


DATA ARTICLE OPEN ACCESS

GloSAT LATsdb: A Global Compilation of Land Air Temperature Station Records With Updated Climatological Normals From Local Expectation Kriging

Michael Taylor¹ | Timothy J. Osborn¹  | Kathryn Cowtan² | Colin P. Morice³ | Philip D. Jones¹ | Emily J. Wallis¹ | David H. Lister¹

¹Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, UK | ²Department of Chemistry, University of York, York, UK | ³Met Office, Exeter, UK

Correspondence: Timothy J. Osborn (t.osborn@uea.ac.uk)

Received: 15 March 2025 | **Revised:** 11 July 2025 | **Accepted:** 16 July 2025

Funding: This work was supported by Natural Environment Research Council (NE/S015582/1 and NE/S015566/1).

Keywords: climatology | CRUTEM | GloSAT | kriging | land air temperature | normals | weather stations

ABSTRACT

To accurately determine multi-centennial trends in climate data records of the Earth's surface temperature, measurements are commonly analysed in the form of anomalies relative to a climatological reference period such as the World Meteorological Organization (WMO) 1961–1990 baseline. One of many climate-monitoring challenges is that weather records of land surface temperature can be short, typically of the order of several years or decades, and often do not sufficiently overlap the reference period to allow calculation of the climatological normals needed to convert the observations to anomalies. Moreover, the volume of records of this type is increasing due to the rescue of early (pre-baseline) instrumental paper-based records and the growing prevalence of newer (post-baseline) weather stations. To address this, we apply a method to estimate the climatological normal for each calendar month of temperature time series that do not have sufficient data during the baseline period, using an approximation to local expectation kriging with station holdout (LEK). This exploits the information in neighbouring time series to estimate the expected mean level of short series of observations. We apply the method to a global database of monthly land air temperature at 11865 stations based on CRUTEM5 but with the acquisition of an additional 1233 station series including some

Dataset Details:

This dataset is called GloSAT (Global Surface Air Temperature) LAT (land air temperature) sdb (station database), and comprises monthly mean temperature observations and climatological normals for meteorological stations worldwide. This release, v1.0, contains data for 11,865 stations and extends the CRUTEM5 1850–2021 station data (Osborn et al. 2021) back to 1781. Mid-latitude stations have been adjusted to compensate for some thermometer exposure changes. Many stations with either no data or incomplete data during the 1961–1990 baseline period have their climatological normals and standard deviations estimated via a kriging algorithm. The dataset is available from Zenodo at doi.org/10.5281/zenodo.14888902 and the Climatic Research Unit at crudata.uea.ac.uk.

Identifier: (DOI) [10.5281/zenodo.14888902](https://doi.org/10.5281/zenodo.14888902)

Creator: Climatic Research Unit (CRU), School of Environmental Sciences, University of East Anglia, Norwich, UK.

Dataset correspondence: t.osborn@uea.ac.uk

Title: GloSAT LATsdb v1.0: a global compilation of land air temperature station records with updated climatological normals from local expectation kriging.

Publisher: Zenodo.

Publication year: 2025.

Resource type: Dataset.

Version: 1.0.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 Crown copyright and The Author(s). *Geoscience Data Journal* published by Royal Meteorological Society and John Wiley & Sons Ltd. This article is published with the permission of the Controller of HMSO and the King's Printer for Scotland.

that extend back to 1781, and with mid-latitude stations adjusted for exposure bias arising from the transition to Stevenson screens. We evaluate the LEK-based normals using climatological normals calculated directly from the station observations. Using this method, we obtain estimated normals for 2699 stations that did not previously have normals and we improve the estimated normals for a further 2611 which had previously been estimated from incomplete data. Finally, we demonstrate how incorporating these thousands of previously unused station observation fragments affects hemispheric temperature averages. Pre-1850 data—primarily from Europe—show a modest warming trend but pronounced multidecadal variability that is greater than after 1850. The additional stations improve spatial coverage by a few percent in recent decades and raise pre-1860 Northern Hemisphere temperature estimates by approximately 0.1°C.

1 | Introduction

Weather stations measuring air temperature on land are at different elevations above sea level, and different countries calculate average monthly temperatures using different methods and formulae. To avoid biases that could result from these differences, monthly average temperatures are converted to anomalies from a climatological reference period where the coverage is best (see Appendix A for a discussion of the history that led to the WMO using 1961–1990 as the preferred reference period). This entails computation of the climatological reference period average for each station where the average temperature for each calendar month is called a climatological normal (more commonly referred to as a ‘normal’). Because many stations do not have complete records for the 1961–1990 period, several methods have been developed to estimate 1961–1990 averages from neighbouring records or using other sources of data (see Osborn and Jones 2014 and references therein).

A new methodology is needed to compute climatological normals for short, or long but gappy, instrumental measurement time series of land surface air temperature that do not overlap the climatological reference period sufficiently or at all. Short time series are a common occurrence. Early observers often carefully recorded temperature measurements several times a day continuously over the course of several years at a single location (see for example the records taken by the amateur observer Dr. E. A. Holyoke from 1789 to 1826 in Massachusetts, USA, described in van der Schrier and Jones 2008). These data records provide snapshots of the weather centuries ago during the pre-industrial period and could be important anchors for long-term trend calculations. In recent decades, automatic weather stations (AWS), often located in inhospitable environments such as high mountains (Matthews et al. 2020) or in Antarctica (Wang et al. 2023), provide series that fill in spatial gaps but with little or no data during the reference period. In other cases, an instrument may operate for only a limited time or have had to be moved or replaced. In Antarctica, for example, temperature measurements at the Halley Research Station have moved with the station itself (King et al. 2021). The same is true of instruments at drifting ice stations in the Arctic (Przybylak and Wyszynski 2020).

As a result, the time series at a given geolocation, while important in terms of spatial sampling of the global temperature, can be very short. Even long time series may be punctuated by gaps in their record arising from an array of causes such as instrument failure or local impediments associated with staffing or regional

conflict. Whatever the cause, whenever available observations do not sufficiently span the 1961–1990 climatological reference period, absolute temperature measurements are unable to be accurately converted to anomalies and used in many of the gridded temperature analyses such as HadCRUT5 (Morice et al. 2021), impacting the local accuracy of the estimated temperature anomalies. Some global temperature analyses already address this issue by using methods that do not rely on converting each station record to anomalies from a fixed reference period. For example, the Berkeley Earth land temperature dataset (Rohde et al. 2013) adjusts the mean level of each station to an expected temperature anomaly based on other stations in the region.

There are a number of ways that climatological normals might be estimated when data gaps prevent their direct calculation from the station's data. These typically combine the incomplete station data with information on local temperature changes from another source, such as reanalysis data (Way and Bonnaventure 2015; Gillespie et al. 2021), calibrated palaeoclimate data (e.g., Briffa et al. 2013), climate model simulations (Mahony et al. 2022) or gridded observational datasets (Jones and Moberg 2003; Way and Bonnaventure 2015). However, each method is sensitive to the choice of the alternative data source and is limited by the reliability of the alternative data source for representing local temperature change. What is needed is a robust way to calculate climatological normals directly from available observations, such as regression against well-correlated neighbouring stations, as used by Perry and Hollis (2005) to estimate normals for UK stations. Such an approach is developed and evaluated here. The rationale for doing this is two-fold: first, to increase the spatio-temporal coverage of station anomalies contributing to gridded global surface temperature anomaly datasets, and second, to improve estimates of existing normals where they are currently estimated from incomplete data during the climatological reference period.

The rest of this data paper is organised as follows. In Section 2, we explain how we produced the dataset. We present the input and auxiliary datasets used and the quality control methods we have applied. We then describe the workflow for computation of the climatological normals for the global station database using local expectation kriging (LEK), together with the parameterisation and methodology we have adopted for its implementation. In Section 3, we present an evaluation of the outcomes of the dataset production approach, including the climatological normals produced with (LEK) and the coverage improvements arising from our approach. In Section 4, we describe the location and format of the dataset. In Section 5, we outline how the dataset might be used for climate applications.

2 | Dataset Production

Figure 1 presents a flowchart of the steps and procedures used to produce GloSAT LATsdb.

Newly acquired GloSAT station temperature observation time series were merged with the latest CRUTEM5 station database (Osborn et al. 2021) to extend the record back to 1781. This station database was exposure bias adjusted (EBA) for known screen type changes using an exposure bias model for stations in the mid-latitudes (Wallis et al. 2024). Quality control was then performed and a first reliable year (FRY) flag was applied for each station. Normals and standard deviations (SD) were then computed following the CRUTEM5 approach (Osborn et al. 2021) and outliers were detected and flagged (Osborn et al. 2021) to produce the primary input dataset GloSAT_eba. This was then processed with local expectation kriging (LEK: see Section 2.5). Kriging was applied to individual clusters of the global station archive, partitioned by a hierarchical cluster analysis (Section 2.6). The resultant LEK processed clusters were then merged. Finally, normals and SD estimated from the LEK values were combined with those available in GloSAT_eba to infill gaps in the climatological normals and

SDs for the station database. Together, these form the GloSAT LATsdb dataset.

