature difference to estimate the likely uncertainty in atmospheric stability. Details can be found in (Kent et al., 2013) with updates in (Cornes et al., 2020).

Adjustment of MAT observation to a standard height above sea level

The previous approach provides only a likely distribution of measurement height and height adjustment uncertainty. Correlated and uncorrelated contributions were estimated separately, and the correlation struc-

ture was not captured. For those observations that could not be linked to a measurement height in WMO Publication 47 (Kent et al., 2007), default heights were chosen which after 1945 accounted for the regional variation in typical ship heights. However, this had the consequence that a ship with a known ID and unknown height would change its estimated height depending on which 5° grid cell it occupied at a particular time. Here, the development of an ensemble of heights permits correlated uncertainty to be handled as follows:

- In each ensemble member a ship with known ID will keep the same measurement height, at least within a calendar year
- In each ensemble member the stability estimate will be the same for nearby ships within the same 5° grid cell and 10-day period.

Prior to height information being available for individual observations from WMO Publication 47 (mostly pre-1973 before callsigns are available to assign WMO Publication 47 entries to ships in ICOADS), observation heights are estimated from the literature following Kent et al. (2013). A 200-member ensemble is created to reflect uncertainty in measurement heights; the parameters for the ensemble are shown in Table 4. Per ship, the ensemble varies the height four times to reflect the change in average ship height over time (parameter prefix "Height" in Table 4). The exact timing of the start of the height changes also forms a component of the ensemble to reflect uncertainty in that date (parameter prefix "Year" in Table 4).

From 1970 onwards, the height of a ship is determined by joining the ICOADS ID/callsign with the corresponding WMO Publication 47 entry. Sometimes, the height of the dry bulb thermometer above the summer load line i.e. thermometer height, is used. However, if the thermometer height is missing, then the height of the barometer, anemometer or the height of the visual observing height is used. If the height of the ship's thermometer (inferred from WMO Publication 47) is known, the assumed uncertainty in that height is ± 1 m.

For missing heights after 1970, the height is sampled from the distribution of known heights from WMO Publication 47. Height information is grouped by 5° grid box, 10° grid box, country of registration, oceanic basin, 10° latitude band, vessel length and vessel type. For IDs without a height, the height is sampled across a selection of heights where a ship has matching information. Observations without ship ID information are

assigned a pseudo-ID and placed in a sub-group for which an ensemble of possible heights is generated. The overall estimated thermometer height for ships in ICOADS is shown in Fig. 2 of Cornes et al. (2020).

To adjust MAT measurements to a standard reference height, and quantify uncertainty in adjustments, we implement an ensemble version of the height adjustment method described in Cornes et al. (2020). For each ensemble member, a single random sample of stability parameters (temperature scaling parameter and Monin-Obukov length, Biri et al., 2023) is taken in each 5° grid cell and 10-day period from a pool of 10,000 parameter sets, which comprises 5,000 samples taken from the distribution of random uncertainty components and 5,000 taken from the systematic components (see Cornes et al., 2020). A fixed sample over each 10-day period within a month is used. This is designed to replicate the expected constancy of stability parameters within the synoptic timescale. For each sample of stability parameters and height estimates, the temperature data were adjusted to a height of 2 m above sea-level.

2.2.3 Adjustment of observations during World War 2

The adjustment of data from ICOADS Decks 245 (UK Royal Navy Ships) and 195 (U.S. Navy Ship Logs) during the period 1942-1945 follows the method described in Cornes et al. (2020) to calibrate observations from these decks to those in other decks. For GloSATMAT the adjustments and uncertainty values have been calculated for both daytime and nighttime data. While the uncertainty in these adjustments was previously considered to be entirely correlated by ship track, the World War 2 adjustments have been partitioned so that 1/3 of the total uncertainty is modelled as correlated by ship track and 2/3 is uncorrelated. This partitioning addresses a problem with the error covariance matrices, which would otherwise have not always been positive semidefinite and hence could not have been used in further calculations.

2.2.4 Diurnal heating bias adjustment

The Cropper et al. (2023) implementation of the Berry et al. (2004) heating bias model was used to quantify the expected diurnal influence on MAT of energy storage and release by ship superstructures. On an annual basis, we fit a set of heating bias model coefficients to every uniquely identifiable ship track used from ICOADS following processing to improve the association of ship identifiers with individual ships as described in Cornes et al. (2020). In addition, some ship IDs prior to 1850 were not unique and were assigned new unique IDs using additional information contained within the ship report. Where this was possible the

data for these problematic IDs were retained, otherwise the data were excluded. The estimates of the heating bias are based on the difference from MAT and the underlying trend in nighttime temperature, defined as temperature between one hour after sunset until one hour after sunrise. Hence, each MAT observation is adjusted by the difference between the estimated bias and the nighttime mean MAT. As such, the full diurnal cycle is removed from the data and GloSATMAT should be considered a nighttime-equivalent dataset.

As described in Cropper et al. (2023), 2500 alternate sets of heating bias model coefficients are determined from the full suite of ICOADS ships from 1856-2020. For each individual ship, 2500 realisations of the heating model coefficients are generated, and an ensemble of 60 of these, minimising several different cost functions, are selected as the best fit for each ship. The ensemble means of the adjustments over the 60 different ensemble members becomes the heating bias adjustment.

The data requirements for each observation to apply the heating bias adjustment are position, time, cloud cover and relative wind speed. For a ship track to be included, we require at least 12 observations where the underlying nighttime MAT can be determined. If a ship track has partially missing cloud and/or wind speed values, these values are sampled from the 1961-1990 climatological distribution of cloud and wind derived from ICOADS after the 60 best model coefficient combinations have been selected. If a ship track has completely missing cloud and/or wind, the sampling occurs during the selection of the heating bias model coefficients.

Almost all observations pre-1856 lack cloud cover data and many ships do not meet the criterion of having 12 observations for which the underlying nighttime MAT can be determined. In this period, we relax this requirement so that all observations in this period that pass QC can be included. For many ships this means that the coefficients for the heating bias model cannot be fitted as there is no target nighttime information. In this situation we use an ensemble of heating bias model coefficients computed from ships that do have the required data, where it is likely that these ships are typical during this period. The correlated uncertainty for MAT from ships fit this way is fixed at 0.45°C, the 97.5th quantile over the 1857 - 1870 period.

2.3 The GloSAT GSAT reference analysis

2.3.1 LSAT Anomaly Grids

A 200-member ensemble of gridded LSAT fields has been constructed from station temperature series by applying the ensemble gridding procedure described in Morice et al. (2012). This includes ensemble sampling for homogenization uncertainty, exposure bias uncertainty, station climatology uncertainty and urbanization uncertainty, which are accompanied by analytical estimates of measurement and sampling uncertainties as in Morice et al. (2012). The exposure bias uncertainty model of Morice et al. (2012) is retained despite the addition of new exposure bias adjustments as the Wallis et al. (2024) adjustments are available for a subset of stations located at mid-latitudes only and the adjustments are uncertain. The HadCRUT5 climatology uncertainty model is retained (Morice et al., 2012), with the HadCRUT5 uncertainty model for interpolated normals used for LEK normals (Taylor et al., 2025). For these interpolated normals, the station climatology uncertainty, $\sigma_{\mathrm{clim}}(m)$, for month calendar month m is a function of the standard deviation of monthly temperatures for the calendar month, $\sigma_{var}(m)$, with $\sigma_{\mathrm{clim}}(m) = \sigma_{var}(m)/\sqrt{15}$. The resulting uncertainty is equivalent to that for a station with 15 years of available monthly averages within the climatology period.

[16] RR comment 3

Table 5 shows the mapping of uncertainty sources~~error model terms~~ to the uncertainty model's ensemble members and error covariance matrices. These ensemble members and error covariance matrices form the inputs for a non-interpolated merged analysis and the resulting 200-member LSAT anomaly ensemble grids are then used as inputs to the Gaussian process based infilling method described in Morice et al. (2021) and previously used to create the HadCRUT5 dataset (see section 2.3.3).

2.3.2 MAT Anomaly Grids

A 200-member marine air temperature ensemble has been generated based on the 200-member ensemble of diurnally adjusted all hours marine air temperature observations adjusted to a 2 m reference height. Each ensemble member is gridded separately after the method described in Cornes et al. (2020) using the ship observation ensembles that sample uncertainty in height adjustments and World War 2 (WW2) adjustments. As in CLASSnmat v1, these gridded fields are initially generated as monthly actual MAT values and then climatological averages are subtracted from the gridded data to produce the anomaly fields. Missing values in the GloSATMAT climatology fields are filled using a thin-plate spline with the smoothing parameter set

to zero. In this way, the spline acts as an exact interpolator, and hence where grid-cell values exist these are reproduced exactly in the interpolated field. Missing cells are interpolated from non-missing cells in the neighbourhood, and this allows anomaly values to be calculated in regions where there are insufficient values over the climatological period to construct a 30-year mean value. The gridded anomaly values are calculated by subtracting this climatology from the gridded actual temperature fields for the respective month of the year. As a final QC check on the gridded data, any grid-cell MAT anomalies greater than 10 °C or less than -10 °C are removed. It should be noted that these climatologies differ slightly from the values used in the earlier climatological QC (see section 2.2.1)~~of the marine data due to that being a different stage of the processing chain~~.

[17] R1 min5

Following the example of the CLASSnmat dataset, uncorrelated, systematic, and sampling uncertainties are encoded into monthly error covariance matrices, excluding terms included in the ensemble. Uncertainty in diurnal heating adjustments is encoded into these error covariance matrices and treated as uncorrelated between observations. A summary of uncertainty model components is provided in Table 5. The 200-member gridded MAT ensemble and accompanying error covariance matrices are provided as inputs into the Gaussian process based infilling method (previously used to create the HadCRUT5 dataset, Morice et al., 2021) to generate a 200-member infilled MAT anomaly ensemble.

2.3.3 LSAT and MAT interpolated analyses

Interpolated analyses of LSAT and MAT are produced using the Gaussian process analysis system of Morice et al. (2012). This analysis method takes ensemble gridded temperature anomaly fields as inputs, modelling systematic error structures in observed air temperature anomalies. Additional observational error structure is provided to the analysis in error covariance matrices. Full details of the method are provided in Morice et al. (2021).

Separate LSAT and MAT analyses are produced. Ensemble and error covariance inputs to the analysis equations are summarised in Table 5. The LSAT analysis error model structure matches that used in the HadCRUT5 data set, with the structure of individual uncertainty components described in Morice et al. (2021). The structure of terms in the MAT, encoding error structure into the 200-member MAT analysis ensemble and accompanying error covariance matrices, are described in Section 2.2 and in Cornes et al. (2020) and Cropper et al. (2023).

Analysis estimation follows the method of Morice et al. (2021) using the HadCRUT5 analysis processing system. As in Morice et al. (2021), the Gaussian process models monthly temperature anomaly fields as the sum of a monthly field mean and a spatially varying Gaussian process. The Gaussian process model uses a Matérn covariance function that models covariances between locations on the Earth's surface as a function of Euclidian distance. As for HadCRUT5, the Matérn covariance function's smoothing parameter is set to $\eta = 1.5$. Parameters representing the standard deviation of temperature anomaly variability σ and spatial decorrelation length scales ρ are estimated separately for land and ocean analyses as the average of maximum likelihood estimates for monthly fields from 1961 to 1990. The resulting parameter estimates for the LSAT analysis are $\sigma = 1.2$ ˚C, $\rho = 1300$ km and for the MAT analysis are $\sigma = 0.65$ ˚C, $\rho = 1550$ km. For comparison, these LSAT parameters estimates for GloSATref are the same as for HadCRUT5, but the GloSATref MAT parameter estimates differ from those for SST in HadCRUT5 which have a lower amplitude ($\sigma = 0.6$ ˚C) and shorter length scale ($\rho = 1300$ km).

As in the HadCRUT5 data set, the analyses are masked in regions of weak observational constraint, defined by a metric of one minus the ratio of posterior to prior variance of the spatial model estimates at each analysis grid cell. The analysis is masked where this metric takes a value less than $\alpha = 0.25$ (see Morice et al. (2021) and its supporting information for discussion).

2.3.4 Merging LSAT and MAT analyses

The GSAT ensemble grids are produced as weighted averages of MAT and LSAT anomaly ensemble grids. The weighting scheme follows the HadCRUT5 method, with weighting based on the fraction of land area in each grid cell. As in HadCRUT5, for the interpolated analysis, sea ice regions are treated as if they were land in the weighting scheme and a minimum weighting of 25% land is placed on grid cells that are directly observed by land meteorological stations. These choices are retained from HadCRUT5 to aid comparison of the climate diagnostics based on SST and MAT through comparison of HadCRUT5 and the GloSAT reference analysis.

3 Results and Discussion

3.1 Global and hemispheric temperature anomaly series

Figure 1 shows a comparison between GloSATref and HadCRUT5 interpolated global and hemispheric analyses including uncertainty ranges and percentages of global and hemispheric coverage. Each of these two analyses is produced using the HadCRUT5 analysis methodology. Differences in global annual average time series between these two analyses reflect the differences in input LSAT and MAT/SST data sets and their uncertainties combined with the interaction between anomaly patterns and data coverage in those input data sets and the spatial interpolation methods.

GloSATref, like HadCRUT5, represents the major changes we expect to see in the global mean. The pre-1850 record, despite its increased uncertainty, captures the strong cooling associated with major volcanic eruptions. The overall warming of global temperature is clear, although GloSATref warms slightly less overall. Features noted in previous comparisons of SST and NMAT large-scale averages remain prominent in the GloSATref comparison to HadCRUT5, for example the lower temperatures in the early 1900s and in recent years (Cornes et al., 2020; Gulev et al., 2021; Chan et al., 2024).

For their period of overlap, currently 1850 to 2021, there is broad agreement in the global and hemispheric means. While areal coverage of the northern hemisphere is similar in the two data sets, GloSATref has slightly reduced southern hemispheric coverage compared to HadCRUT5. GloSATref is on average warmer than HadCRUT5 from the start of the latter dataset in 1850 until 1880 and then more notably cooler on average than HadCRUT5 until the early 1910s, especially in the Southern Hemisphere. **[18]** R3 min4

Variability in GSAT and hemispheric averages from GloSATref is high in the period before the start of HadCRUT5 in 1850, as expected from the lower data coverage. There are two periods of strong cold anomalies which result from the cooling of the atmosphere following the eruptions of major volcanos: Laki in 1783-1784 (Zambri et al., 2019), an eruption of unknown source in 1808, Tambora in 1815, and further eruptions in the 1820s and 1830s (Brönnimann et al., 2019). It is likely that the global mean effect of Laki is over-estimated in this time series (Figure 1a) because it is influenced by the location of the available observations in the 1780s; this effect can be seen in model simulations masked to the GloSATref data coverage (Ballinger et al., submitted).

3.2 Temperature anomaly maps and data coverage

The most obvious feature in the 20-year averages of SAT anomalies shown in Fig. 2 is the increase in temperature over time, particularly at high northern latitudes. The period before about 1820 is particularly cold, likely due to volcanic activity. Warm anomalies are present in higher latitudes of the North Atlantic in the 20-year averages from 1821-1840 to 1881-1900~~in the 20-year averages from 1821-1840 to 1900~~ and the South Atlantic from 1821-1840 to 1861-1880. These might be indicative of multidecadal variability. The Arctic region is relatively warm in the 1920s and 1930s (Hegerl et al., 2018). Despite the new adjustments to MAT for biases present during WW2, this period may still be unrealistically warm in GloSATref according to a new analysis of SST data (Chan et al., 2024).

[19] "R1 min6"

While use of MAT allows GloSATref to extend prior to 1850, there is nevertheless a marked reduction in data availability prior to about 1855 (see global coverage metrics in Fig. 1 and 20-yearly coverage maps in Supplement Fig. S1 and Fig. S3). Before 1850, the majority of weather station LSAT series are situated in Europe and the eastern coast of North America, limiting northern hemisphere coverage of the GloSATref analysis fields. Combined with an absence of station records in the southern hemisphere, this contributes to increased uncertainty in global and hemispheric series. MAT coverage is predominantly restricted to the Atlantic and Indian Oceans, reflecting primary trade routes. For these regions it is possible to make estimates of regional temperature anomalies based on GloSATref analysis fields using reasonable criteria for data availability (Fig. 2). The Pacific has highly limited sampling in this early period. Non-uniform global coverage in the early record may lead to a sampling bias in global average SAT estimates in this period. Uncertainties in global and hemispheric means are therefore larger in this early period as both LSAT and MAT are sparsely observed, but the uncertainty estimates themselves are likely less robustly quantified~~also likely to be more uncertain~~.

[20] R1 min7

Observation coverage improves over time accompanied by increased global coverage of the gridded analysis fields. In the second half of the 1800s there is a marked increase in station LSAT series availability, with analysis estimates available by the end of the century for most land regions excluding Antarctica, interior regions of Africa, and northern regions of South America including the Amazon. By the 1880s measurements from all oceans are available, although with varying degrees of data availability. While sufficient data is available to estimate average temperature anomalies across much of the Pacific at this time (Fig. 2), data

[21] R1 min8

coverage in the analysis is limited in regions of the western and southern Pacific. Data coverage for the Southern Ocean and high northern latitudes is highly limited.

In the early 20th century most of the missing grid cells are in the Southern Ocean and over Antarctica. Underpinning land station data coverage remains limited for much of Africa, South America, interior regions of Eastern Asia and high latitude regions of North America and Eurasia, although the available observation coverage permits analysis estimates for these regions. The Antarctic becomes represented in the analysis in the 1950s. Coverage for MAT peaks in the 1970s and 1980s and has declined markedly since (Berry and Kent, 2016; Kent and Kennedy, 2021). Observation coverage of the Southern Ocean and southern extents of the Pacific, Atlantic and Indian Ocean remains reduced in GloSATref in comparison to HadCRUT5 in recent decades (see Supplement Fig. S1-S4). SST observations from drifting buoys contribute notably to differences in marine data coverage in these regions but there are also fewer ship observations of MAT than of SST.

As in the preceding CRUTEM5 station database (see discussion in Osborn et al., 2021), the number of LSAT series is relatively high and stable from the early 1960s to the early 2010s, with a sharp reduction in the last decade due in part to delays in access to data. However, in CRUTEM5, the number of station series actually used in the gridding showed a decline from the 1970s onwards, falling to below 90% of its 1970s peak from 1991 onwards (figure 7 of Osborn et al., 2021) and this is mostly because new stations installed since the 1970s did not have the data necessary to estimate their 1961-1990 normals. In the current study, the use of kriging (section 2.1.3) to estimate missing values from neighbouring stations for the purpose of estimating normals has partly ameliorated this issue: the number of stations used for gridding is 90% of its 1970s peak as late as 2011 (compared with 1991 in CRUTEM5). It is worth noting, though, that these additional stations that can now be used for the gridded dataset are commonly in grid cells with other LSAT stations; therefore, their importance is that they reduce the sampling uncertainty at the grid cell level rather than extending coverage of the gridded dataset in the final decades.

Differences in temperature anomalies over land between GloSATref and HadCRUT5 (Fig. 3) result from the increased number of stations used to create the grids, primarily from use of the LEK method, and from early-record mid-latitude screen bias adjustments. The effects of screen bias adjustments are evident in anomaly difference maps prior to the 1930s, most notably with GloSATref containing cooler anomalies across Eurasia in the 1890-1909 and 1910-1929 panels and in western North America in the 1850-1869

panel. This difference is largest in summer and autumn months when northern hemisphere screen biases are largest, for which HadCRUT5 does not include adjustments. From the 1930s onwards, differences between GloSATref and HadCRUT5 anomalies over land are minimal.

3.3 Effects of MAT and LSAT updates in global series

Global average anomaly time series for GloSATref, GloSATLAT and GloSATMAT are shown in Fig. 4 in comparison to methodologically related data sets, showing the effects of data and methodological changes contributing to GloSATref. Additional land and marine data set comparisons are shown in Supplement Fig. S7 and S8 including a broader range of data sets.

[22] Zeke Hausfather comment

Averaging globally over the land domain for the non-interpolated GloSATLAT data (noting that coverage becomes increasingly limited during the first 100 years) and comparing to the closely related CRUTEM5 data set (Fig.~~Figure~~ 4a,d) shows that the Wallis et al. (2024) exposure bias adjustments cool LSAT global annual averages in the 19th and early 20th centuries by less than 0.1 °C, primarily through mitigation of summer warm biases which are larger than the annual bias adjustments. Differences from CRUTEM5 in the 21st century are less than 0.04 °C of either sign. These differences result from the expanded set of station records included in anomaly grids.

Comparing the all-hours GloSATMAT to related NMAT data sets, CLASSnmat v2 uses the same methodology and input data sources as GloSATMAT but, like its predecessor CLASSnmat v1, excludes daytime observations and starts in 1880. The main reason for differences between the two versions of CLASSnmat is the addition of extra data sources (see Sect. 2.2.1, Fig 4b,e). CLASSnmat v2 is therefore most similar to GloSATMAT. For most of the period 1850-1879 where there are only estimates from GloSATMAT and HadSST4, GloSATMAT is warmer than HadSST4 on average by about 0.1˚C, which may suggest either an incomplete removal of diurnal heating biases in GloSATMAT or a residual cold bias in HadSST4. This relative warmth compared to HadSST4 is also seen for the first decade of the NMAT records. The early 20th century shows rapid variations in the difference between the MAT datasets and HadSST4. Recent analysis has suggested that most SST data products, including HadSST4, may be biased cold during the early 20th century (Sippel et al., 2024; Chan et al., 2024); further work is required to understand the different temperature records in this period. During WW2, GloSATMAT shows cold temperature anomalies relative to both HadSST4 and the NMAT-based estimates. This may suggest that the new bias adjustments are an improve-

ment (Sect. 2.2.3) as a new analysis by Chan et al. (2024) indicates HadSST4 may contain residual warm biases during WW2.

The smaller warming trend in NMAT relative to SST data has been noted for some time (for an overview see Cornes et al., 2021; Gulev et al., 2021). This feature is also apparent in the MAT data presented here (Fig. 4) and is shown to start in the late 1950s. SST and air temperature anomaly differences in this period are larger in tropical regions but are more similar during El Nino periods (not shown). Climate models and atmospheric reanalyses show the opposite relationship between SST and MAT trends (Gulev et al., 2021) supported by a proposed physical constraint (Richardson, 2023). Further analysis of the in situ marine observations is therefore required.

[23] R3 min5

Global average temperature series for GloSATref and HadCRUT5 analyses are shown in Fig. 4c (as Fig. 1 but without the uncertainty range), and their difference in Fig. 4f. Over their common period, comparison with averages derived from their underpinning single domain data sets indicates that differences in time variation of global temperature anomaly series for GloSATref and HadCRUT5 primarily arise from their marine components, with early record LSAT exposure bias adjustments having a smaller effect.

3.4 Comparison to other GMST data sets

The global average temperature time series for the GloSAT GSAT analysis is shown in Fig. 5 alongside those of current GMST analyses. Fig. 5b gives a more detailed view of their differences from GloSATref. As noted in Section 1, the GMST data sets differ in methods used and underpinning observation data although there are overlaps and commonalities between them in both respects.

Also shown is the global average temperature anomaly time series for the median of the PAGES2k multi-proxy ensemble reconstructions (PAGES2k, 2019). The PAGES2k ensemble was calibrated using the Cowtan and Way (2014) spatially infilled analysis of HadCRUT4 data over 1850–2000, so that (by weighting and scaling the proxy records) it resembles the overall warming and some of the variability of the infilled HadCRUT4. The pre-1850 period is not directly constrained in this way and it is for comparison with GloSATref during this period that it has been included.

There is general agreement between all the estimates, with particularly good agreement on the representation of interannual variability among the instrumental-based datasets. From 1850 to around 1885, there is a wider scatter between the various datasets, with some warmer and some cooler than GloSATref reflecting

the greater uncertainty in these early observational datasets. This is also a period when these other datasets lie outside the GloSATref 95% confidence range more often~~, especially the recently-published DCENT data set which has new treatments of inhomogeneities (Chan et al., 2024)~~.

From 1888 to 1913, GloSATref is consistently cooler than the SST-based datasets and they all lie outside the GloSATref 95% confidence range at times during this period except for the HadCRUT5-based Calvert (2024) dataset. This difference is notable because there is evidence that this early 20th century cool period is biased cold in the SST datasets (Sippel et al., 2024), yet GloSATref is cooler still by 0.1 to 0.2 °C. The SST-based datasets are more similar to each other than to GloSATref due to differences between SST and MAT (see Fig. 4e). Differences over 1888 to 1913 (Fig. 5) are most prominent for the DCENT and COBE-STEMP3 data sets (Chan et al., 2024; Ishii et al., 2025), with each notably warmer than GloSATref over 1888 to 1913 and exhibiting excursions beyond the GloSATref 95% confidence intervals throughout the first half of the 20th century. These two recently published data sets have new treatments of inhomogeneities. Each uses land station LAT in their respective SST bias adjustment schemes. While the specifics of their adjustment methods differ, each method acts to reduce differences between detrended global average SST and LSAT anomalies. These two data sets also add SST data-source-based adjustments that impact early 20th century warming: COBE-STEMP3 includes an adjustment to address a truncation error in a prominent underpinning data source in this period, the KOBE collection (Chan et al., 2019), while DCENT uses an SST adjustment scheme based on nation, ICOADS deck and measurement method (Chan and Huybers, 2021). The PAGES2k ensemble median palaeoclimate reconstruction (PAGES2k, 2019) is warmer than GloSATref and much closer to DCENT and COBE-STEMP3 during this early 20th century period.

From the 1930s to the end of the 20th century, differences between GloSATref and GMST data sets based on ERSST v5 (GISTEMP and NOAAGlobalTemp) are on average smaller than differences with HadSST4 based data sets (HadCRUT5 and Berkeley Earth). This is likely related to the more direct use of NMAT observations to bias adjust ERSST v5 SSTs than HadSST4 SSTs. From the 1950s onwards the dominant difference is the slightly lower trend in MAT relative to SST noted earlier. This leads to GMST data sets lying above the GloSATref confidence range in recent years.