2.1 | Input Weather Station Data

The CRUTEM5 1850–2021 station database (version CRUTEM5.0.0.1.0, updated from Osborn et al. 2021) was extended back to 1781 using early station records already in the Climatic Research Unit (CRU) archives. The CRU archives held one or more pre-1850 observations for 335 CRUTEM5 stations, and these had already been subject to simple quality checks at the time they were acquired; where there was some doubt over the homogeneity of early data, a ‘first reliable year’ had been specified to indicate that earlier data should not be used. This is preferable to removing the earlier data, since it allows future studies to analyse the earlier data and potentially improve their homogeneity.

The total of 10,632 weather station records (excluding 7 with incomplete location information) in the CRUTEM5 sdb was subsequently increased by 1233 through new acquisitions, and many more were improved by acquiring data to fill in short data gaps

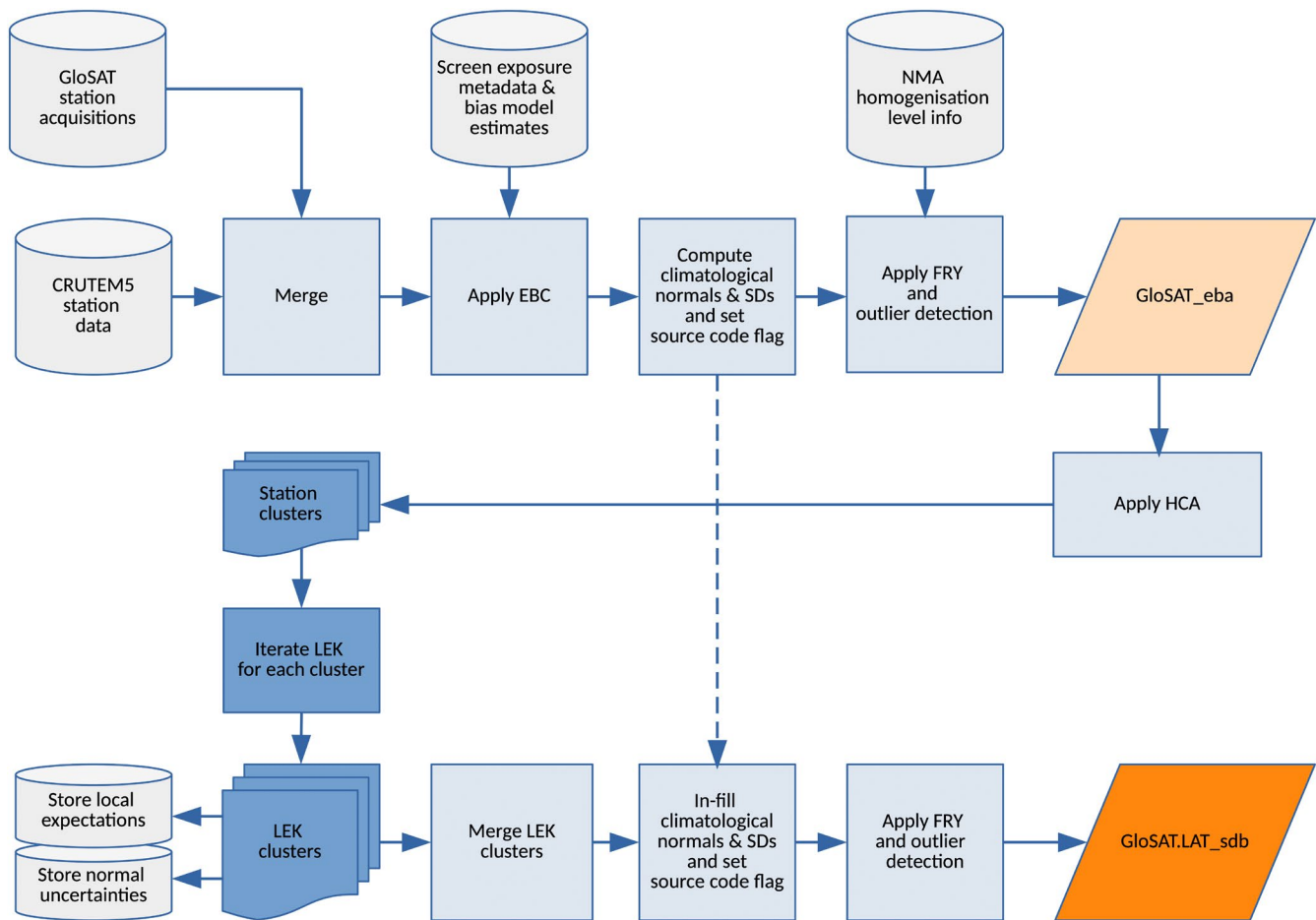


FIGURE 1 | Processing flowchart for production of the GloSAT land air temperature (LAT) station database (sdb). Standard software flowchart symbols are used (drum: Database; rectangle: Process; stacked shape: Process applied multiple times to data subsets; parallelogram: Output). New GloSAT acquisitions are merged with the CRUTEM5 station database to back-extend the record to 1781. This is exposure bias adjusted (EBA; see Wallis et al. 2024) and quality controlled by applying a first reliable year (FRY) flag and removing outliers (Osborn et al. 2021) to produce the primary input dataset GloSAT_eba. Hierarchical cluster analysis (HCA) is used to partition this global archive prior to processing with local expectation kriging (LEK) from which climatological normals and standard deviations are computed to generate the output dataset GloSAT LATsdb.

or to replace them with series that have improved homogeneity. These acquisitions and updates are summarised in Table 1. The spatio-temporal coverage of stations in the extended database is shown in Figure 2.

2.2 | Exposure Bias Adjustments

The transition from earlier, non-standard, thermometer screen types to the Stevenson screen standard has been found to introduce an exposure bias in many early temperature observations (Parker 1994; Wallis et al. 2024). These biases were primarily introduced due to differences in the amount of solar radiation affecting the thermometer between screen types, with the largest biases generally associated with the transition from free-standing ‘open’ exposures during summer months. Wallis et al. (2024) characterised the exposure bias using 54 parallel measurement series and developed regression models to predict the seasonal bias for differing classes of exposure as a function of annual temperature, top of atmosphere and/or surface shortwave downward solar radiation. The Wallis et al. (2024) exposure bias adjustments were applied to $n = 1960$ mid-latitude stations at the monthly timescale in the GloSAT station database, to produce GloSAT_eba. The adjustments were applied to stations whether or not they had been homogenised at source, unless the station was known to have already been corrected specifically for exposure bias. The bias adjustments vary in magnitude and sign across months and with pre-Stevenson screen exposure types, but the average effect is that small positive adjustments are applied to Northern Hemisphere stations pre-1880, and larger negative mean adjustments are applied to Northern and Southern Hemisphere stations between 1882–1934 and 1856–1900, respectively. The exposure bias adjusted dataset GloSAT_eba is the primary input to the LEK processing chain shown in Figure 1.

2.3 | Quality Indicators

The CRUTEM approach to homogenisation of station data is adopted for GloSAT, namely that we do not apply global, statistical algorithms to identify and correct for inhomogeneities; instead we preferentially use data series that have been homogenised at source (by national or regional initiatives) and we represent remaining inhomogeneities in the CRUTEM error model (e.g., the urbanisation, exposure and homogenisation error terms; see Morice et al. 2021). This approach complements other datasets (e.g., NOAA: Huang et al. 2020; or Berkeley Earth: Rohde and Hausfather 2020) which apply homogenisation algorithms globally, allowing multiple global temperature datasets to sample some of the structural uncertainty arising from the choice of how to deal with inhomogeneities.

For GloSAT LATsdb we have, however, made a simple assessment of the expected level of homogenisation for stations within the database. This is based on the source codes provided for each station (a caveat being that some stations may have been derived from multiple sources over the 40 years that the CRUTEM station database has been maintained and updated, so the source code provides only a broad indication of

the original or main source in some instances, and the source of the most recent values in other cases). For each data source, we conducted a broad-brush assessment of whether we judge that the data provider carried out a homogeneity assessment (and possibly applied adjustments to reduce inhomogeneities). Based on this, we assigned one of five homogenisation indicators to each data source (and for some sources we also provide the time period over which homogeneity has been assessed, so observations outside that period are more likely to contain uncorrected inhomogeneities).

HOM00 indicates that the source either did not assess homogeneity or it is unknown if they did. HOM01 is assigned where it is unknown if the source assessed the homogeneity but the data are from an authoritative source, including a national meteorological service (NMS) or from World Weather Records (WWR). The other three codes apply to stations that have had some inhomogeneities identified and corrected: at the source but the method used is unknown to us (HOM02), via documented methods (HOM03) or by CRU, after receiving the data, using documented methods (HOM04). For the GloSAT LATsdb, this broad-brush homogeneity assessment is not used further (other than to partially inform the assignment of a first reliable year (FRY) for some station time series), but it is provided because it may be useful for users of the database or for future homogeneity assessments.

After assembly of the station database, the same checks for physically implausible values (accounting for station latitude, elevation and month of the year) and statistical outliers (based on the interquartile range but with a relaxed criteria if regional anomalies of the same sign occur) as described by Osborn et al. (2021) for CRUTEM5 were applied. We provide two versions of the station database: one with the full data and one with values that failed these checks removed.

2.4 | CRUTEM5 Normals

As part of CRUTEM5 processing (Osborn and Jones 2014; Osborn et al. 2021), climatological normals are computed—independently for each month of the year—directly from the station data when there are sufficient (≥ 15 out of 30) monthly observations during 1961–1990 for that month. For stations with insufficient observations to meet the minimum data criteria, some normals were obtained from external sources (mostly the WMO), from an earlier version of the dataset, or estimated by infilling from neighbours or from earlier reference periods with adjustments made for the expected difference (Osborn and Jones 2014), and flagged with an identifying normal category code (codes 2 to 5 in Table 2). Nevertheless, normals were not obtained or estimated for 2649 stations in CRUTEM5 and therefore none of those stations could contribute to the production of the CRUTEM5 gridded global dataset (Osborn et al. 2021). If we adopted the same CRUTEM5 approach with the additional stations assembled for GloSAT LATsdb, the number of series without normals (and therefore unused for producing the gridded GloSATref dataset; Morice et al. 2025) would have grown to 3430 (column GloSAT_eba in Table 5). The timing and location of the stations without normals is shown by white in the left-hand column of Figure 2.

TABLE 1 | Summary of new acquisitions and updates applied to CRUTEM.5.0.1.0 to create the GloSAT land surface air temperature station database (GloSAT LATsdb). 'New' indicates series that were not previously in CRUTEM sdb.

Region	Count	Source	Type
New acquisitions			
Global	439	NCEI World Weather Records (WWR)	New series
Global	149	NCEI Monthly Climatological Data for the World	New series
Canada	346	Environment Canada	New homogenised series
Germany	14	DWD	New series
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), Armagh (UK 1796–2021) Paris (France 1658–2019), Bordeaux (France 1851–2021) Perpignan (France 1836–2021), Gorkij (Russia 1881–1989) Tianjin (China 1890–2021), St Helena (1892–2021) Ascension (1923–2021), Nassau (Bahamas 1811–2021) Tenkodogo (Burkina Faso 1951–1991) Cucuta (Colombia 1961–2021) Antananarivo (Madagascar 1889–2021)
Routine updates			
Global	7810	CLIMAT	Updated series for 2020–2021
Australia	112	BoM	Updated ACORN2 series for 2019–2021
Canada	434	Environment Canada	Updated series and improved homogeneity
Chile	314	Chilean Centre for Climate & Resilience Research	Updated series for 2016–2021
Denmark, Faroes, Greenland	17	Danish Meteorological Institute	Updated series for 2017–2020
Iceland	127	Icelandic Meteorological Office & Trausti Jónsson	Updated series and improved homogeneity

(Continues)

TABLE 1 | (Continued)

Region	Count	Source	Type
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

Although the absolute number is highest in recent decades, as a proportion of the total available data it also has a peak (at about 20%) in the early 1800s.

We also note that where there are only partially complete (<30) observed values during 1961–1990 (but still ≥ 15 to allow a normal to be calculated) the estimated normal has increased uncertainty (Brohan et al. 2006) but less well known is that it can also result in a small underestimate of global-mean warming: Calvert (2024) quantifies this underestimate as 0.003°C . Furthermore, the externally-sourced (e.g., WMO) normals and those previously estimated by some type of infilling are not easily reproducible with new data and are considered to be less reliable (Brohan et al. 2006).

Taking these issues together, therefore, it is desirable to develop an approach that infills data gaps for the purpose of estimating normals which will improve the utilisation of stations without 1961–1990 data, improve reproducibility, and reduce both random and systematic (e.g., underestimated warming) errors. Data gaps throughout the full length of each station's timeseries could potentially be infilled to address issues caused by changing observational coverage over time. However, this coverage effect and associated errors are already accommodated in the HadCRUT5 analysis approach (Morice et al. 2021) so in the current study we limit our use of infilling to the purpose of making better estimates of station normals. The following section presents an approach for achieving this.

2.5 | Local Expectation Kriging (LEK)

In the current application, we want to determine a baseline for short station fragments that do not overlap with or are incomplete over the 1961–1990 baseline period. This will ensure that their inclusion will not bias the resulting record due to the appearance and disappearance of those stations. The baseline must align each fragment with the temperature field inferred from longer records, while still allowing the fragment to inform the temperature field about local and short-term variability.

Kriging is a spatial interpolation method that computes the best linear unbiased (minimum error variance) estimates at unsampled locations by weighting nearby observations based on spatial covariance and distances (Cressie 1990). For monthly mean near surface temperatures, the covariance function is modelled as an exponential kernel that depends on the great circle distance between pairs of stations. We chose Kriging because it is a well-established Gaussian process model that provides a simple

and mathematically well understood solution to estimating the value of a field at locations where no observations are available, and that accounts for both the information content of nearby observations and redundancy amongst closely spaced stations. A holdout approach is needed because we want to evaluate how well the observations at a given station can be estimated from surrounding stations. The Kriging estimator enables the approximate holdout calculation.

Thus, Kriging is employed to reconstruct an estimate of the temperature field at the location of each station using data from all other stations. In the case of GloSAT LATsdb, the temperature field at each monthly time step results from 11,865 stations located unevenly across the global land surface and with incomplete monthly observations spanning the years 1781–2021. Kriging allows the data gaps in the global station temperature record to be infilled. The data-driven gap-filling capability provided by Kriging thereby allows for both the estimation of previously absent climatological normals as well as the improvement of normals computed from formerly partially complete data.

2.5.1 | Kriging With Hold out

Conceptually, a target station can be held-out from a cluster of local stations, and a temperature series at that location can be reconstructed by kriging. This is referred to as 'kriging with hold out' (Pang et al. 2023). However, the kriging calculation requires a matrix inversion (with dimensions equal to the number of stations) in order to determine the amount of independent information a neighbouring station provides for estimating the temperature series at the desired location. This matrix inversion must be recomputed whenever the selection of included stations changes and requires a matrix inversion for every station and every month in the reconstruction. While this is practical for small station networks, when working with hundreds or thousands of stations it becomes impractical.

A more computationally tractable approach for large networks of stations like CRUTEM5 or GloSAT LATsdb is to hold out batches of (say) 10% of the stations at a time and compute temperature series for those stations from the remaining stations in the network. This approach reduces the computational overhead to 10 matrix inversions per time step, but at the cost of losing some station information for reconstructing the time series at a given location. Selection of the batches is, however, problematic as neighbouring stations should belong to different batches, which must change over time as the available stations change.

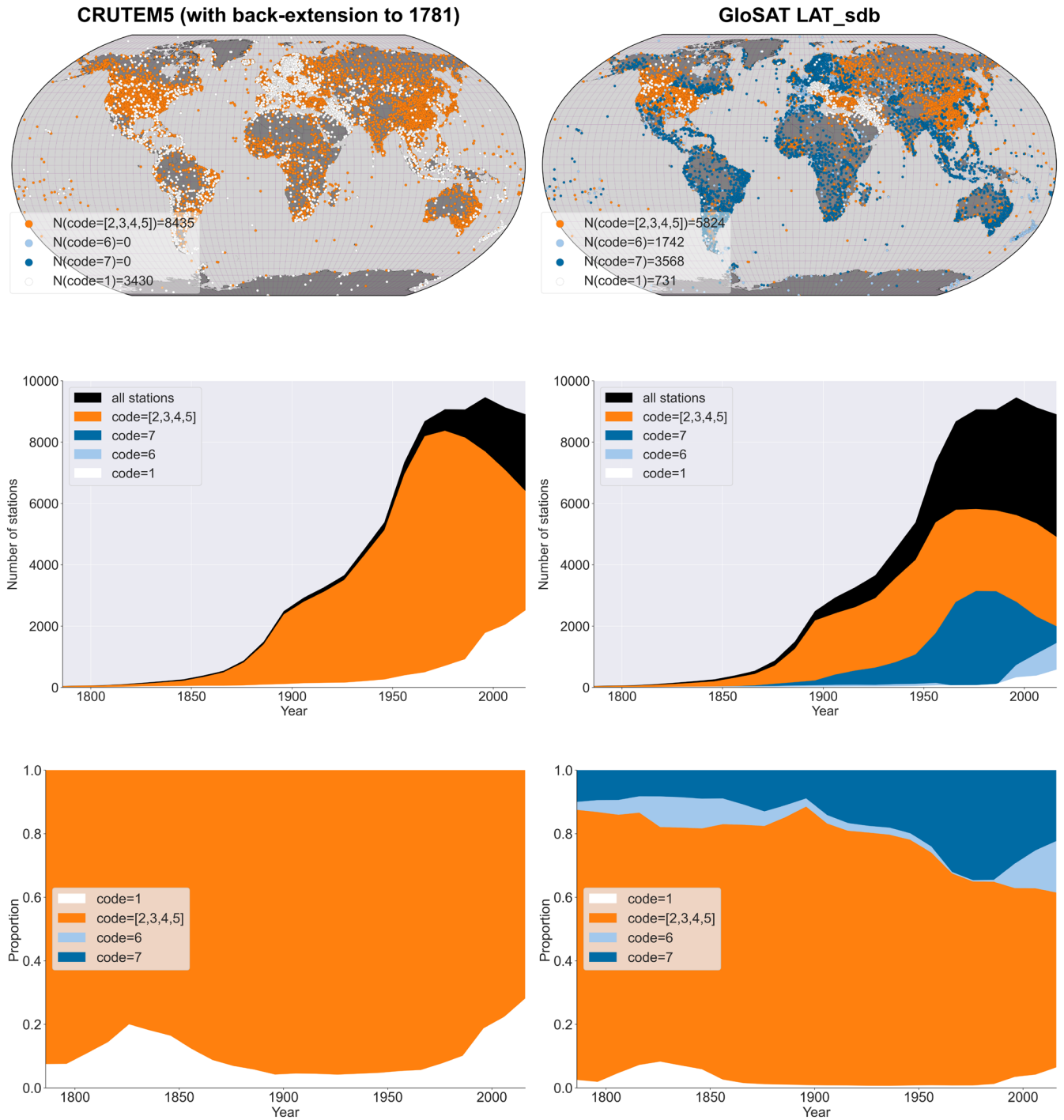


FIGURE 2 | Spatio-temporal distribution of CRUTEM5 (left column) and GloSAT LAT (right column) stations with and without climatological normals, to illustrate the improvements enabled by the local expectation kriging. Top row: Spatial distribution of stations coloured by normal category code; middle row: Time-evolving decadal counts of station normals by category code; bottom row: Time-evolving decadal percent of station normals by category code. The category codes for the normals are given in Table 2: Stations without normals (code 1 = white); with normals calculated directly from observed values or indirect sources (codes 2 to 4 = orange); with normals calculated by combining some observations with some LEK values (code 7 = dark blue); with normals calculated solely from LEK values (code 6 = pale blue).