Before 1850, the only comparison is with data from the palaeoclimate record. During this period, GloSATref shows much greater variability than it does after 1850 (arising especially from the land component: Fig. 4) and much greater variability than the PAGES2k median global temperature reconstruction (Fig. 5a), with

the PAGES2k median extending outside both 2.5% and 97.5% confidence limits of GloSATref at times (Fig. 5b and Supplement Fig. S6). There is a possible step in the PAGES2k-GloSATref difference around 1820 (Fig. 5b) but this is the type of feature that occurs when differencing a highly variable series with a smoother lower variance series, so it is not strong evidence for a discontinuity in either GloSATref or PAGES2k. There are multiple reasons for the differences between GloSATref and PAGES2k prior to 1850. The variance of the PAGES2k median may be too small, due to averaging over multiple possible realisations (PAGES2k ensemble range is not shown here, but is shown in Supplement Fig. S6) as discussed by PAGES2k (2019) and Anchukaitis and Smerdon (2022). It is likely that pre-1850 variability in the GloSATref global series is overestimated by much reduced global measurement sampling (Fig. 2), which increases uncertainty and variance in the global-mean estimate. This is especially the case for the land data, which is limited almost entirely to Europe before 1820. Coverage uncertainty estimates for the GloSAT series (based on sub-sampling late 20th century and early 21st century ERA5 reanalysis fields to the observed locations) is indeed larger during this period but may not fully represent this uncertainty. The underlying measurements in the early 19th century were taken from less standardised instruments and measurement platforms and will likely have larger uncorrected errors which contribute to increased variance. Part of the enhanced variability is likely due to strong volcanic activity, with the GloSATref series cooling around the times of the unidentified 1809 volcanic eruption and the 1815 eruption of Mount Tambora, and to a lesser degree around the time of the 1830s eruptions. It is possible that this volcanic response is not fully captured in the PAGES2k median (Anchukaitis and Smerdon, 2022). ~~HadCRUT5 and especially GloSATref (which provides the coolest estimates at that time) are both cooler than the PAGES2k ensemble median during the early 20th century period, in contrast to Chan et al. (2024).~~

[24] R3 maj1

[25] R3 maj1

[26] moved two paragraphs earlier where this period is now discussed at greater length and rephrased to fit

4 Conclusions

A new global surface air temperature (GSAT) anomaly dataset has been produced, GloSATref. This is the first such dataset to use marine air temperature (MAT) instead of sea surface temperature (SST) in combination with land surface air temperature (LSAT). Use of MAT has allowed extension of the global measurement-based record back to 1784, with a land-only analysis back to 1781, adding nearly 70 years when compared to datasets using SST as their marine component which start in 1850 or later. The construction of a MAT-

based dataset required several challenges to be overcome, most notably the derivation and application of adjustments to account for warm biases due to daytime heating of the ship and sensor environment (Cropper et al., 2023). Similarly, for air temperature observations made over land, new estimates of bias were required to be able to use observations made before the wide-spread adoption of Stevenson-style screens to shelter the thermometers from the direct or indirect effect of solar radiation (Wallis et al., 2024). These adjustments will also improve the existing LSAT record (Osborn et al., 2021) before about 1930.

Extending the instrumental near surface temperature record earlier in time by nearly 70 years gives an estimate of temperature changes associated with a period of strong volcanic activity between the 1780s and the 1820s, albeit requiring consideration of limitations in early global data coverage~~albeit without full global coverage~~. These data are a new resource for study of the climate of this early instrumental period (e.g. as in Ballinger et al. (submitted)), and provide an additional line of evidence for understanding global near surface temperature change and variability alongside existing GMST data sets.

[27] R3 Maj 2

Instrumental monitoring of GSAT is contingent on sustained observation of air temperature over ocean and land. At present, the marine air temperature observing network is less robust than that for SST, with numbers of MAT observations having declined since the 1990s (Kent and Kennedy, 2021). Similarly, the LSAT record is dependent on maintenance of meteorological station networks and international data exchange. Support for international data infrastructure to meet the requirements of the construction of long term climate records remains essential (Folland et al., 2001; Kent et al., 2019; Li et al., 2021). Equally important is support for data rescue through imaging and digitisation of observations to improve global sampling and of metadata to improve understanding of observing methods throughout the record (Brohan et al., 2009; Brönnimann et al., 2018; Luterbacher et al., 2024).

Global datasets based on SST have been produced for more than 30 years and have improved as understanding of the characteristics of the observations become better understood and methods of data set construction advance. Similar improvements can be expected with records based on MAT if resources permit. This new global surface air temperature analysis provides an additional line of evidence of changes in global temperature alongside existing GMST data sets and reanalyses.

Data availability. The GloSATref analysis and its constituent datasets are available via the CEDA archive:

[28] Added CEDA archive locations for GloSAT data sets

- The GloSATref analysis: https://catalogue.ceda.ac.uk/uuid/a2519624a593402a83246bd359d098be/~~https://www.metoffice.gov.uk/hadobs/glosatref/~~;
- The GloSATLAT data set and station files: https://catalogue.ceda.ac.uk/uuid/ef237f578329487eb02fb42f9db56bb2/ ~~https://www.metoffice.gov.uk/hadobs/glosatref/~~;
- The GloSATMAT data set: https://catalogue.ceda.ac.uk/uuid/e6251bf935304cfbb9c9269dc7757a35/;

Data used in comparison plots

- CRUTEM.5.0.2.0 (Osborn et al., 2021): https://www.metoffice.gov.uk/hadobs/crutem5/. Accessed 2024-01-10.
- CLASSnmat v1 (Cornes et al., 2020): https://catalogue.ceda.ac.uk/uuid/5bbf48b128bd488dbb10a56111feb36a/ Accessed 2023-06-13.
- CLASSnmat v2 (Cornes et al., 2024): https://catalogue.ceda.ac.uk/uuid/306246329ae04eb3b2299446d911530a/ ~~https://doi.org/10.6084/m9.figshare.27015661~~. Accessed 2024-09-27.
- HadCRUT.5.0.2.0 (Morice et al., 2021): https://www.metoffice.gov.uk/hadobs/hadcrut5/. Accessed 2024-01-10.
- HadSST.4.0.1.0 (Kennedy et al., 2019): https://www.metoffice.gov.uk/hadobs/hadsst4/. Accessed 2024-05-21.
- ERSST v6 (Huang et al., 2025): https://www.ncei.noaa.gov/pub/data/cmb/ersst/v5/2023.ersst.v6. Accessed 2025-05-02.
- NOAAGlobalTemp v6 (Huang et al., 2022): Huang, B., X. Yin, M. J. Menne, R. Vose, and H. Zhang, NOAA Global Surface Temperature Dataset (NOAAGlobalTemp), Version 6.0.0 subset used: aravg.ann.land_ocean.90S.90N.v6.0.0.202407.asc. NOAA National Centers for Environmental Information. https://doi.org/10.25921/rzxg-p717. Accessed 2024-09-10.
- GISTEMP v4 (Lenssen et al., 2019): https://data.giss.nasa.gov/gistemp/. Accessed 2024-09-10.
- Berkley Earth (Rohde et al., 2013a, b; Rohde and Hausfather, 2020): https://www.berkeleyearth.org. Accessed 2024-09-10.

[29] DOI not yet functional: https://dx.doi.org/10.5285/a

[30] DOI not yet functional: https://dx.doi.org/10.5285/e

[31] DOI not yet functional: https://dx.doi.org/10.5285/e

[32] DOI not yet functional: https://dx.doi.org/10.5285/3

- Kadow Version 6. (Kadow et al., 2020): https://doi.org/10.5281/zenodo.11262704. Accessed 2024-03-01.
- China MST2.0: China MST2.0-Imax (Sun et al., 2022): download from https://figshare.com/articles/dataset/China-MST2_0_datasets/16929427/4. Data set accessed 2024-09-10.
- Calvert (Calvert, 2024): https://doi.org/10.26050/WDCC/HadCRU_MLE_v1.2.Accessed 2024-09-10.
- DCENT, DCSST, DCLAT v1.0 (Chan et al., 2024): https://doi.org/10.7910/DVN/NU4UGW, Accessed 2024-09-10.
- PAGES2k (PAGES2k, 2019): https://doi.org/10.25921/tkxp-vn12. Accessed 2023-04-14.
- COBE-STEMP3, COBE-LSAT3, COBE-LSAT3 (Ishii et al., 2025): https://climate.mri-jma.go.jp/pub/archives/Ishii-et-al_COBE-SST3

Author contributions. CPM wrote the first draft of the paper with contributions from ECK, RCC, NR, JJK, TJO, EH and APS. The processing and analysis of the land observations was performed by MT, EJW, KC and TJO, that for the marine observations by TC, RCC, DIB, BRR, PRT and ECK. CPM, JW, JJK and NR constructed the combined gridded analysis. All authors reviewed the final text.

Competing interests. None

Disclaimer. None

Acknowledgements. We thank David Lister and Phil Jones (Climatic Research Unit, UEA) for updates and acquisitions to the CRUTEM station database.

Financial Support. Authors were supported by the Global Surface Air Temperature (GloSAT) Natural Environmental Research Council (NERC) project under grants NE/S015647/2 (CPM DIB, RCC, TC, JJK, NR,

JW, ECK), NE/S015582/1 (TJO, MT, EJW), NE/S015566/1 (KC), NE/S015698/1b (APS) and NE/S015574/1 (EH, PT). In addition CPM, JJK, NR and JW were supported by the Met Office Hadley Centre Climate Programme funded by DSIT, BRR was supported by NERC Grant NE/R015953/1 and EH was supported by the National Centre for Atmospheric Science (NCAS).

Table 2. Summary of new acquisitions and updates applied to CRUTEM.5.0.1.0 to create the GloSAT land surface air temperature station database (sdb). Place and country names are as given by the data source.

Region	No. series	Source	Details
Routine updates			
Global	7810	CLIMAT	Updated series for 2020-2021
Australia	112	BoM	Updated ACORN series for 2019-2021
Canada	434	Environment Canada	Updated series and improved homogeneity
Chile	314	Chilean Centre for Climate and Resilience Research	Updated series for 2016-2021
Denmark, Faroe Islands, and Greenland	17	Danish Meteorological Institute	Updated series for 2017-2020
Iceland	127	Icelandic Meteorological Office and Trausti Jónsson	Updated series and improved homogeneity
New Zealand	7	NIWA	Updated series for 2018-2021
Russia	604	GHCN-Daily	Updated series for 2018-2021
Switzerland	13	MeteoSwiss	Updated existing series
USA	1218	USHCN	Updated series and improved homogeneity
New acquisitions			
Global	439	NCEI World Weather Records (WWR)	New series not previously in CRUTEM sdb
Global	149	NCEI Monthly Climatological Data for the World (MCDW)	New series not previously in CRUTEM sdb
Canada	346	Environment Canada	New, homogenised series not previously in CRUTEM sdb
Germany	14	DWD	New series not previously in CRUTEM sdb
Improved series (e.g. earlier extensions, gaps filled, improved homogeneity)			
Global	2618	NCEI World Weather Records (WWR)	Gaps filled, especially for 2011-2016
China	322	CMA	Replacements with improved homogeneity
France	49	MeteoFrance	Mostly post-1950 additions to existing series
Germany	68	DWD	Replacements with improved completeness
Switzerland	14	MeteoSwiss	Replacements with improved homogeneity
Global	15	Multiple sources (papers, archives)	Jersey (1894-2019), Dublin (Ireland 1831-2021), Reading (UK 1908-2019), Perpignan (France 1836-2021), Bordeaux (France 1851-2021), Paris (France 1658-2019), Gorkij (Russia 1881-1989),Tenkodogo (Burkina Faso 1951-1991), Cucuta (Colombia 1961-2021), Nassau (Bahamas 1811-2021), St Helena (1892-2021), Ascension (1923-2021), Armagh (UK 1796-2021), Tianjin (China 1890-2021), Antananarivo (Madagascar 1889-2021)

Table 3. Station counts per normal category code in CRUTEM5 and GloSATref. Of those stations coded as have missing normals, 26 do have a normal in up to 5 calendar months. Climatology uncertainty models for data derived, WMO and extrapolated normal are described in Brohan et al. (2006) and Morice et al. (2012).

Normal Code	Category of normal	CRUTEM station count	GloSATref station count	Climatology uncertainty model used
1	Normals missing	26492~~2656~~	731~~738~~	Not applicable
2	Estimated using previous infilling methods	100	11	Extrapolated
3	WMO	25	1	WMO
4	Calculated directly from observations	7855	5806	Data
5	Taken from previous data set version	3	6~~7~~	Data
6	Estimated solely from LEK	N/A	1742	Extrapolated
7	From combination of data and LEK	N/A	3568	Extrapolated
Total	**All usable types [categories 2-7]**	**7983**	**1113~~411135~~**	

Table 4. Parameters used to construct the GloSATMAT height ensemble.

Parameter	Sampling distribution
Year: start of increasing heights around 1870	Normal: mean 1870, standard deviation 3 years
Year: start of World War 2 heights	Uniform: 1939 $\pm$ 2
Year: end of World War 2 heights	Uniform: 1946 $\pm$ 5
Height: constant until around 1870	Normal: mean 6 m, standard deviation 2 m
Height: reached at start of World War 2	Normal: mean 12 m, standard deviation 2 m
Height: change at start of World War 2	Normal: mean -1 m, standard deviation 2 m
Height: reached by 1973	Normal: mean 16 m, standard deviation 2 m

Table 5. Representation of uncertainty model components in LSAT and MAT error models.