2.5.2 | Approximate Kriging With Hold out

Here, we present a new approach to this problem that provides an approximation to the hold out approach using a single matrix inversion for each time step. In kriging, the linear estimator \hat{y}_0 ('local expectation') at a given station location x_0 is equal to the

linear weighted average of the measured values y_i at the other station locations x_i :

$$\hat{y}_0 = \sum_{i=1}^n w_i y_i \quad (1)$$

TABLE 2 | Normal category code definitions for climatological normals and standard deviations (SD) together with the estimate of the uncertainty on the normal. Category codes 6 and 7 are new compared with CRUTEM5.

Code	Normal & SD	Normal uncertainty= $f \cdot SD$
1	Missing	
2	Estimated using previous infilling methods	As HadCRUT5, $f = 1 / \sqrt{15}$
3	WMO	As HadCRUT5, $f = 0.3$
4	Calculated directly from station data	As HadCRUT5, $f = 1 / \sqrt{n}$; $n \geq 15$ values in 1961–1990
5	Taken from previous dataset version	As HadCRUT5, $f = 1 / \sqrt{n}$; n values in 1961–1990
6	Estimated solely from LEK	$f = 1 / \sqrt{15}$
7	Calculated from a combination of station data and LEK	$f = 1 / \sqrt{15}$

where n is the number of stations included. The weights w_i are obtained by minimising the expected value of the estimation variance (i.e., the square of the error from the unknown true value y_0). In the approximate hold out approach, the kriging calculation is performed as usual including all of the stations present in a given timestep, and then the weight for the station itself is set to zero and the remaining weights scaled up so that they sum to 1, to retain the unbiasedness of the estimator. The resulting weights are not identical to those obtained by holdout, but retain the key features of weighting stations according to their distance from the location of interest while down-weighting clusters of stations which do not contribute independent information. The weights which minimise the error term $E\{(y_0 - \hat{y}_0)^2\}$ are determined by inverting the covariance matrix of the stations and multiplying by the vector of covariances between the station locations and the station of interest:

$$w_i = \sum_j [\text{Cov}(y_i, y_j)]^{-1} \text{Cov}(y_j, y_0) \quad (2)$$

An additional equation, $\sum_i w_i = 1$, is included using the method of Lagrange multipliers to ensure that the resulting estimator in Equation (1) is unbiased (Wackernagel 2003). Since the vector of covariances $\text{Cov}(y_j, y_0)$ is a row of the matrix $\text{Cov}(y_i, y_j)$, the resulting vector of weights always has $w_0 = 1$ and $w_{j>0} = 0$, which provides no information as to how to weight the remaining stations having removed station zero. In the absence of errors this tells us that the station itself is the best estimator of the temperature series at the location of the station, however this provides no information as to how to weight the remaining stations if that station is removed.

This can be addressed by introducing an error term to reflect the fact that, due to localised effects such as exposure and measurement error, station data are never perfect observations of the true temperature field even at the station's own location. We achieve this by inflating the diagonal terms of the covariance matrix in Equation (2) by adding a value τ^2 representing the effect of additive measurement and local representativity errors that are uncorrelated between stations, giving rise to:

$$w_i = \sum_j [\text{Cov}(y_i, y_j) + \tau^2 \delta_{ij}]^{-1} \text{Cov}(y_j, y_0) \quad (3)$$

This leads to a vector of weights which includes contributions from all local stations, so that re-weighting to remove the station itself becomes possible. It has the additional benefit of tending to stabilise the matrix inversion.

The value of τ would normally be the uncertainty in the observations; however, this depends on the covariance matrix also being on an absolute scale. With no noise term, the expected values determined by the kriging calculation are invariant to the scale of the covariance matrix. The scale of station variability is not otherwise needed for this calculation so the kernel was chosen to have a maximum value of 1 at the station location. In this case τ becomes the station noise as a fraction of station variability. A value $\tau = 0$ corresponds to the error free case, which does not allow the estimation of station weights, while a value $\tau \gg 1$ leads to a covariance matrix that is approximately diagonal and so weights the remaining stations according to their covariance with the station of interest, ignoring any dependence between the remaining stations (e.g., due to nearby stations being mostly clustered in one direction). The value of τ was therefore chosen empirically to be just large enough to lead to a stable matrix inversion and reweighting even for sparse station networks: this criteria led to the choice of $\tau = 0.1$. This procedure for the fast approximation of kriging with hold out will be referred to throughout the remainder of the work as Local Expectation Kriging (LEK).

Near surface (2 m) temperature observations at weather stations show correlations which decrease with distance with a characteristic length scale L which depends on region and direction, as well as whether observations are sampled at the monthly, yearly, or decadal timescale (Jones et al. 1997). We model the spatial distribution of covariance in Equation (3) with an exponential function (because it is controlled by a single characteristic distance parameter and better reflects observed patterns of station covariance than other commonly used kernels):

$$C_{ij} = \text{Cov}(x_i, x_j) = C(d) = C_0 e^{-d(x_i, x_j)/L} \quad (4)$$

where C_0 can be the variance of the data (i.e., of the spatial field at station locations), but is arbitrarily set to one because the weight calculation is invariant to scaling of the covariance function. Another reason for this choice of kernel is that any Gaussian process model whose covariance function depends only on $x_i - x_j$, like the exponential function in Equation (4), is guaranteed stationarity (Bishop 2006). We prescribe $L = 900$ km a priori, guided by the empirical findings of Jones et al. (1997) and Cowtan et al. (2018).

The matrix C_{ij} is of size $n \times n$ and takes $\mathcal{O}(n^3)$ operations for its inversion, so when dealing with large global station networks (here we have $> 10,000$ stations) the matrix inversion is still impractical (Barber 2012). However, given that $L = 900$ km, stations far away from the location whose temperature is being estimated have negligible influence. Therefore, we apply LEK to smaller clusters of stations separately, with the clusters chosen from their geographical distribution using Hierarchical Cluster Analysis (HCA) as described in the next section. LEK provides temperature estimates at each station within a cluster, and repeating the analysis across all clusters provides information for the global network of stations.

2.6 | Hierarchical Cluster Analysis (HCA)

We use agglomerative hierarchical clustering with complete-linkage to build the clusters needed to partition the global station network (Everitt et al. 2011). Each station initially starts in its own cluster (a single ‘leaf’ of the dendrogram). The two clusters separated by the shortest distance are then merged into one cluster. This step is repeated many times until all stations are grouped into the target number of clusters.

Different clustering methods determine the distance between clusters in different ways. For complete-linkage clustering, the distance between two clusters is given by the distance between the farthest separated pair of stations (one from each cluster). The method is, therefore, also known as ‘farthest neighbour clustering’. At each step, the pair of clusters with the shortest link (i.e., the shortest of all the farthest neighbours between all cluster pairs) are merged. The method is robust because it considers all pairs of stations within all pairs of clusters at each step. This does make it computationally expensive, though we minimise this by pre-calculating the distance matrix for all station pairs and merging its rows and columns as clustering progresses.

To assess the runtime needed to perform LEK on the global station archive on the JASMIN high performance computing (HPC) facility at the Centre for Environmental Data Analytics (CEDA), we took a station dense part of the network comprising several thousand stations in the contiguous United States and created clusters increasing in size from 2 to 2048 stations and timed the LEK computation.

We found that there was an approximately linear relation between doubling the cluster size and a five-fold increase in runtime. As mentioned in Section 2.5.2, this is due to the computational load associated with inversion of the linear system of equations in Equation (3). To limit parallel LEK runs (one run per cluster) at HPC facilities to $\mathcal{O}(\text{day})$ it was necessary to limit the maximum number of stations in a given cluster to $n_{\max} = 700$ stations. Since the station archive GloSAT LATsdb comprises 11,865 stations, an equipartition with $n = 700$ stations per cluster would correspond approximately to 17 clusters. The spatial degrees of freedom associated with the instrumental record is reported to be ~ 40 (Jones et al. 1997), suggesting a smaller value of n_{\max} may be reasonable.

Another consideration is that the ‘rich get richer’ behaviour inherent to agglomerative clustering tends to lead to uneven cluster sizes (though complete-linkage clustering is less susceptible to this than the single-linkage approach). To account for this, those clusters exceeding a maximum cluster size n_{\max} were iterated again with agglomerative clustering until no cluster contained more than n_{\max} stations. For $n_{\max} = 700$, the resulting partition of the global archive was found to contain 50 clusters while for $n_{\max} = 400$, 73 clusters were obtained.

To assess the stability of this clustering approach and the variability in the number of stations per cluster, a number of experiments were conducted. Table 3 summarises the HCA agglomerative clustering statistics for initial target numbers of clusters ranging from 10 to 50 and presents the total number of clusters after the 2nd iteration together with the median number of stations per cluster and the interquartile range (IQR). Table 3 shows that $n_{\max} = 400$ leads to a more stable number of clusters (~ 70) independent of the initial target number of clusters. The IQR is also stable, consistently equal to approximately double the median number of stations per cluster. Limiting the cluster size to 700 stations (close to the maximal computational load) leads to less cluster fragmentation and less inter-cluster variability in the number of stations per cluster but a less stable result as a function of the initial target number of clusters. Visual inspection of the spatial distribution of the clustering for an initial target of 30 or 40 clusters led us to favour 40 as this clustering matches more closely the spatial extent of Köppen climate zones (Beck et al. 2018) and known national meteorological service networks. Figure 3

TABLE 3 | Hierarchical Cluster Analysis statistics for initial target numbers of clusters ranging from 10 to 50.

1st iteration		2nd iteration: Max (cluster size) = 400		2nd iteration: Max (cluster size) = 700		
Ncluster	Ncluster	Median	IQR (stations/	Ncluster	Median	IQR (stations/
		(stations/	cluster)		(stations/	cluster)
		cluster)	cluster)		cluster)	cluster)
10	140	37	101	68	90	251
20	77	117	244	32	332	258
30	75	113	238	44	251	360
40	73	126	223	50	209	299
50	76	114	224	58	152	244

Note: In the first set of experiments (columns 2–4) the maximum cluster size allowed is 400 stations, and in the second set (columns 5–7) the maximum allowed is 700 stations. In both cases, a 2nd iteration of HCA is applied to any clusters exceeding this maximum in the 1st iteration. The total number of clusters after the 2nd iteration is reported together with the median number of stations per cluster and its interquartile range (IQR).

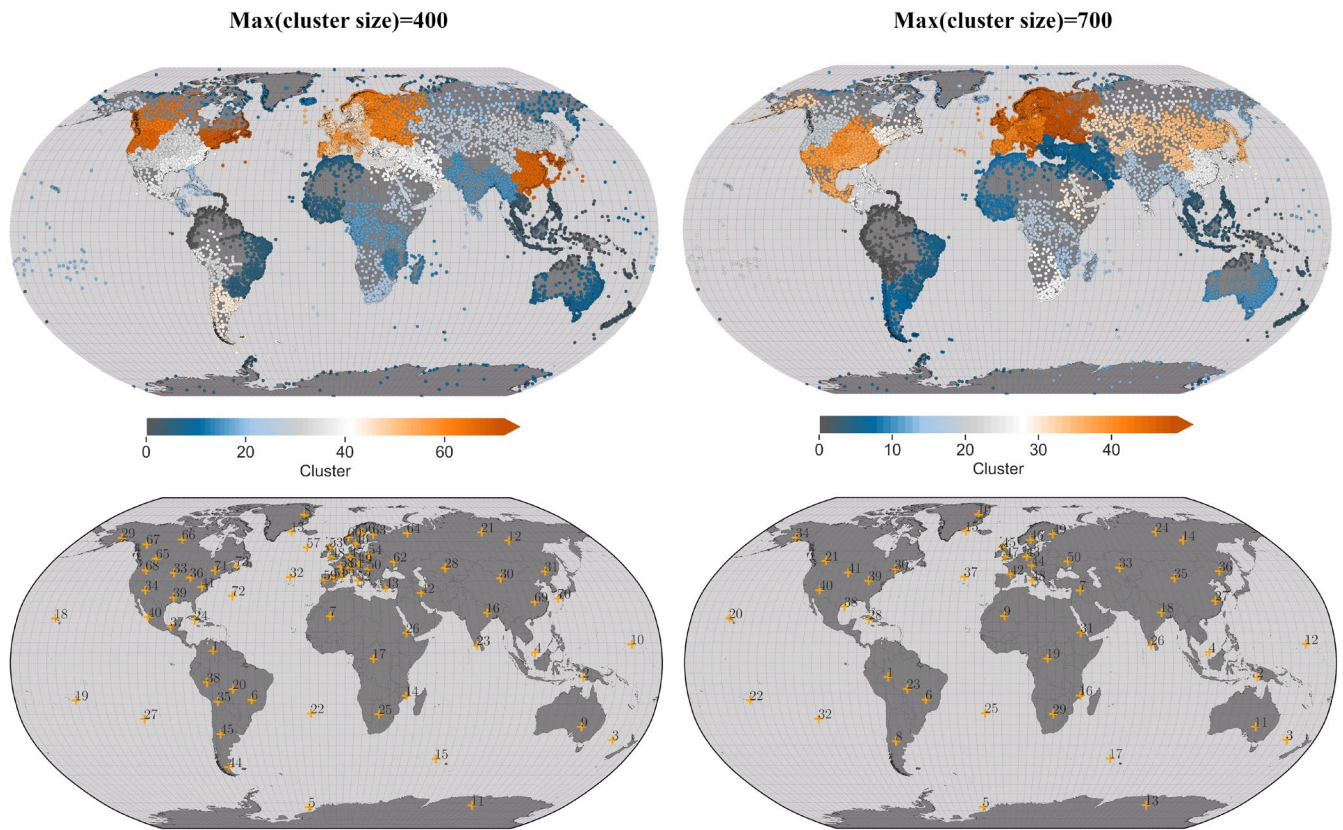


FIGURE 3 | Spatial distribution of the station clusters (top) and their centroid locations (bottom) for two partitions where the maximum cluster size is limited to (left) 400 stations (73 clusters) and (right) 700 stations (50 clusters).

shows the spatial distribution of the clusters and their centroid locations for the two partitions where the maximum cluster size is limited to 400 stations and 700 stations.

To perform the clustering, we therefore initialised the agglomerative HCA algorithm with a target of 40 clusters and set $n_{max} = 400$ in the second iteration. We used the python scikit-learn AgglomerativeClustering method applied to the distance matrix D to build the model as follows:

```
model = AgglomerativeClustering(n_clusters=40,
affinity='precomputed', linkage='complete',
distance_threshold=None, compute_distances=True,
compute_full_tree=True).fit(D)
```

In the next sub-section we address the issue of kriging across cluster boundaries.

2.6.1 | Cluster Halo Analysis

Since kriging is driven most strongly by neighbouring station timeseries, it is necessary to assess the impact of cluster boundaries—i.e. the effect of a long neighbouring reference series ending up in an adjacent cluster. To assess this, for each cluster, we use the distance matrix to create a halo of stations surrounding the cluster extending beyond the cluster

boundary with stations ranked in increasing order by distance. A second run of LEK for each of these clusters augmented with halos was then made. Then, the root mean squared (RMS) difference between station normals computed by LEK for clusters with the halo and the same clusters without the halo is computed. This allows us to quantify the size of the cluster boundary effect.

Since the clusters projected onto the globe are non-circular, it is not possible to construct halos by radially extending the clusters by a constant distance (for example, the spatial correlation length scale L). Furthermore, the variability in the number of stations per cluster arising from the varying spatial density of stations worldwide, means that a data-adaptive approach is needed to set the halo size. A robust solution is to refer to the distance matrix with stations ranked by distance while also setting an upper bound to the allowed number of stations added to the halo because of computational limits.

Stations in the halo are then selected according to the following criteria:

1. Their distance from the edge of the cluster is less than the LEK correlation length scale $L = 900$ km.
2. The closest 100 stations ranked in increasing order by distance from the cluster.
3. The maximum total cluster + halo membership is limited to 700 stations.

In Section 3.2 of the Dataset Evaluation, we present the results of a sensitivity analysis for the cluster-halo method.

2.7 | Infilling of Climatological Normals and Standard Deviations

With a small number of exceptions (6% of stations—see Section 3.4), LEK provides monthly temperature estimates for each station. For the purpose of evaluating the use of these LEK values to calculate station normals, their average over 1961–1990 for each calendar month was used (a ‘LEK-only’ normal). However, for the purpose of creating the normals for the final GloSAT LATsdb dataset, we retain the ethos of using directly observed data to the full extent possible. Therefore, for stations with complete 1961–1990 data, we calculate the station normal as the average of those complete values. For stations with partial 1961–1990 data, we infill the individual gaps with LEK monthly estimates and then compute the normal as the average of the now complete values. For stations without any observations during 1961–1990, the ‘LEK-only’ normal is used. For stations without LEK estimates, the normal is either left as missing, calculated from incomplete actual values or previous estimates (from WMO etc.) are retained.

Table 4 summarises these criteria. Although the LEK method also provides an associated uncertainty for the ‘LEK-only’ normal, we have taken a conservative approach that follows closely the previous CRUTEM5 error model, which assumes that normals not calculated directly from observed values are more uncertain (being similar to the uncertainty if they were calculated from 15 or fewer values). Table 2 provides the normal category code definitions and the estimation of the uncertainty on the normals. Category codes 6 and 7 are new compared with CRUTEM5.

The HadCRUT5 Analysis also requires estimates of the monthly temperature standard deviation (SD) for each station, as part of the error model (Morice et al. 2021). Following the long-standing CRUTEM approach (Osborn and Jones 2014), SD is calculated using all observations during 1941–1990, provided a minimum

of 15 values is available. If fewer are available, the period is expanded to encompass the full length of the dataset, but if fewer than 10 observed values are available, then the SD of the LEK estimates is used instead.