Analysis domain	Uncertainty term	Analysis input	Error structure
LSAT	Urbanisation bias	Ensemble	One sided piecewise trend; correlated between stations to represent uncertainty in large scale averages (Morice et al., 2012)
	Exposure bias	Ensemble	Two-sided early record bias reducing to zero in mid-20th century; correlated between stations to represent uncertainty in large scale averages (Morice et al., 2012)
	Homogenization error	Ensemble	Random step change model, independent between stations (Morice et al., 2012)
	Measurement error	Error covariance	Random error uncorrelated between stations/grid cells (Morice et al., 2012)
	Grid cell sampling error	Error covariance	Random error uncorrelated between stations/grid cells (Morice et al., 2012)
MAT	Height adjustment / stability uncertainty	Ensemble	Per-ship height uncertainty and per 5° grid-cell/10-day period stability uncertainty encoded into ensemble grids
	Climatology uncertainty	Not used	Not used
	WW2 bias	Error covariance	Per-ship/year uncertainty during the period 1942-1945 for ships from Decks 195 and 245 (Cornes et al., 2020)
	Diurnal adjustment error	Not used	Included in measurement random error
	Measurement random error	Error covariance	Random error uncorrelated between ship/grid cells (Cornes et al., 2020)
	Measurement bias error	Error covariance	Correlated uncertainty between ships/grid cells (Cornes et al., 2020)
	Grid cell sampling error	Error covariance	Per-ship errors encoded into grid error covariance matrices (Cornes et al., 2020)

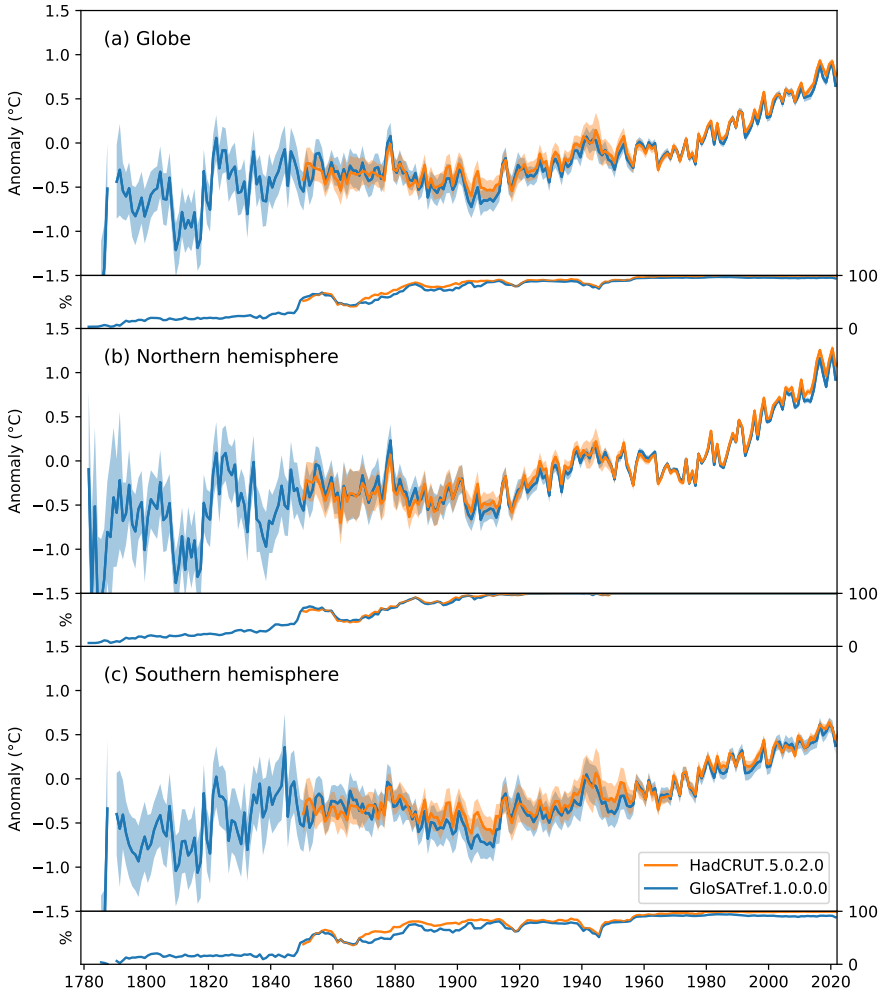

Figure 1. Global and hemispheric average temperature anomaly time series for the GloSATref and HadCRUT5 global analyses (°C, relative to 1961-1990), together with the % of areal grid coverage for (a) the full globe (b) the northern hemisphere and (c) the southern hemisphere. Hemispheric series omit years in which there are no data available in the respective hemisphere. Global averages are computed as an average of northern and southern hemispheric series, requiring data to be available in both hemispheres. Methods for computation of time series and their uncertainties are provided in (Morice et al., 2021).

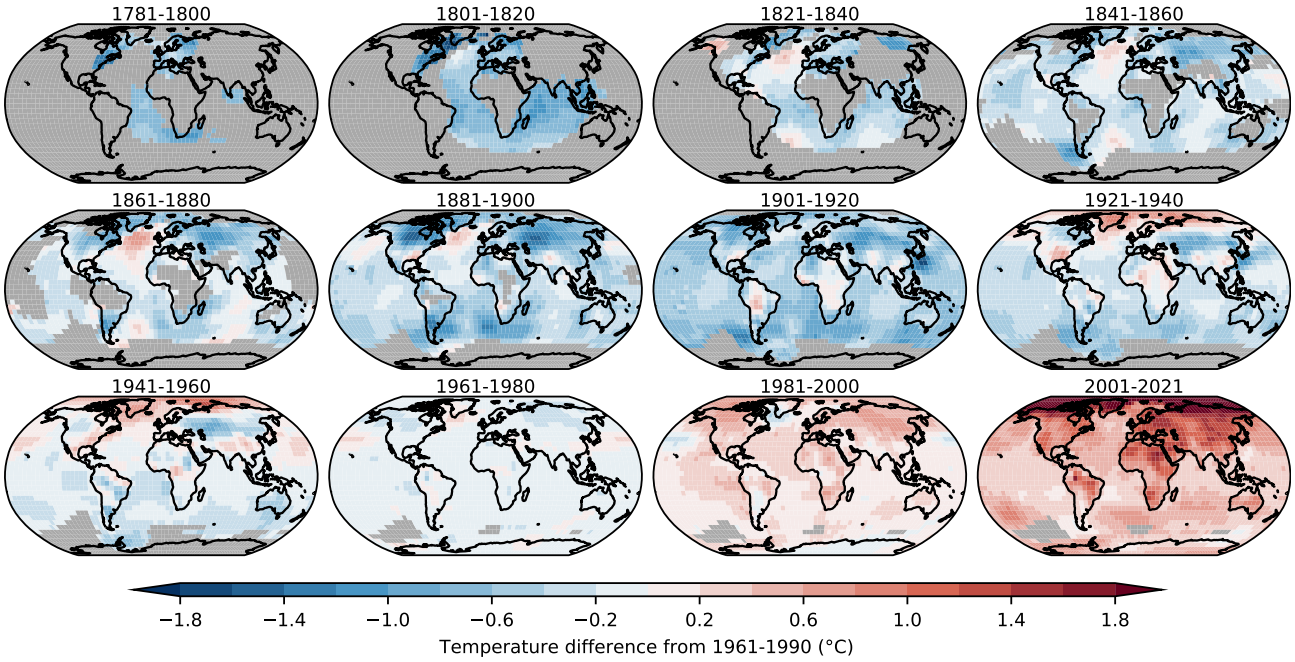

Figure 2. 20-year average SAT for the GloSATref analysis (°C relative to 1961-1990, final panel shows 21-year average). Each panel averages available gridded data to quarterly, annual and then 20-year (21-year) averages, requiring data in at least two quarters to form an annual average and at least 10 annual averages to produce a 20-year average.

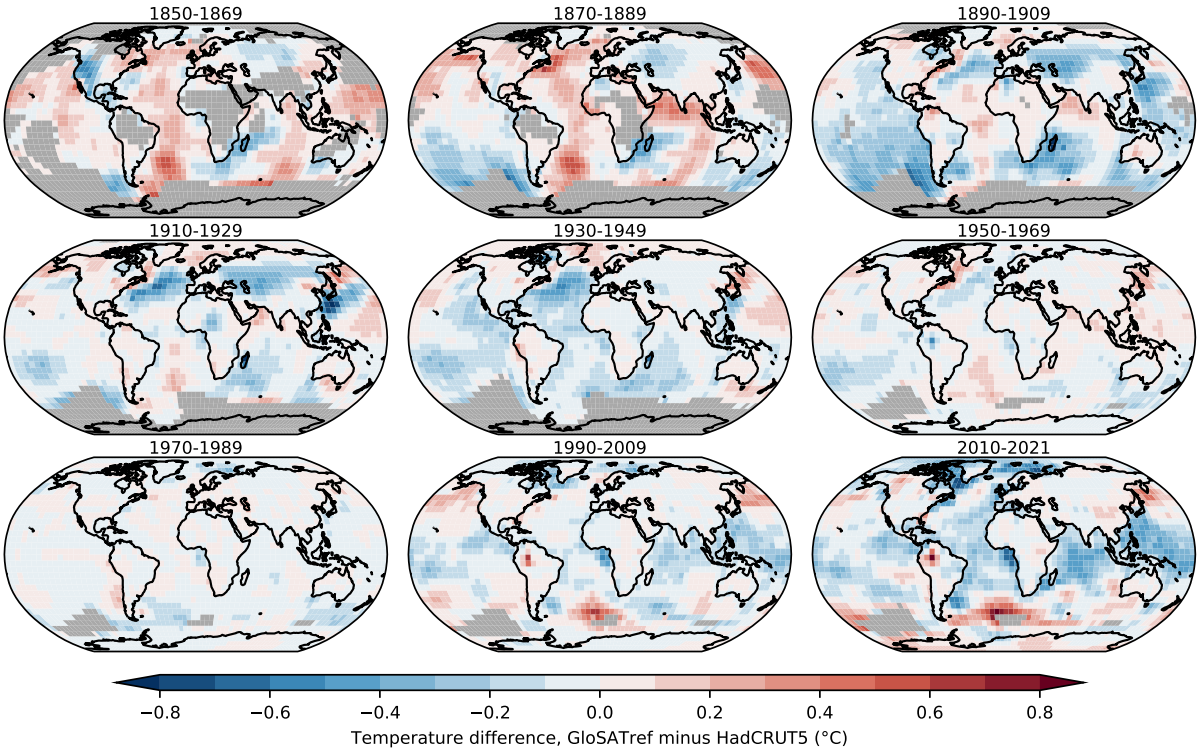

Figure 3. Difference in 20-year average anomalies, relative to 1961-1990, between the GloSATref and HadCRUT5 analyses (°C, final panel shows 12-year average). Blue colours indicate where HadCRUT5 is colder than GloSATref. Red indicates where HadCRUT5 is warmer.

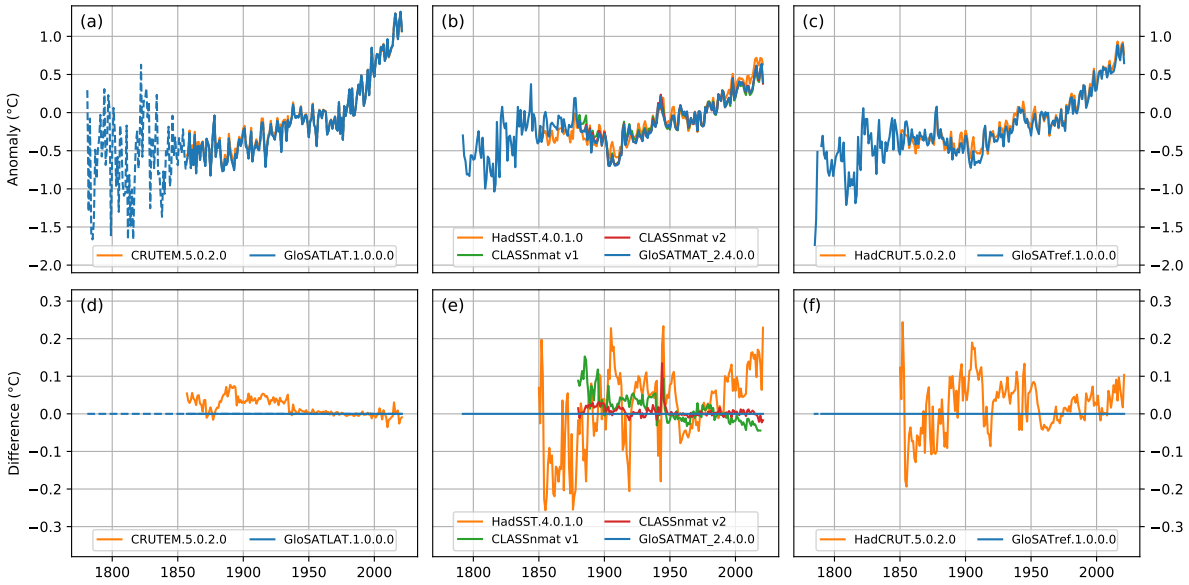

Figure 4. Top row: global average temperature anomalies (°C) relative to 1961-1990 for (a) LSAT, (b) SST/NMAT/MAT, and (c) GMST/GSAT. Bottom row: differences in global average 1961-1990 temperature anomaly time series for (d) LSAT minus CRUTEM.5.0.2.0 LSAT, (e) SST/NMAT/MAT minus HadSST.4.0.1.0 SST, and (f) GMST/GSAT minus HadCRUT.5.0.2.0. The dashed line for GloSATLAT extends the series prior to 1856 by omitting the requirement for at least 5 populated grid cells in each hemisphere, with only Northern Hemisphere station data present in GloSATLAT grids prior 1854. None of the LSAT/SST/NMAT/MAT data sets in panels (a, b, d, e) use spatial interpolation while spatial interpolation is used in GSAT/GMST analyses shown in panels (c, f). Plotted series use all available gridded data without co-location to common data coverage.