3 | Dataset Evaluation

In this section we present and assess the outcomes of the dataset production approach, considering first the agreement between ‘LEK-only’ normals and those calculated directly from co-located data and any dependence on latitude or temperature. Then we present a sensitivity analysis of the hierarchical cluster analysis with halos and an exploration of the biases avoided by infilling incomplete 1961–1990 data. Finally, we show the impact of incorporating the new station data and the LEK-based normals on the spatio-temporal coverage of the data and on large scale hemispheric averages of the mean temperature. Unless otherwise stated, LEK was applied to the 73 clusters obtained with a maximum cluster size of 400 and two iterations of HCA (i.e., those shown in the left column of Figure 3).

3.1 | Evaluation of LEK-Based Normals

Figure 4 shows by scatter plots and regression lines the overall level of agreement between co-located ‘LEK-only’ normals and those calculated directly from the GloSAT_eba 1961–1990 observations (which in many cases are incomplete, though here we require at least 15 values). Although there are a small number of cases with larger differences, the overall agreement is very strong and there is no systematic bias. The normals averaged across each calendar month (Figure 4) show close to zero bias between ‘LEK-only’ and ‘data-only’ normals.

For the calculation of standard deviations, Figure 4 (bottom row) shows there is a slight bias towards lower ‘LEK-only’ SDs than ‘data-only’ SDs across the whole set and in most individual months, but it is very small. The large scatter between ‘LEK-only’ and ‘data-only’ SDs is the reason why SDs calculated directly from the data are preferred (Section 2.7).

TABLE 4 | Criteria for calculating climatological normals, applied from left to right.

LEK normal?	Number of 1961–1990 observed values	Existing normal?	Final normal	Category code
No	$n \geq 15$	N/A	Calculate from the available 1961–1990 observed values	4
No	$n < 15$	Yes	Use existing normal and its category code	2, 3, 4 or 5
No	$n < 15$	No	Missing	1
Yes	$n = 30$	N/A	Calculate from the complete 1961–1990 observed values	4
Yes	$n = 0$	N/A	Use LEK normal	6
Yes	$0 < n < 30$	N/A	Calculate from 1961 to 1990 timeseries that infills observed gaps with LEK estimates	7

TABLE 5 | Station counts per normal category code.

Code	Category of normal	Number of stations	
		GloSAT_eba	GloSAT LATsdb
1	Missing	3430	731
2	Estimated using previous infilling methods	91	11
3	WMO	26	1
4	Calculated directly from station data	8306	5806
5	Taken from previous dataset version	12	6
6	Estimated solely from LEK	0	1742
7	Calculated from a combination of station data and LEK	0	3568
Total [2–5]	All stations with non-LEK normals	8435	5824
Total [2–7]	All stations with normals	8435	11,134
Total	All stations	11,865	11,865

The differences between the ‘LEK-only’ and ‘data-only’ normals are further explored in Figure 5, which shows the LEK minus GloSAT_eba difference for co-located climatological normals binned by: (a) latitude and (b) temperature. These tests were made to explore the possibility of greater bias or greater uncertainty in normals at latitudinal extremes, or at sparsely observed latitudes such as the ocean-dominated southern mid-latitudes. Each bin contains LEK–data comparisons for many stations and these are summarised by the means, interquartile range, and 95% range of the differences, along with the root mean squared (RMS) difference. There are no clear dependencies of the mean bias on either station latitude or station normal temperature, and 95% of individual differences are less than 0.5°C at almost all latitudes and temperatures (see the 2.5%–97.5% ranges in Figure 5). There are some localised maxima in the RMS differences (e.g., just south of the equator and also where mean temperature is just below freezing) but the interquartile ranges are not notably larger, so the elevated RMS values are likely due to a very small number of stations with larger LEK–data differences in those bins.

3.2 | Evaluation of the Cluster Halo Method

For computational reasons, the full station database has to be split into geographically defined clusters, and separate LEK calculations are performed on each cluster. This raises the concern that the normals for stations near the edge of a cluster will be estimated primarily using information from neighbouring stations in

one direction rather than from all sides of the station. This can be addressed by including an additional ‘halo’ of stations around the cluster that are used to improve the LEK estimates near the edge of the cluster, but with increased computational cost. A similar concern arises naturally where a cluster is bounded by an ocean region, though in this case, the issue cannot be resolved by including extra stations as there are none over the ocean. The halo analysis can nevertheless help to quantify the typical size of the boundary effect, which is useful for the coastal boundaries too.

Figure 6 shows two example clusters for parts of the Maritime Continent (top) and South Asia (bottom) with halo stations in blue. The cluster stations are coloured according to the magnitude of the RMS (Root Mean Square) difference between the ‘LEK-only’ normals calculated with and without the halo. As expected, the impact of including a halo is limited to stations near the cluster boundaries. However, even close to the cluster boundaries, the RMS difference (calculated over the 12 monthly normals) from including a halo is rarely larger than 0.1°C. This demonstrates that LEK can be used to estimate normals even close to the boundaries of a cluster of stations.

We next compare the impact of including a halo on the RMS difference between the ‘LEK-only’ and ‘data-only’ normal for two different maximum cluster sizes (Figure 7). Although we have seen in the previous figure some examples where the halo slightly modifies the LEK-based estimate of the normal, when considered across the whole station database the inclusion of a halo barely improves their agreement with the ‘data-only’ normals. The halo makes a small improvement at the lower quartile (i.e., for stations where the LEK approach was already closely matching the normals calculated directly from observations), and for the mean RMS for the smaller clusters (which makes sense because splitting the dataset into smaller clusters will increase the number of stations near a cluster boundary). These results also show slightly closer agreement with larger clusters, with up to a 7% reduction in RMS difference. Given that the ‘data-only’ normals are not the truth (since they suffer from their own limitations such as many being based on partially complete data), this modest increase in agreement is not sufficient to justify using the larger cluster sizes given the increase in computational cost.

3.3 | Previous Biases From Incomplete 1961–1990 Data

The CRUTEM5 approach allows ‘data-only’ normals to be calculated from as little as 15 values during 1961–1990, with no restriction on how those values are spread through the 30-year period (see Osborn and Jones 2014 for restrictions applied in some earlier versions of CRUTEM). Suppose we only have values for 1961–1975 at station A and only values for 1976–1990 at station B. If, like the global-mean, both stations are warming, then the normal calculated for station A will be cooler than its full 1961–1990 mean, while the normal for station B will be warmer. Subtracting each normal from its station’s values will make station A’s anomalies slightly too warm and station B’s slightly too cool. Because station A is more likely to have pre-1961 data and station B is more likely to have post-1990 data, the overall effect is to artificially warm earlier anomalies and

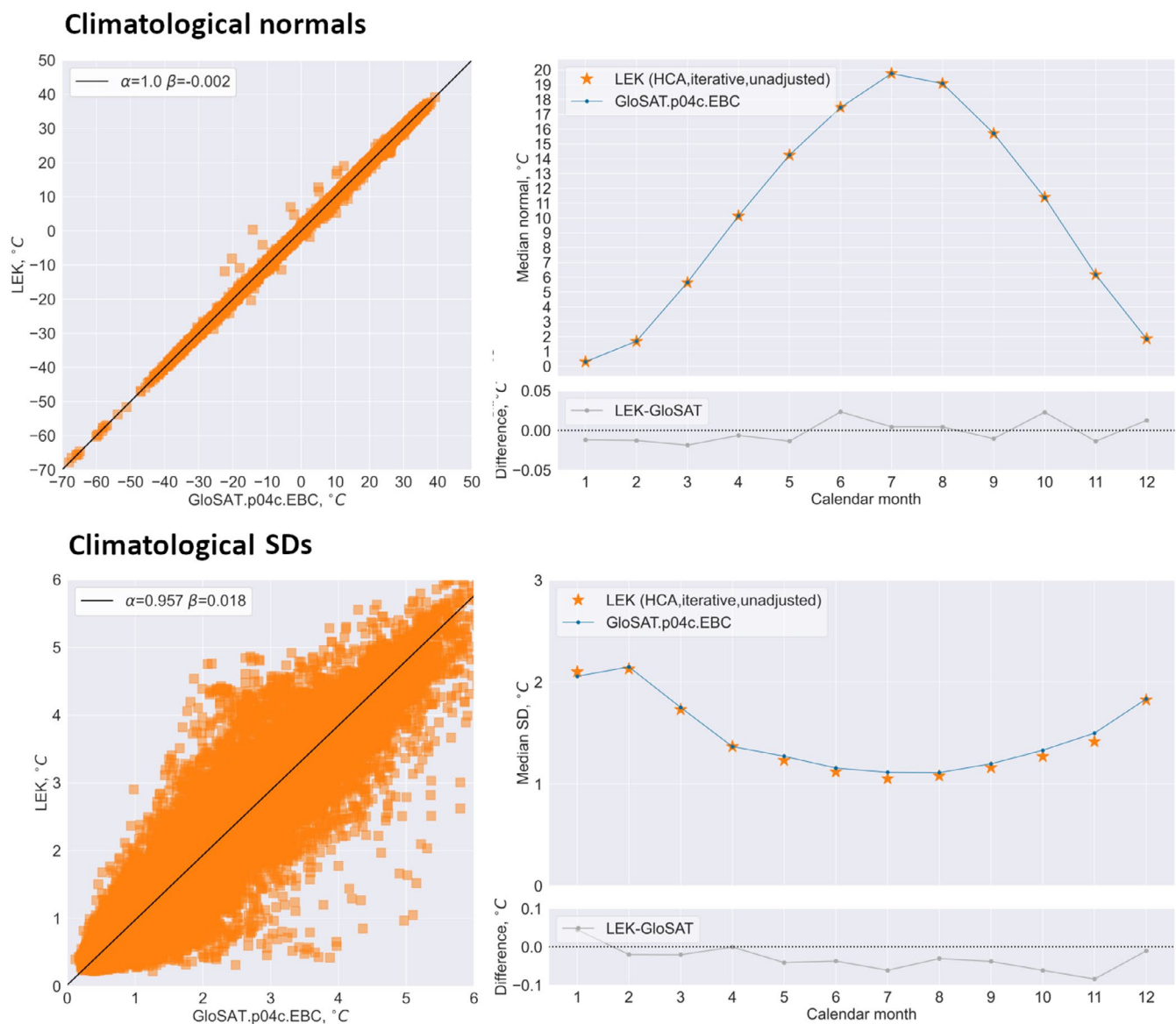


FIGURE 4 | Comparison of climatological normals (top) and SDs (bottom) calculated from LEK values with those calculated from co-located station data. Left column: Scatter plots (LEK-based estimates on y-axes, station data estimates on x-axes; slope and intercept of linear fit are annotated). Right column: Medians over all stations by month, with differences in medians below.

cool later anomalies, and artificially reduce the overall warming. These examples are extreme cases, and Calvert (2024) estimates the overall effect on global mean temperature to be very small (0.003°C reduction in long-term warming). Nevertheless, our use of LEK to infill data gaps will remove this artefact by supplementing the observed values with LEK-based estimates so that the normals are calculated from the full 30 years (both these example stations would follow the normal category code 7 method shown in Table 4).

The fingerprint of this effect is illustrated in Figure 8, showing the mean difference between LEK-based normals and ‘data-only’ normals as a function of the mean year of the available observations within the 1961–1990 period. A mean year of 1968 can only be obtained for cases similar to station A (i.e., with only 1961–1975 observations), while a mean year of 1983 can only be obtained for cases like station B. A mean year of 1975 could arise for many reasons, for example, complete data or only

15 values spread evenly across the reference period. Though the variation of individual station differences is large (shading in Figure 8 shows the mean difference ± 1 SD), there is a consistent pattern across all 12 months of the LEK-based normal exceeding the ‘data-only’ normal for stations with earlier data and being lower for stations with data only later in the reference period. The overall effect between the extremes (station A to station B) is around 0.1°C when aggregated over all months and stations, similar to the rate of global warming during this period. LEK infilling serves to remove this issue arising from data gaps during the 1961–1990 reference period.

3.4 | Improved Spatio-Temporal Representation

As noted earlier, 3430 stations in the extended database would not have had normals using the CRUTEM5 approach (Table 5) and would not have been used further to create a gridded

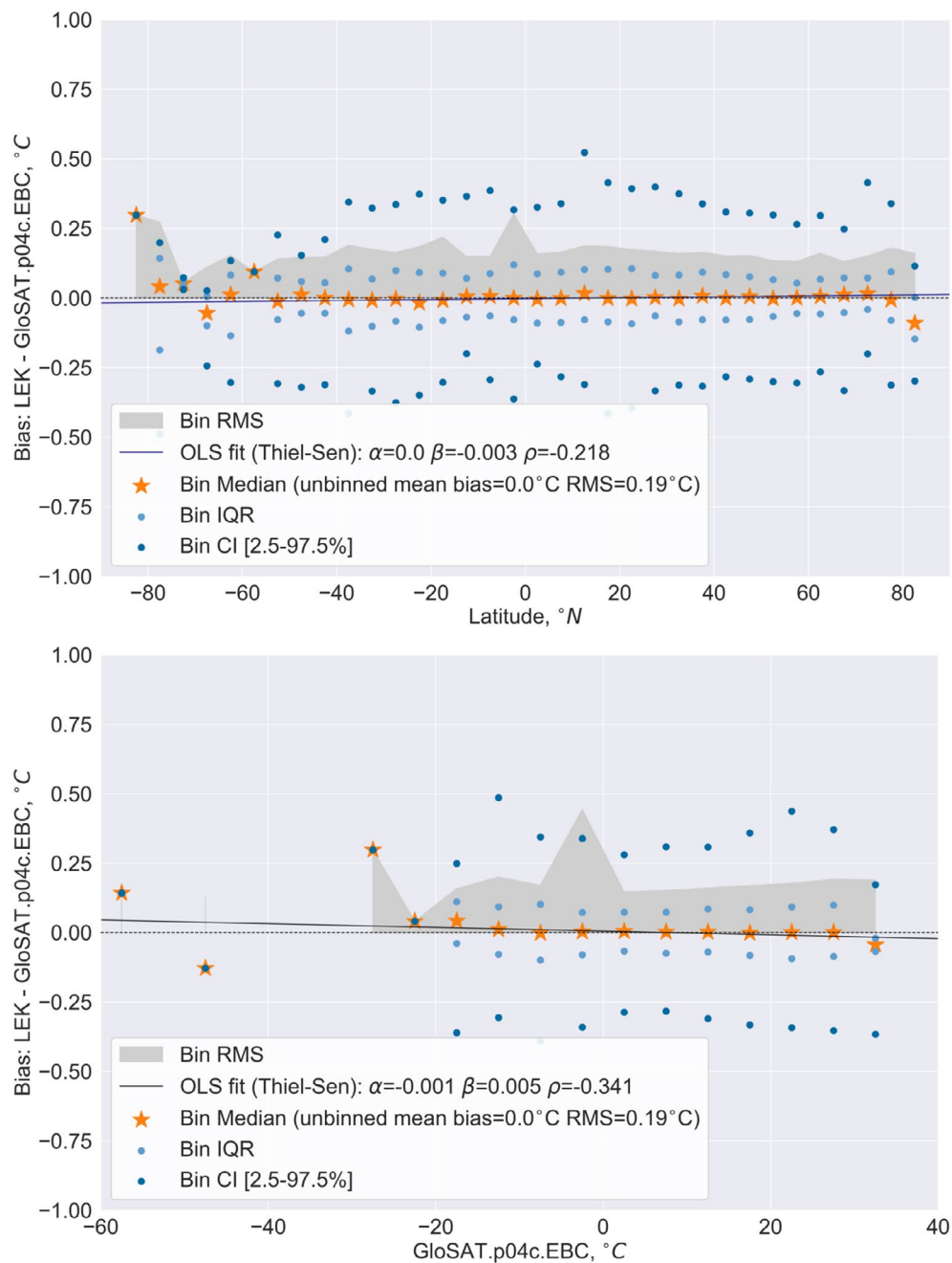


FIGURE 5 | Comparison of climatological normals calculated from LEK values with those calculated from co-located station data as a function of (a) latitude and (b) temperature. Stations are grouped into 5° latitude bins or 5°C temperature bins. The difference between LEK and station normals for all stations within each bin is then summarised by the 2.5 and 97.5 percentiles, the lower and upper quartiles, the median, and their root mean squared difference. Regression lines (solid) show the lack of dependence on latitude or temperature.

temperature anomaly dataset (such as GloSATref: Morice et al. 2025). Now, however, using LEK to estimate missing normals has reduced this to 731, and the number of stations that can be converted to temperature anomalies and used for gridding has increased from 7983 in CRUTEM5 (Osborn et al. 2021) to 8435 with the new acquisitions, and now to 11,134 with the LEK process (see Table 5 and reduction in *white* station locations and counts from left column to right column of Figure 2). The remaining 731 stations for which normals are still not estimated either have no near neighbours, are very short (325 span less than a decade) or have limited temporal overlap with nearby stations that have data during 1961–1990. These reasons prevent

a successful application of kriging for estimating values during 1961–1990.