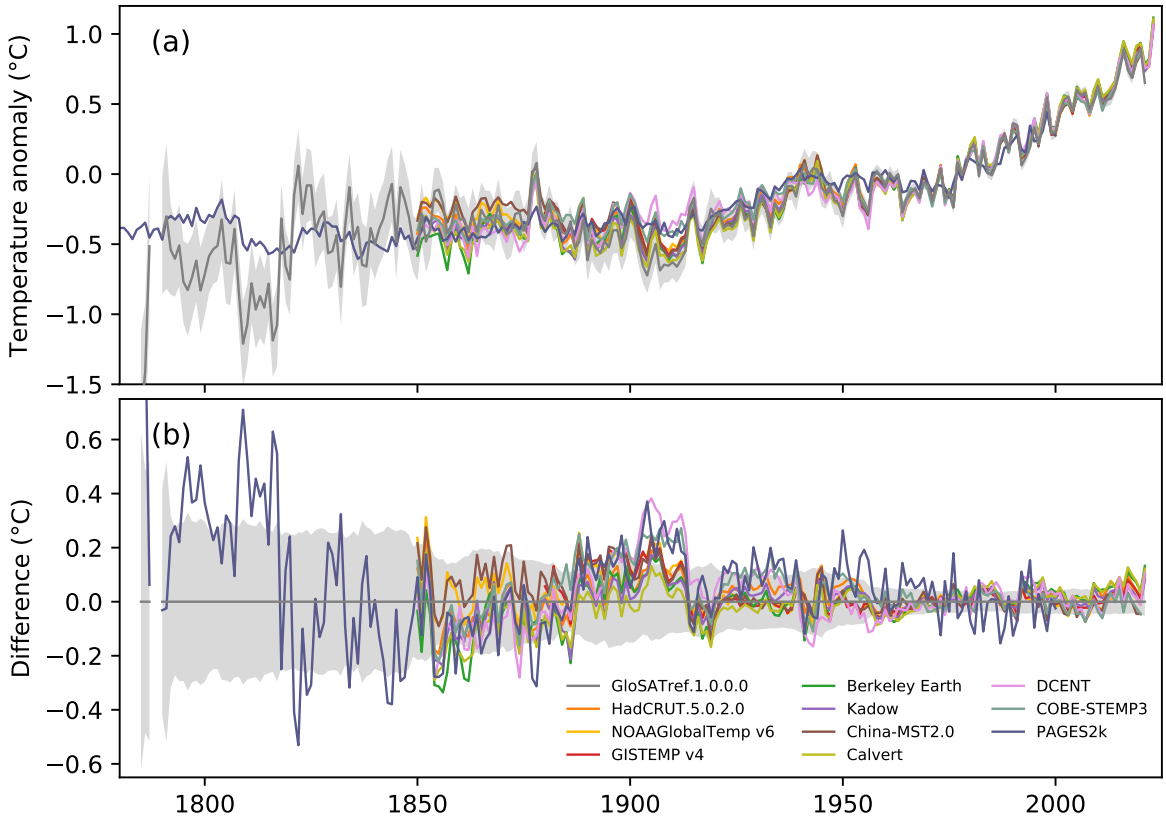

Figure 5. (a) GSAT and GMST (°C relative to 1961-1990) derived from instrumental data sets and the ensemble median of the PAGES2k palaeoclimate reconstructions and (b) differences between GloSATref and instrumental GMST and the PAGES2k ensemble median (each series minus GloSATref). Solid lines show the ensemble means or medians where the dataset provides an ensemble; grey shading is the 2.5% to 97.5% confidence interval for the GloSATref series only. All instrumental series are January-December annual averages. PAGES2k series are representative of April-March annual averages.

References

Anchukaitis, K. J. and Smerdon, J. E.: Progress and uncertainties in global and hemispheric temperature reconstructions of the Common Era, Quaternary Science Reviews, 286, 107 537, https://doi.org/10.1016/j.quascirev.2022.107537, 2022.

Ashcroft, L., Gergis, J., and Karoly, D. J.: A historical climate dataset for southeastern Australia, 1788–1859, Geoscience Data Journal, 1, 158–178, https://doi.org/10.1002/gdj3.19, 2014.

Ballinger, A., Schurer, A., Hegerl, G., Dittus, A., Hawkins, E., Cornes, R., Kent, E., Marshall, L., Morice, C., Osborn, T., and Rayner, N.A. Rumbold, S.: Importance of beginning industrial-era climate simulations in the eighteenth century, Environmental Research Letters, submitted.

Berry, D. I. and Kent, E. C.: Air–Sea fluxes from ICOADS: the construction of a new gridded dataset with uncertainty estimates, International Journal of Climatology, 31, 987–1001, https://doi.org/10.1002/joc.2059, 2011.

Berry, D. I. and Kent, E. C.: Assessing the health of the in situ global surface marine climate observing system, International Journal of Climatology, 37, 2248–2259, https://doi.org/10.1002/joc.4914, 2016.

Berry, D. I., Kent, E. C., and Taylor, P. K.: An Analytical Model of Heating Errors in Marine Air Temperatures from Ships, Journal of Atmospheric and Oceanic Technology, 21, 1198–1215, https://doi.org/10.1175/1520-0426(2004)021<1198:aamohe>2.0.co;2, 2004.

Biri, S., Cornes, R. C., Berry, D. I., Kent, E. C., and Yelland, M. J.: AirSeaFluxCode: Open-source software for calculating turbulent air-sea fluxes from meteorological parameters, Frontiers in Marine Science, 9, https://doi.org/10.3389/fmars.2022.1049168, 2023.

Böhm, R., Jones, P. D., Hiebl, J., Frank, D., Brunetti, M., and Maugeri, M.: The early instrumental warm-bias: a solution for long central European temperature series 1760–2007, Climatic Change, 101, 41–67, https://doi.org/10.1007/s10584-009-9649-4, 2009.

Bottomley, M., Folland, C. K., Hsiung, J., Newell, R. E., and Parker, D. E.: Global Ocean Surface Temperature Atlas (GOSTA), Meteorological Office (Bracknell) and Massachusetts Institute of Technology, 1990.

Brohan, P., Kennedy, J. J., Harris, I., Tett, S. F. B., and Jones, P. D.: Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850, Journal of Geophysical Research: Atmospheres, 111, https://doi.org/10.1029/2005jd006548, 2006.

Brohan, P., Allan, R., Freeman, J. E., Waple, A. M., Wheeler, D., Wilkinson, C., and Woodruff, S.: Marine Observations of Old Weather, Bulletin of the American Meteorological Society, 90, 219–230, https://doi.org/10.1175/2008bams2522.1, 2009.

Brönnimann, S., Brugnara, Y., Allan, R. J., Brunet, M., Compo, G. P., Crouthamel, R. I., Jones, P. D., Jourdain, S., Luterbacher, J., Siegmund, P., Valente, M. A., and Wilkinson, C. W.: A roadmap to climate data rescue services, Geoscience Data Journal, 5, 28–39, https://doi.org/10.1002/gdj3.56, 2018.

Brönnimann, S., Franke, J., Nussbaumer, S. U., Zumbühl, H. J., Steiner, D., Trachsel, M., Hegerl, G. C., Schurer, A., Worni, M., Malik, A., Flückiger, J., and Raible, C. C.: Last phase of the Little Ice Age forced by volcanic eruptions, Nature Geoscience, 12, 650–656, https://doi.org/10.1038/s41561-019-0402-y, 2019.

Brunet, M., SaladiÉ, O., Jones, P., SigrÓ, J., Aguilar, E., Moberg, A., Lister, D., Walther, A., Lopez, D., and Almarza, C.: The development of a new dataset of Spanish Daily Adjusted Temperature Series (SDATS) (1850–2003), International Journal of Climatology, 26, 1777–1802, https://doi.org/10.1002/joc.1338, 2006.

Brunet, M., Asin, J., Sigró, J., Bañón, M., García, F., Aguilar, E., Palenzuela, J. E., Peterson, T. C., and Jones, P.: The minimization of the screen bias from ancient Western Mediterranean air temperature records: an exploratory statistical analysis, International Journal of Climatology, 31, 1879–1895, https://doi.org/10.1002/joc.2192, 2011.

Calvert, B. T. T.: Improving global temperature datasets to better account for non-uniform warming, Quarterly Journal of the Royal Meteorological Society, https://doi.org/10.1002/qj.4791, 2024.

Chan, D. and Huybers, P.: Correcting Observational Biases in Sea Surface Temperature Observations Removes Anomalous Warmth during World War II, Journal of Climate, 34, 4585 – 4602, https://doi.org/10.1175/JCLI-D-20-0907.1, 2021.

Chan, D., Kent, E. C., Berry, D. I., and Huybers, P.: Correcting datasets leads to more homogeneous early-twentieth-century sea surface warming, Nature, 571, 393–397, 2019.

Chan, D., Gebbie, G., Huybers, P., and Kent, E. C.: A Dynamically Consistent ENsemble of Temperature at the Earth surface since 1850 from the DCENT dataset, Scientific Data, 11, https://doi.org/10.1038/s41597-024-03742-x, 2024.

Cornes, R., Berry, D., Junod, R., Kent, E. C., and Rayner, N.: Sidebar 2.1. Night marine air temperature [in "State of the Climate in 2020"], Bull. Amer. Meteor., 102, S39–S41, https://doi.org/10.1175/BAMS-D-21-0098.1, 2021.

Cornes, R., Kent, E., and Cropper, T.: CLASSnmat v2, https://doi.org/10.6084/m9.figshare.27015661, 2024.

Cornes, R. C., Kent, E., Berry, D., and Kennedy, J. J.: CLASSnmat: A global night marine air temperature data set, 1880–2019, Geoscience Data Journal, 7, 170–184, https://doi.org/10.1002/gdj3.100, 2020.

Cowtan, K. and Way, R. G.: Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends, Quarterly Journal of the Royal Meteorological Society, 140, 1935–1944, https://doi.org/10.1002/qj.2297, 2014.

Cowtan, K., Hausfather, Z., Hawkins, E., Jacobs, P., Mann, M. E., Miller, S. K., Steinman, B. A., Stolpe, M. B., and Way, R. G.: Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures, Geophysical Research Letters, 42, 6526–6534, https://doi.org/10.1002/2015gl064888, 2015.

Cropper, T. E., Berry, D. I., Cornes, R. C., and Kent, E. C.: Quantifying Daytime Heating Biases in Marine Air Temperature Observations from Ships, Journal of Atmospheric and Oceanic Technology, 40, 427–438, https://doi.org/10.1175/jtech-d-22-0080.1, 2023.

Folland, C. K., Rayner, N. A., Brown, S. J., Smith, T. M., Shen, S. S. P., Parker, D. E., Macadam, I., Jones, P. D., Jones, R. N., Nicholls, N., and Sexton, D. M. H.: Global temperature change and its uncertainties since 1861, Geophysical Research Letters, 28, 2621–2624, https://doi.org/10.1029/2001gl012877, 2001.

Freeman, E., Woodruff, S. D., Worley, S. J., Lubker, S. J., Kent, E. C., Angel, W. E., Berry, D. I., Brohan, P., Eastman, R., Gates, L., Gloeden, W., Ji, Z., Lawrimore, J., Rayner, N. A., Rosenhagen, G., and Smith, S. R.: ICOADS Release 3.0: a major update to the historical marine climate record, International Journal of Climatology, 37, 2211–2232, https://doi.org/10.1002/joc.4775, 2017.

Gulev, S. K., Thorne, P. W., Ahn, J., Dentener, F. J., Domingues, C. M., Gerland, S., Gong, D.and Kaufman, D. S., Nnamchi, H. C., Quaas, J., Rivera, J. A., Sathyendranath, S., Smith, S. L., Trewin, B., von Schuckmann, K., and Vose, R. S.: Changing State of the Climate System, pp. 287–422, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021.

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

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne, M. J., Smith, T. M., Vose, R. S., and Zhang, H.-M.: Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5): Upgrades, Validations, and Intercomparisons, Journal of Climate, 30, 8179–8205, https://doi.org/10.1175/jcli-d-16-0836.1, 2017.

Huang, B., Yin, X., Menne, M. J., Vose, R., and Zhang, H.-M.: Improvements to the Land Surface Air Temperature Reconstruction in NOAAGlobalTemp: An Artificial Neural Network Approach, Artificial Intelligence for the Earth Systems, 1, https://doi.org/10.1175/aies-d-22-0032.1, 2022.

Huang, B., Yin, X., Boyer, T., Liu, C., Menne, M., Rao, Y. D., Smith, T., Vose, R., and Zhang, H.-M.: Extended Reconstructed Sea Surface Temperature, Version 6 (ERSSTv6). Part I: An Artificial Neural Network Approach, Journal of Climate, 38, 1105 – 1121, https://doi.org/10.1175/JCLI-D-23-0707.1, 2025.

Ishii, M., Nishimura, A., Yasui, S., and Hirahara, S.: Historical High-Resolution Daily SST Analysis (COBE-SST3) with Consistency to Monthly Land Surface Air Temperature, Journal of the Meteorological Society of Japan. Ser. II, 103, 17–44, https://doi.org/10.2151/jmsj.2025-002, 2025.

Jones, G. S.: "Apples and Oranges": On comparing simulated historic near-surface temperature changes with observations, Quarterly Journal of the Royal Meteorological Society, 146, 3747–3771, https://doi.org/10.1002/qj.3871, 2020.

Jones, P. D. and Moberg, A.: Hemispheric and Large-Scale Surface Air Temperature Variations: An Extensive Revision and an Update to 2001, Journal of Climate, 16, 206–223, https://doi.org/10.1175/1520-0442(2003)016<0206:halssa>2.0.co;2, 2003.

Jones, P. D., New, M., Parker, D. E., Martin, S., and Rigor, I. G.: Surface air temperature and its changes over the past 150 years, Reviews of Geophysics, 37, 173–199, https://doi.org/10.1029/1999rg900002, 1999.

Jones, P. D., Lister, D. H., Osborn, T. J., Harpham, C., Salmon, M., and Morice, C. P.: Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010, Journal of Geophysical Research: Atmospheres, 117, https://doi.org/10.1029/2011jd017139, 2012.

Junod, R. A. and Christy, J. R.: A new compilation of globally gridded night-time marine air temperatures: The UAHN-MATv1 dataset, International Journal of Climatology, 40, 2609–2623, https://doi.org/10.1002/joc.6354, 2020.

Kadow, C., Hall, D. M., and Ulbrich, U.: Artificial intelligence reconstructs missing climate information, Nature Geoscience, 13, 408–413, https://doi.org/10.1038/s41561-020-0582-5, 2020.

Kennedy, J. J., Rayner, N. A., Atkinson, C. P., and Killick, R. E.: An Ensemble Data Set of Sea Surface Temperature Change From 1850: The Met Office Hadley Centre HadSST.4.0.0.0 Data Set, Journal of Geophysical Research: Atmospheres, 124, 7719–7763, https://doi.org/10.1029/2018jd029867, 2019.

Kent, E. C. and Kennedy, J. J.: Historical Estimates of Surface Marine Temperatures, Annual Review of Marine Science, 13, 283–311, https://doi.org/10.1146/annurev-marine-042120-111807, 2021.

Kent, E. C., Woodruff, S. D., and Berry, D. I.: Metadata from WMO Publication No. 47 and an Assessment of Voluntary Observing Ship Observation Heights in ICOADS, Journal of Atmospheric and Oceanic Technology, 24, 214–234, https://doi.org/10.1175/jtech1949.1, 2007.

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

Kent, E. C., Rayner, N. A., Berry, D. I., Eastman, R., Grigorieva, V. G., Huang, B., Kennedy, J. J., Smith, S. R., and Willett, K. M.: Observing Requirements for Long-Term Climate Records at the Ocean Surface, Frontiers in Marine Science, 6, https://doi.org/10.3389/fmars.2019.00441, 2019.

Lenssen, N. J. L., Schmidt, G. A., Hansen, J. E., Menne, M. J., Persin, A., Ruedy, R., and Zyss, D.: Improvements in the GISTEMP Uncertainty Model, Journal of Geophysical Research: Atmospheres, 124, 6307–6326, https://doi.org/10.1029/2018jd029522, 2019.

Li, Q., Sun, W., Yun, X., Huang, B., Dong, W., Wang, X. L., Zhai, P., and Jones, P.: An updated evaluation of the global mean land surface air temperature and surface temperature trends based on CLSAT and CMST, Climate Dynamics, 56, 635–650, https://doi.org/10.1007/s00382-020-05502-0, 2021.

Luterbacher, J., Allan, R., Wilkinson, C., Hawkins, E., Teleti, P., Lorrey, A., Brönnimann, S., Hechler, P., Velikou, K., and Xoplaki, E.: The Importance and Scientific Value of Long Weather and Climate Records; Examples of Historical Marine Data Efforts across the Globe, Climate, 12, 39, https://doi.org/10.3390/cli12030039, 2024.

Menne, M. J. and Williams, C. N.: Homogenization of Temperature Series via Pairwise Comparisons, Journal of Climate, 22, 1700–1717, https://doi.org/10.1175/2008jcli2263.1, 2009.

Menne, M. J., Williams, C. N., Gleason, B. E., Rennie, J. J., and Lawrimore, J. H.: The Global Historical Climatology Network Monthly Temperature Dataset, Version 4, Journal of Climate, 31, 9835–9854, https://doi.org/10.1175/jcli-d-18-0094.1, 2018.

Morice, C. P., Kennedy, J. J., Rayner, N. A., and Jones, P. D.: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, Journal of Geophysical Research: Atmospheres, 117, https://doi.org/10.1029/2011jd017187, 2012.

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., Dunn, R. J. H., Osborn, T. J., Jones, P. D., and Simpson, I. R.: An Updated Assessment of Near-Surface Temperature Change From 1850: The HadCRUT5 Data Set, Journal of Geophysical Research: Atmospheres, 126, https://doi.org/10.1029/2019jd032361, 2021.

Naylor, S.: Thermometer screens and the geographies of uniformity in nineteenth-century meteorology, Notes and Records: the Royal Society Journal of the History of Science, 73, 203–221, https://doi.org/10.1098/rsnr.2018.0037, 2018.

Osborn, T. J., Jones, P. D., Lister, D. H., Morice, C. P., Simpson, I. R., Winn, J. P., Hogan, E., and Harris, I. C.: Land Surface Air Temperature Variations Across the Globe Updated to 2019: The CRUTEM5 Data Set, Journal of Geophysical Research: Atmospheres, 126, https://doi.org/10.1029/2019jd032352, 2021.

PAGES2k: Consistent multidecadal variability in global temperature reconstructions and simulations over the Common Era, Nature Geoscience, 12, 643–649, https://doi.org/10.1038/s41561-019-0400-0, 2019.

Parker, D. E.: Effects of changing exposure of thermometers at land stations, International Journal of Climatology, 14, 1–31, https://doi.org/10.1002/joc.3370140102, 1994.

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A.: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, Journal of Geophysical Research: Atmospheres, 108, https://doi.org/10.1029/2002jd002670, 2003.

Richardson, M., Cowtan, K., and Millar, R. J.: Global temperature definition affects achievement of long-term climate goals, Environmental Research Letters, 13, 054 004, https://doi.org/10.1088/1748-9326/aab305, 2018.

Richardson, M. T.: A Physical Explanation for Ocean Air–Water Warming Differences under CO2-Forced Warming, Journal of Climate, 36, 2857–2871, https://doi.org/10.1175/jcli-d-22-0215.1, 2023.

Rohde, R., Muller, R., , Jacobsen, R., Muller, E., Perlmutter, S., Rosenfeld, A., Wurtel, J., Donald, G., and Wickham, C.: A New Estimate of the Average Earth Surface Land Temperature Spanning 1753 to 2011, Geoinformatics & Geostatistics: An Overview, 01, https://doi.org/10.4172/2327-4581.1000101, 2013a.

Rohde, R., Muller, R., Jacobsen, R., Perlmutter, S., and Mosher, S.: Berkeley Earth Temperature Averaging Process, Geoinformatics & Geostatistics: An Overview, 01, https://doi.org/10.4172/2327-4581.1000103, 2013b.

Rohde, R. A. and Hausfather, Z.: The Berkeley Earth Land/Ocean Temperature Record, Earth System Science Data, 12, 3469–3479, https://doi.org/10.5194/essd-12-3469-2020, 2020.

Sippel, S., Kent, E. C., Meinshausen, N., Chan, D., Kadow, C., Neukom, R., Fischer, E. M., Humphrey, V., Rohde, R., de Vries, I., and Knutti, R.: Early-twentieth-century cold bias in ocean surface temperature observations, Nature, 635, 618–624, https://doi.org/10.1038/s41586-024-08230-1, 2024.

Smith, S. R., Briggs, K., Bourassa, M. A., Elya, J., and Paver, C. R.: Shipboard automated meteorological and oceanographic system data archive: 2005–2017, Geoscience Data Journal, 5, 73–86, https://doi.org/10.1002/gdj3.59, 2018.

Spencer, L., McColl, C., Brohan, P., Wood, K., Allan, R., and Compo, G.: OldWeather3 Marine Data for ICOADS Input, https://doi.org/10.5065/Q54H-3R61, 2019.

Sun, W., Yang, Y., Chao, L., Dong, W., Huang, B., Jones, P., and Li, Q.: Description of the China Global Merged Surface Temperature version 2.0, Earth System Science Data, 14, 1677–1693, https://doi.org/10.5194/essd-14-1677-2022, 2022.

Taylor, M., Osborn, T. J., Cowtan, K., Morice, C. P., Jones, P. D., Wallis, E. J., and Lister, D. H.: GloSAT LATsdb: a global compilation of land air temperature station records with updated climatological normals from local expectation kriging, Geoscience Data Journal, in review, 2025.

Wallis, E. J., Osborn, T. J., Taylor, M., Jones, P. D., Joshi, M., and Hawkins, E.: Quantifying exposure biases in early instrumental land surface air temperature observations, International Journal of Climatology, 44, 1611–1635, https://doi.org/10.1002/joc.8401, 2024.

World Meteorological Organisation: Climatological Normals (CLINO) for the Period 1961-1990, WMO-No. 847, https://community.wmo.int/en/activity-areas/climate-services/climate-products-and-initiatives/wmo-climatological-normals, 1996.

World Meteorological Organisation: Guide to Instruments and Methods of Observation, WMO-No. 8, https://community.wmo.int/en/activity-areas/imop/wmo-no_8, 2018.

Zambri, B., Robock, A., Mills, M. J., and Schmidt, A.: Modeling the 1783–1784 Laki Eruption in Iceland: 2. Climate Impacts, Journal of Geophysical Research: Atmospheres, 124, 6770–6790, https://doi.org/10.1029/2018jd029554, 2019.

S1 Data coverage without statistical infilling

Data coverage of the non-infilled grids of the GloSATref data set is shown in Fig. S1. For comparison, data coverage for the non-infilled grids of the HadCRUT5 data set (Morice et al., 2021) is shown in Fig. S2. These figures provide a comparison of data coverage for the underpinning observational data that contributes to each data set. Differences in data coverage between these two data sets result from (1) the use of marine air temperature observations in GloSATref and sea-surface temperature in HadCRUT5, (2) land data acquisitions in GloSATref that extend the CRUTEM5 station database and (3) the capability to grid additional land station series in GloSATref through the use of new station climatology estimates. Further discussion of data coverage is provided in the main article.

S2 Data coverage of the GloSATref analysis

Data coverage of the GloSATref statistically infilled analysis fields is shown in Fig. S3. For comparison, data coverage for the HadCRUT5 analysis (Morice et al., 2021) is shown in Fig. S4. The two data sets share common statistical analysis methods but differ in observation data used.

Differences in data coverage between the GloSATref and HadCRUT5 analyses result from (1) difference in underpinning observation data coverage described in S1 and (2) differences in fitted statistical model hyperparameters. The marine air temperature analysis parameters of GloSATref result in interpolation over shorter spatial length scales (marine length scale parameter $\rho = 1300$ km) than the sea-surface temperature analysis of HadCRUT5 (marine length scale parameter $\rho = 1550$ km). Land surface air temperature (LSAT) analyses for the the two data sets share the same parameter estimate values (LSAT length scale parameter $\rho = 1300$ km), hence interpolation distances from available observations are broadly equivalent across land and sea ice regions in the two data sets.

S3 LSAT/MAT representation in GloSATref timeseries

The area represented by LSAT and MAT anomalies in the GloSATref analysis varies over time caused by changes in LSAT and MAT observation coverage and changes in sea ice coverage. For global and regional averages computed from the GloSATref analysis this results in varying relative weightings towards land and ocean averages throughout the observed record. The relative weightings towards LSAT and MAT analyses are shown in Fig. S5 for global and hemispheric average timeseries, along with comparisons to weighted averages of LSAT and MAT timeseries using a static weighting for each domain. These comparisons are illustrative of the effects of changing LSAT/MAT weighting only for regions that are represented in the analysis. They do no account for unrepresented regions (although estimates of expected errors from unrepresented regions are included GloSATref uncertainty estimates for global and regional averages through the coverage uncertainty model (Morice et al., 2021)]).

[1] Added Fig. S5

Comparison of global average temperatures indicate that changes in LSAT/MAT weighting have an effect of typically less than 0.1°C for the observed region prior to the 1900s, with individual years prior to the 1850s showing differences approaching 0.3°C. Differences between GloSATref merged analysis series and the fixed weight series are less than 0.02°C following the 1900s. In the LSAT analysis, data is largely unavailable in the southern hemisphere prior to the 1850s. Southern hemisphere series at this time are representative of MAT averages, with differences between the regular merged SAT series and the fixed weight series tending to zero and unable to pro-

vide information about LSAT/MAT sampling error. For both hemispheres, prior to the 1900s, periods of persistent difference between merged analysis and fixed weight series of 0.1 to 0.2°C are coincident with large changes in data availability. For the northern hemisphere form the start on the 1900s onward and for the southern hemisphere from the 1950s onward, differences indicative of the effects of changing LSAT and MAT weighting are less than 0.01°C.

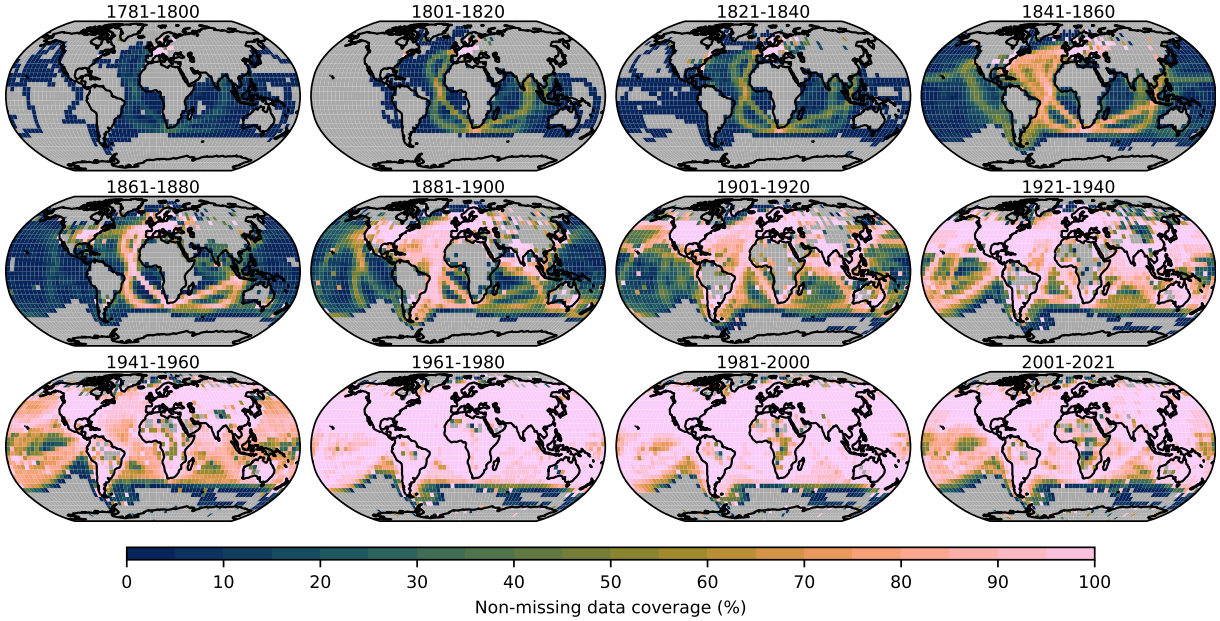

Figure S1. Data coverage for the unfilled GloSATref data set, showing the percentage of months with non-missing data for 20-year periods (final panel shows 21-year period). Grey shaded regions have no data in the corresponding period.

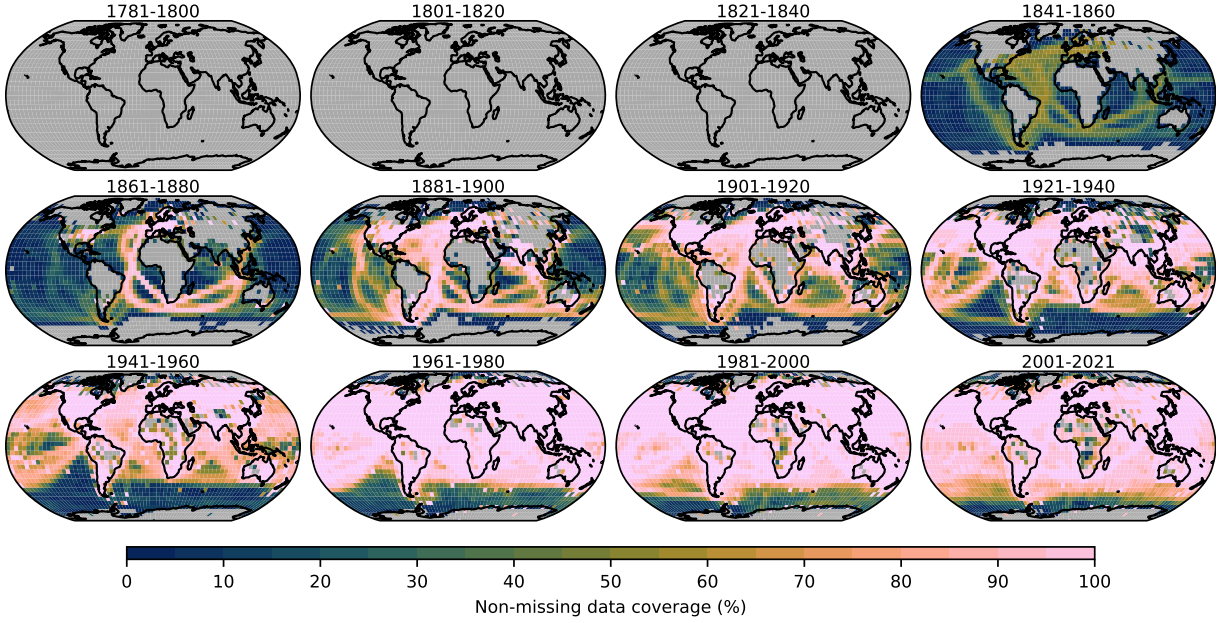

Figure S2. Data coverage in the unfilled HadCRUT5 data set for comparison to Fig. S1, showing the percentage of months with non-missing data for 20-year periods (final panel shows 21-year period). Note that the HadCRUT5 grids contain data from 1850 only.

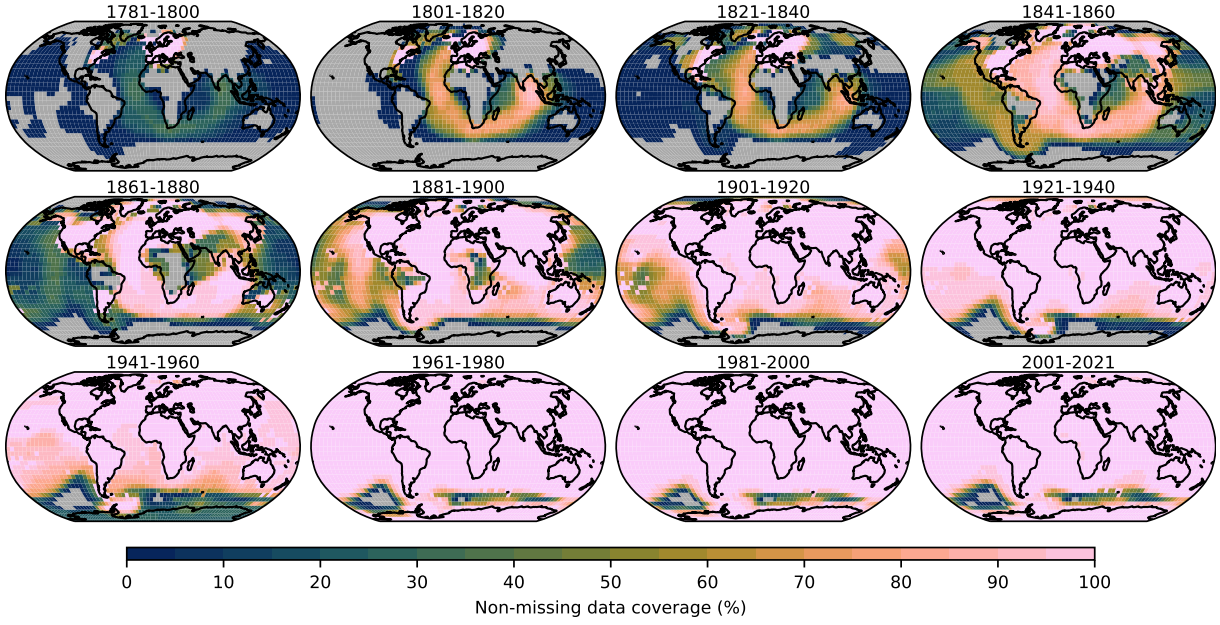

Figure S3. Data coverage for the GloSATref infilled analysis, showing the percentage of months with non-missing data for 20-year periods (final panel shows 21-year period). Grey shaded regions have no data in the corresponding period.

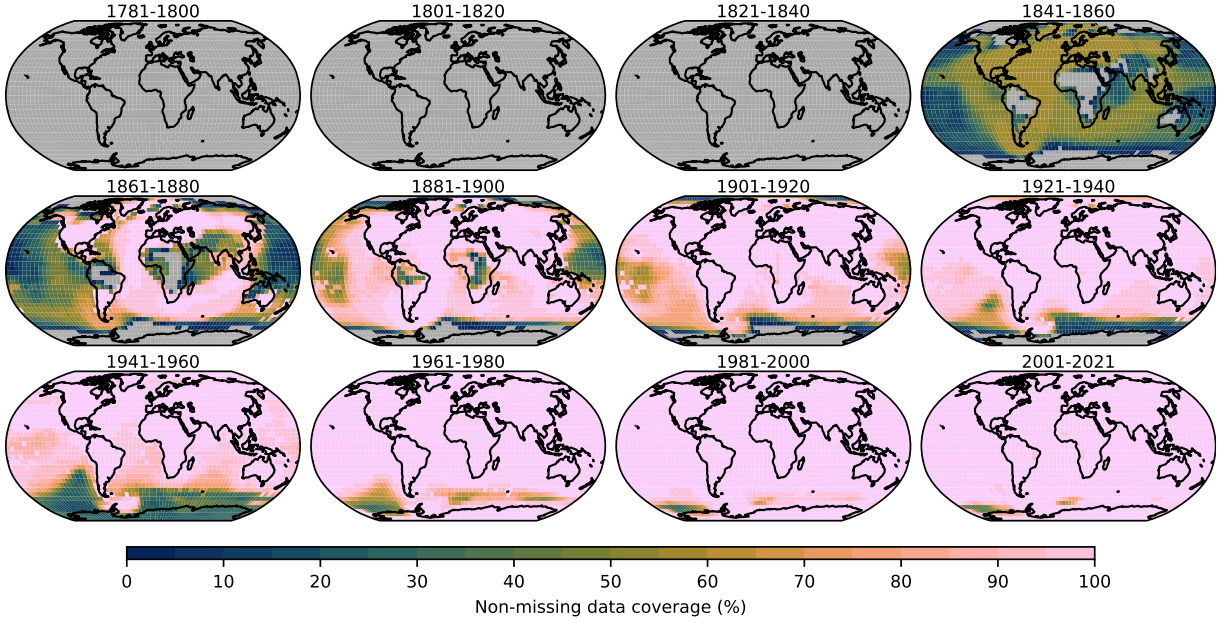

Figure S4. Data coverage in the HadCRUT5 infilled analysis for comparison to Fig. S1, showing the percentage of months with non-missing data for 20-year periods (final panel shows 21-year period). Note that the HadCRUT5 grids contain data from 1850 only.

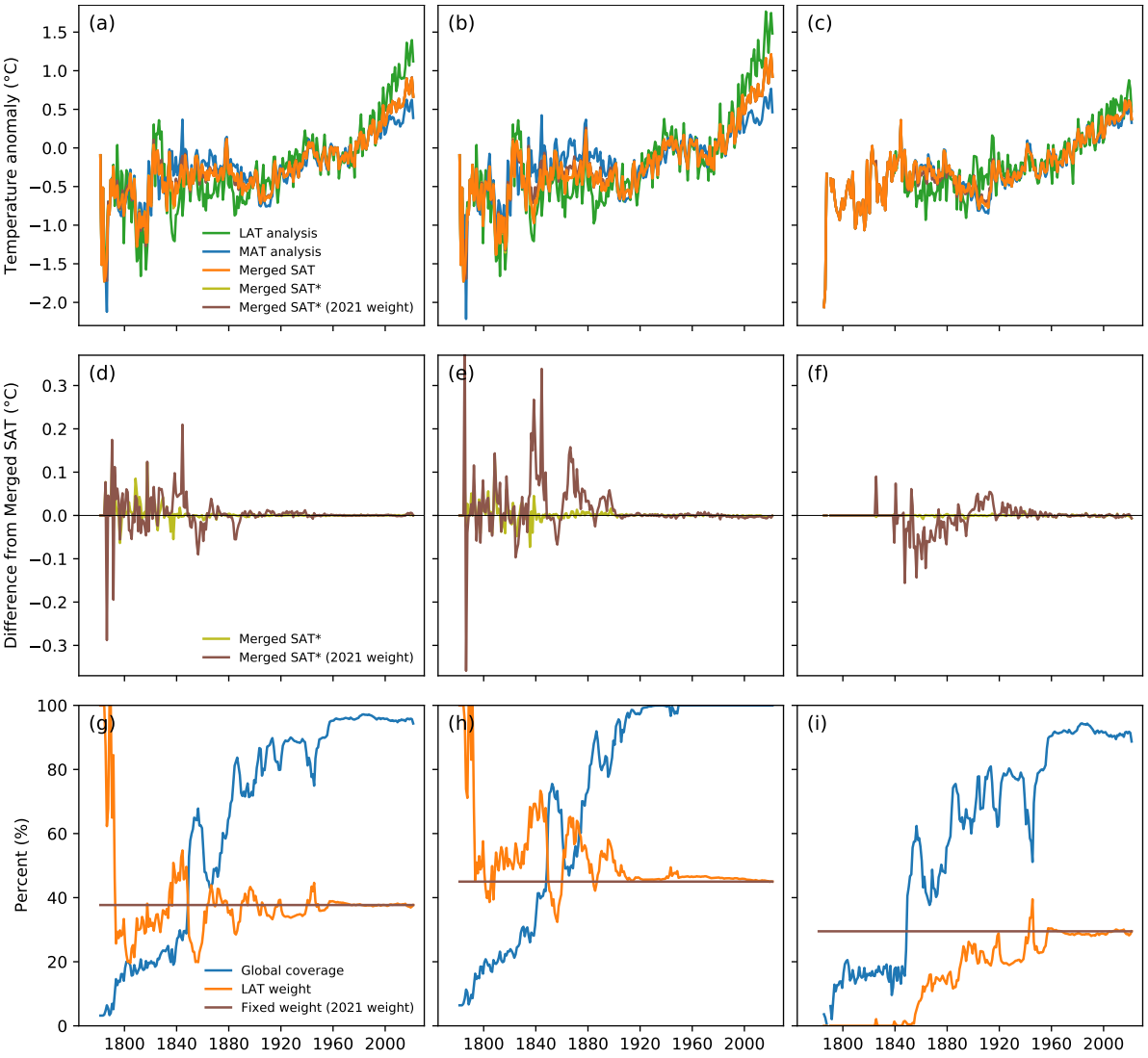

Figure S5. Global and hemispheric annual temperature anomaly time series demonstration the effects of the relative weighting of LSAT and MAT series in the merged analysis. Panels correspond to the following time series: left column (a,d,g), global averages without hemispheric weighting; middle column (b,e,h) northern hemisphere averages, and (c,f,i) southern hemisphere averages. Top row: temperature anomalies, relative to 1961-1990, for MAT and LSAT analysis (including sea ice regions), the merged GST analysis, and variations of the a merged GSAT using different weighting methods. Merged series named with an asterisk are computed as a weighted average of LSAT and MAT series, reverting to single domain analysis series if no data is available for the other domain. Merged SAT* (2021 weighting) series are constructed using a fixed LSAT/MAT weighting for every year applying the weighting from 2021. Merged SAT* applies the varying LSAT/MAT for each analysis year. Differences between Merged SAT and Merged SAT* series demonstrate differences in computing average anomaly series from the merged analysis grids or from a weighted average of SAT/MAT series. Middle row: Differences in SAT series from the Merged SAT series in the corresponding above panels. Bottom row: percentage of total area coverage by analysis grids and relative weighting of LSAT in the Merged SAT series.

S4 Comparisons to paleoclimate reconstructions

Global annual average temperature anomaly series are shown in Fig. S6 in comparison to the PAGES2k multi-method ensemble paleoclimate reconstructions (PAGES2k, 2019). The GloSATref ensemble mean shows greater variability than the PAGES2k median in the early instrumental record. Central estimates exceed the 2.5% and 97.5% confidence limits of each ensemble. Confidence intervals for the GloSATref and PAGES2k ensembles overlap for most years, while departures from confidence intervals occur around the times of the unidentified 1809 volcanic eruption and the 1815 eruption of Mount Tambora. The individual ensemble medians for each paleoclimate reconstruction method have a differing response to these early 19th century volcanoes, with weaker responses than is apparent in the GloSATref analysis global average time series. The GloSATref analysis does, however, have reduced global coverage in this period (see Fig S3), with greater variability resulting from regional sampling.

S5 Extended LSAT and SST/NMAT/MAT comparisons

As in the main article Fig. 4, global average temperature anomaly series over land and marine domains are shown in Fig. S7 and S8, here with additional data sets included that are less closely related to the constituent data sets of the GloSATref analysis. The spread of LSAT data sets is typically greater than the difference between CRUTEM5 and GloSATLAT, likely due to the greater range of input data and processing methods adopted. The Berkeley Earth land data set extends back to 1750 (Rohde et al., 2013a, b), allowing comparison with early record GloSATLAT. The two data sets show similar decadal variability albeit with substantially greater differences between time series prior to the mid-19th century than following. Berkeley Earth uses spatial interpolation while GloSATLAT does not. This contributes to differences between times series for the two data sets. However, as data coverage prior to the mid-19th century is largely confined to Europe and north America, the two data sets are subject to similar regional sampling errors at this time. SST/NMAT/MAT comparisons indicate cooler anomalies in SST data sets in the late 19th century then in GloSATMAT. Early 20th century NMAT/MAT series are cooler than SST, particularly so in comparison to DCSST and COBE-SST3, with recent studies indicating that prexisting SST data sets may be biased cold in this period (Sippel et al., 2024; Chan et al., 2024). The SST data sets exhibit greater warming from SST data sets from the mid-20th century in comparison to CLASSNMAT v1/v2 and GloSATMAT data sets.

[2] Added Fig S7/S8

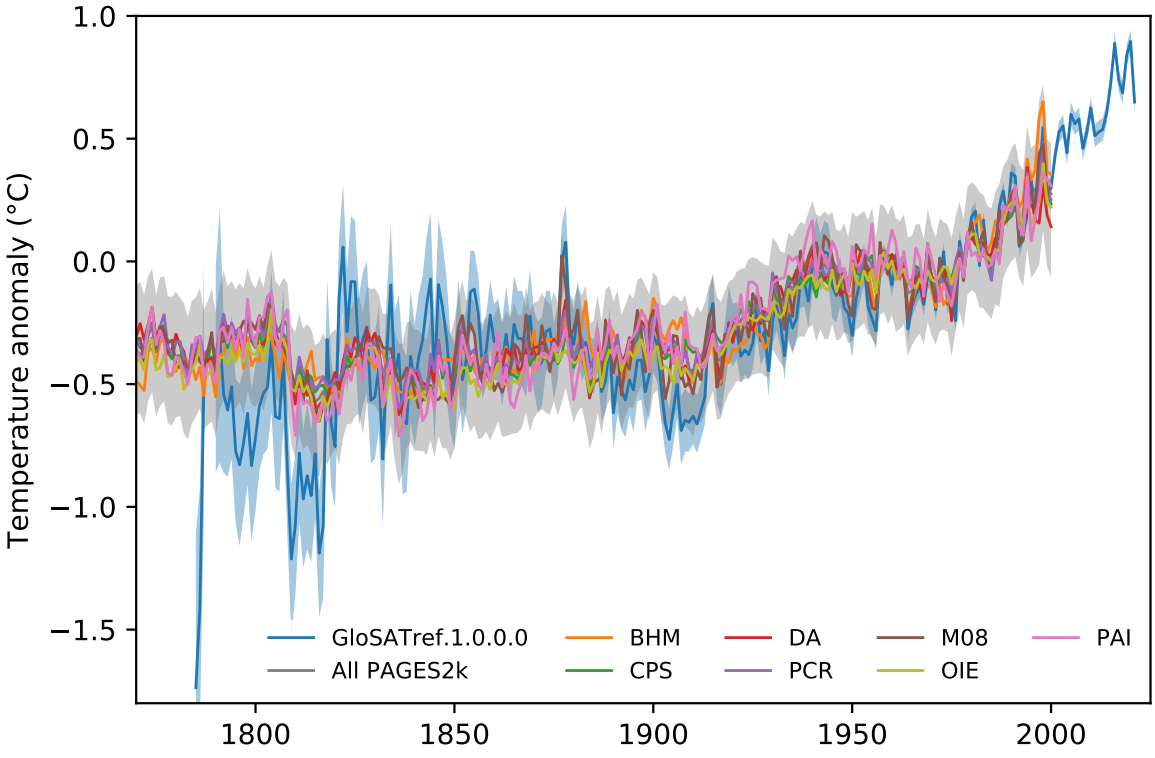

Figure S6. Annual global average temperature anomaly time series for the GloSATref analysis and PAGES2k paleoclimate proxy reconstructions (°C relative to 1961-1990). Uncertainties shown as 95% confidence intervals. Ensemble means for each PAGES2k reconstruction method shown: BHM, CPS, DA, PCR, M08, OIE and PAI.

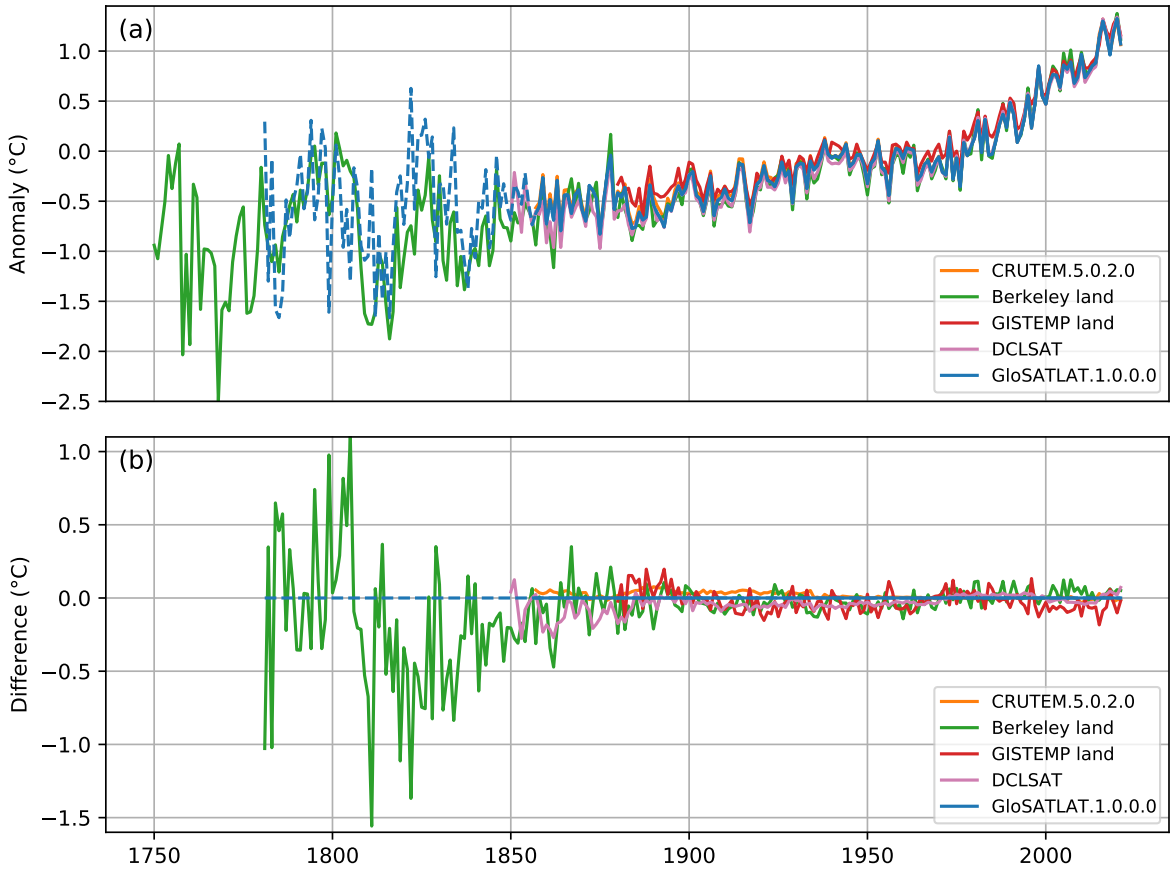

Figure S7. (a) Global average LSAT anomalies (°C) relative to 1961-1990 and (b) differences between each global average LSAT anomaly time series and that of GloSATLAT. The dashed line for GloSATLAT extends the series prior to 1856 by omitting the requirement for at least 5 populated grid cells in each hemisphere, with only Northern Hemisphere station data present in GloSATLAT grids prior 1854. GISTEMP and Berkeley land data sets use spatial interpolation while other data sets do not. DCLSAT series are computed from the gridded fields applying the CRUTEM5 hemispheric weighting scheme of 3/2 northern hemisphere to 1/3 southern hemisphere weighting. Data are not co-located to a common data coverage across data sets.

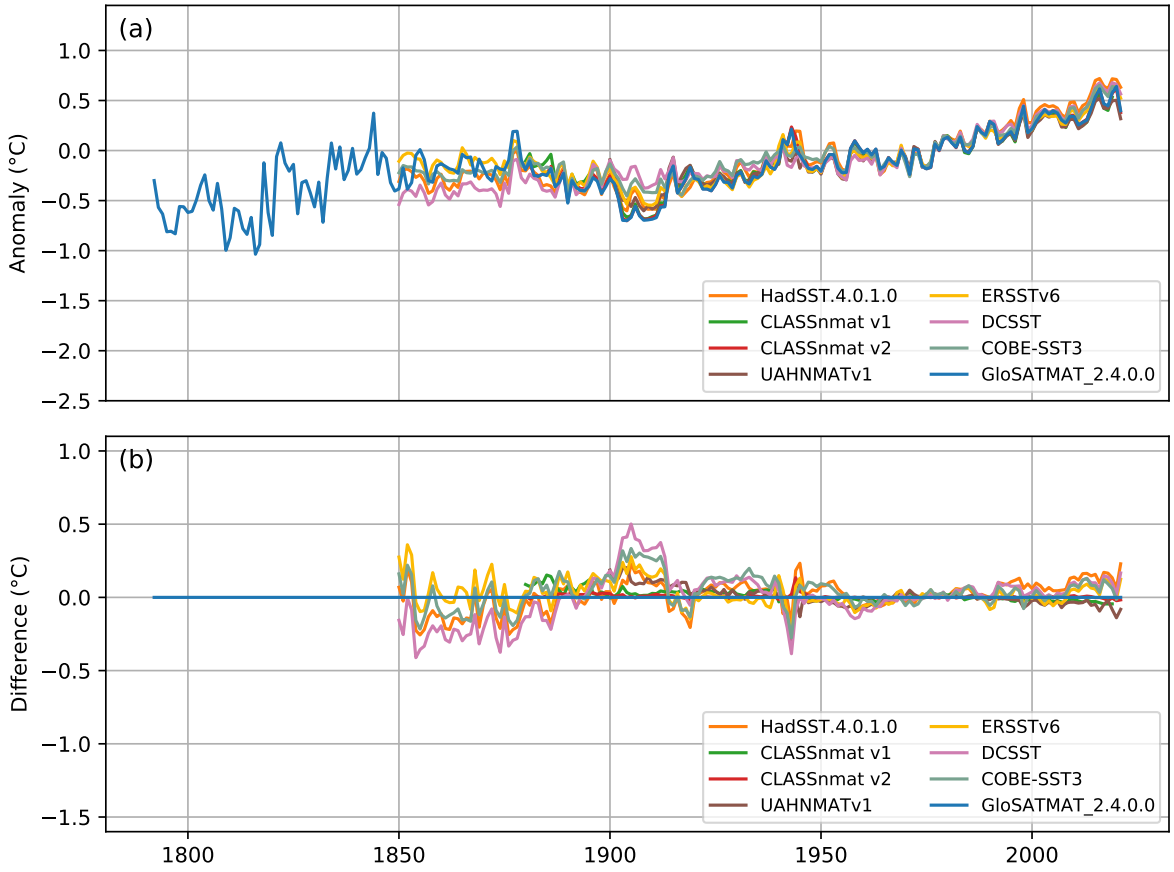

Figure S8. (a) Global average SST/NMAT/MAT anomalies (°C) relative to 1961-1990 and (b) differences between each global average SST/NMAT/MAT anomaly time series and that of GloSATMAT. The plotted data sets do not use spatial interpolation. DCSST series are computed with an area weighted average of the DCSST gridded fields. Data are not co-located to a common data coverage across data sets.

References

Chan, D., Gebbie, G., Huybers, P., and Kent, E. C.: A Dynamically Consistent ENsemble of Temperature at the Earth surface since 1850 from the DCENT dataset, Scientific Data, 11, https://doi.org/10.1038/s41597-024-03742-x, 2024.

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., Dunn, R. J. H., Osborn, T. J., Jones, P. D., and Simpson, I. R.: An Updated Assessment of Near-Surface Temperature Change From 1850: The HadCRUT5 Data Set, Journal of Geophysical Research: Atmospheres, 126, https://doi.org/10.1029/2019jd032361, 2021.

PAGES2k: Consistent multidecadal variability in global temperature reconstructions and simulations over the Common Era, Nature Geoscience, 12, 643–649, https://doi.org/10.1038/s41561-019-0400-0, 2019.

Rohde, R., Muller, R., , Jacobsen, R., Muller, E., Perlmutter, S., Rosenfeld, A., Wurtel, J., Donald, G., and Wickham, C.: A New Estimate of the Average Earth Surface Land Temperature Spanning 1753 to 2011, Geoinformatics & Geostatistics: An Overview, 01, https://doi.org/10.4172/2327-4581.1000101, 2013a.

Rohde, R., Muller, R., Jacobsen, R., Perlmutter, S., and Mosher, S.: Berkeley Earth Temperature Averaging Process, Geoinformatics & Geostatistics: An Overview, 01, https://doi.org/10.4172/2327-4581.1000103, 2013b.

Sippel, S., Kent, E. C., Meinshausen, N., Chan, D., Kadow, C., Neukom, R., Fischer, E. M., Humphrey, V., Rohde, R., de Vries, I., and Knutti, R.: Early-twentieth-century cold bias in ocean surface temperature observations, Nature, 635, 618–624, https://doi.org/10.1038/s41586-024-08230-1, 2024.