Furthermore, many of the existing normals have been improved. Those with normals based on less reliable external or previous estimates (codes 2, 3 and 5; see Section 2.4) have been reduced from 129 to just 18 stations, also improving reproducibility. Those normals calculated only from station observations have been reduced by 2500 from 8306 to 5806. These 2500 would have previously been estimated from partial data during the reference period and now they are based on complete data via LEK infilling, reducing the potential for bias (Section 3.3). They are

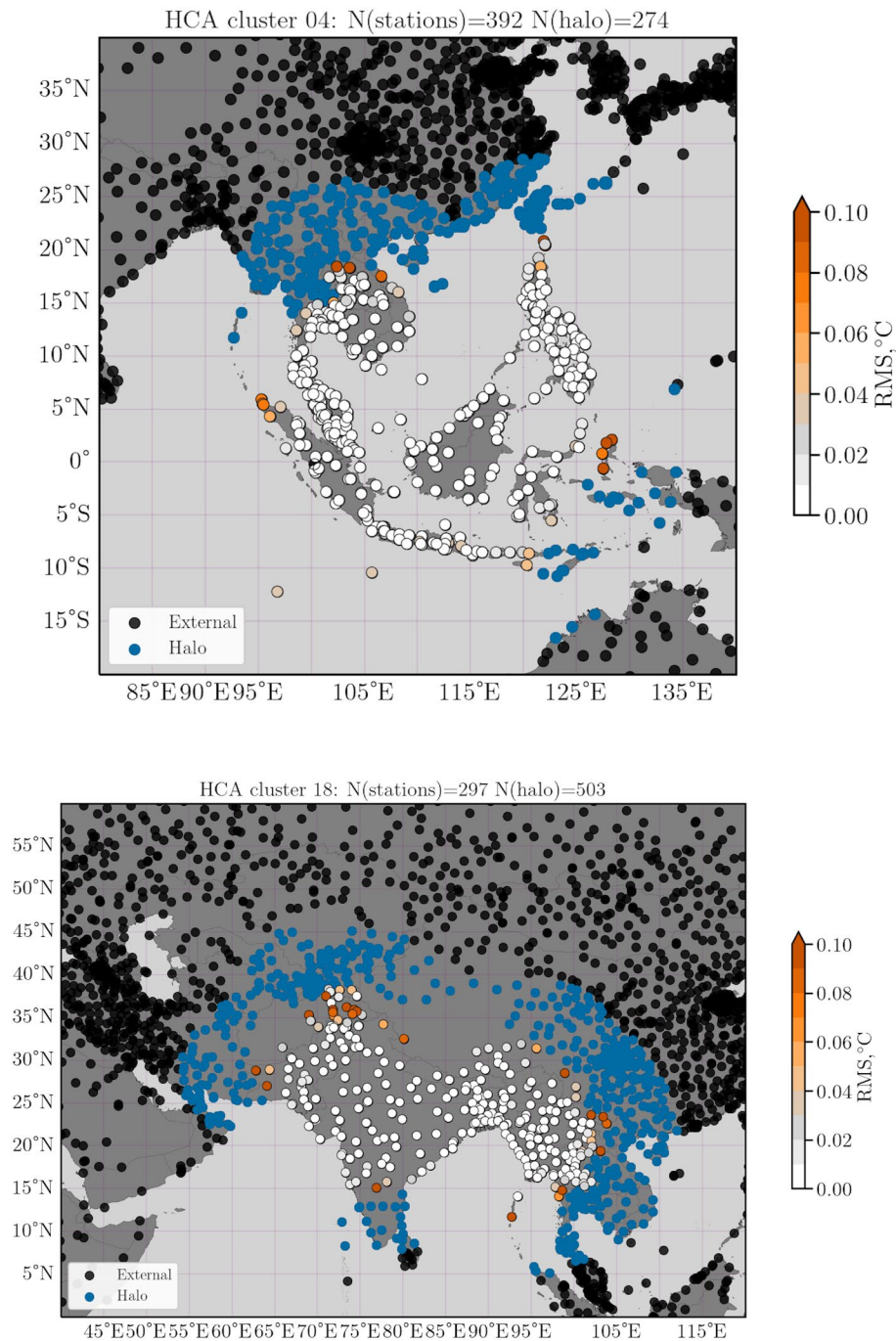


FIGURE 6 | The impact of including a halo of stations around a cluster on LEK normals for two example clusters: Maritime Continent (top) and South Asia (bottom). The halo stations are blue and the cluster stations are coloured according to the magnitude of the RMS (Root Mean Square) difference between the normals estimated by applying LEK with and without the halo. Stations external to both the cluster and halo are black.

represented by stations that have changed from orange to dark blue (code 7) in Figure 2.

There are now 1742 stations with normals estimated solely from LEK. These are mostly post-1990 stations (code 6, pale blue in Figure 2) and help to address the fall off in the number of stations used in gridded datasets (as noted for CRUTEM5 by Osborn et al. 2021). However, they also include some pre-1960 stations and their inclusion will improve coverage in the nineteenth and early twentieth centuries. Thus using LEK to estimate normals contributes to the gains made by data rescue.

The additional station anomaly time series now made available by the use of LEK to aid the estimation of station climatological normals improves station counts across all continents (Table 6). The greatest proportional increases are in Antarctica, where many shorter automatic weather stations can now be used, and South America, where counts of usable stations are nearly doubled.

While this increase in normals (and hence in the series that can be converted to anomalies and used to create gridded datasets) is a positive development, its impact should not be overstated:

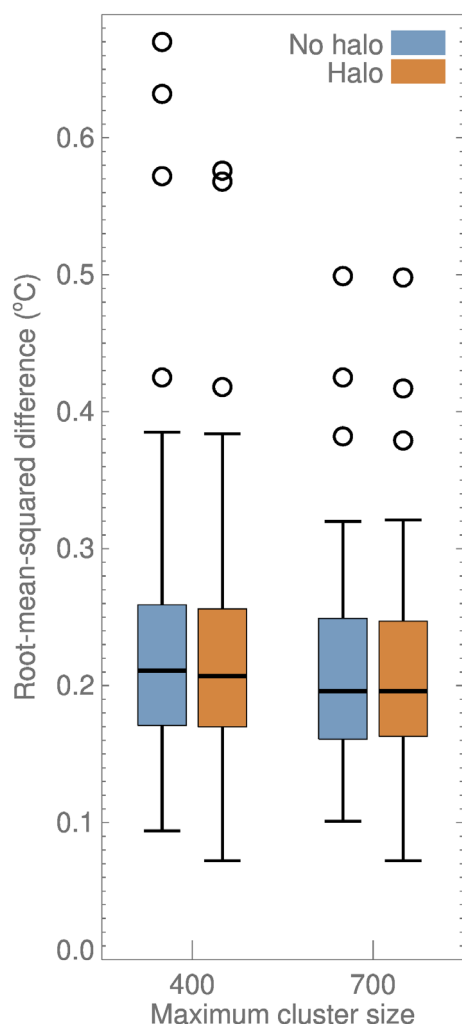


FIGURE 7 | Box plots of the RMS differences between normals calculated from LEK values and normals calculated from co-located station data with (orange) and without (blue) cluster halos. Results are shown for two different clustering options: Maximum cluster size of 400 stations (left) and 700 stations (right). Whiskers are located at 1.5 IQR above and below the quartiles, with outliers marked by open circles.

many of these stations with new LEK-estimated normals are either quite short or are located in station-dense regions of northern Europe and North America, so overall spatial coverage is not expected to increase greatly.

To illustrate the potential for improvement, we present a preliminary view of the impact of the changes reported here (updated, improved and extended station data; adjustment for exposure bias of mid-latitude stations following Wallis et al. 2024; improved and infilled normals using LEK) on hemispheric temperature anomaly timeseries. A more comprehensive presentation is given by Morice et al. (2025), where GloSAT LATsdb is used as the land surface air temperature input to a new gridded dataset of Global Surface Air Temperature (the GloSAT reference analysis) which includes marine air temperature observations and a comprehensive representation of correlated and uncorrelated errors.

Hemispheric land air temperature series have been constructed using the ‘alternative’ gridding method of Osborn et al. (2021)

that takes account of the convergence of lines of longitude towards the poles by allowing stations to contribute to nearby grid cells at high latitudes. For the Northern Hemisphere (NH; Figure 9a), GloSAT LATsdb provides relatively complete 1781–1849 coverage only for Europe, with a smattering of grid cells with a smaller amount of pre-1850 data across Asia and the coasts of North America. Post-1850 data has almost complete data for Europe, USA, southern Canada, India, Japan, and parts of Russia, with least complete data for Africa and some Pacific Islands. The new acquisitions and, especially, the inclusion of LEK-based normals has increased the spatial coverage of data by a small amount throughout the time period, but to a greater degree in recent decades (compare the blue and grey shaded areas in Figure 9b).

The updated database, exposure bias adjustments and inclusion of LEK-based normals make only a small difference to the NH temperature anomaly timeseries (Figure 9c) beyond the obvious extension back to 1781. The very limited spatial coverage in the early period (Figure 9a) contributes to the more pronounced variability before 1850 (Figure 9c). The individual influences on the NH temperature series can be seen in the differences shown in panel (d) of Figure 9. The new acquisitions cooled the pre-1870 anomalies by around 0.1°C (black curve) compared with CRUTEM5. Exposure bias adjustments cool the 1880–1935 annual temperatures by about 0.06°C (orange curve), though the impact is larger on summer temperature (not shown here, but see figure 12 of Wallis et al. 2024). Inclusion of stations with LEK-estimated normals slightly warms the pre-1860 NH temperature series (difference between blue and orange curves) because it allows the inclusion of series with, on average, slightly less negative anomalies than those of the long stations that have continuous data from the early period to the 1961–1990 baseline.

GloSAT LATsdb has no pre-1850 coverage in the Southern Hemisphere (SH; Figure 10a), but the LEK-estimated normals have provided better post-1990 coverage (Figure 10b; compare the blue and grey shaded areas) by allowing the inclusion of many Antarctic stations with insufficient reference period data to calculate normals directly from the data. The extra coverage has only a small effect on the SH series (though temporarily warming it by almost 0.1°C around 2010; blue curve in Figure 10d), while exposure bias adjustments have slightly cooled pre-1895 anomalies (orange curve). Again, see Wallis et al. (2024) for reasons. The net effect of new data, exposure bias adjustments, and LEK-estimated normals is negligible on the SH mean temperature anomaly (Figure 10c).

4 | Dataset Access

4.1 | Licence

The dataset has been produced for the GloSAT project (www.glosat.org) and is released into the public domain with an Open Government Licence (<https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>). Users are free to use this dataset and only need to acknowledge the source of the information.

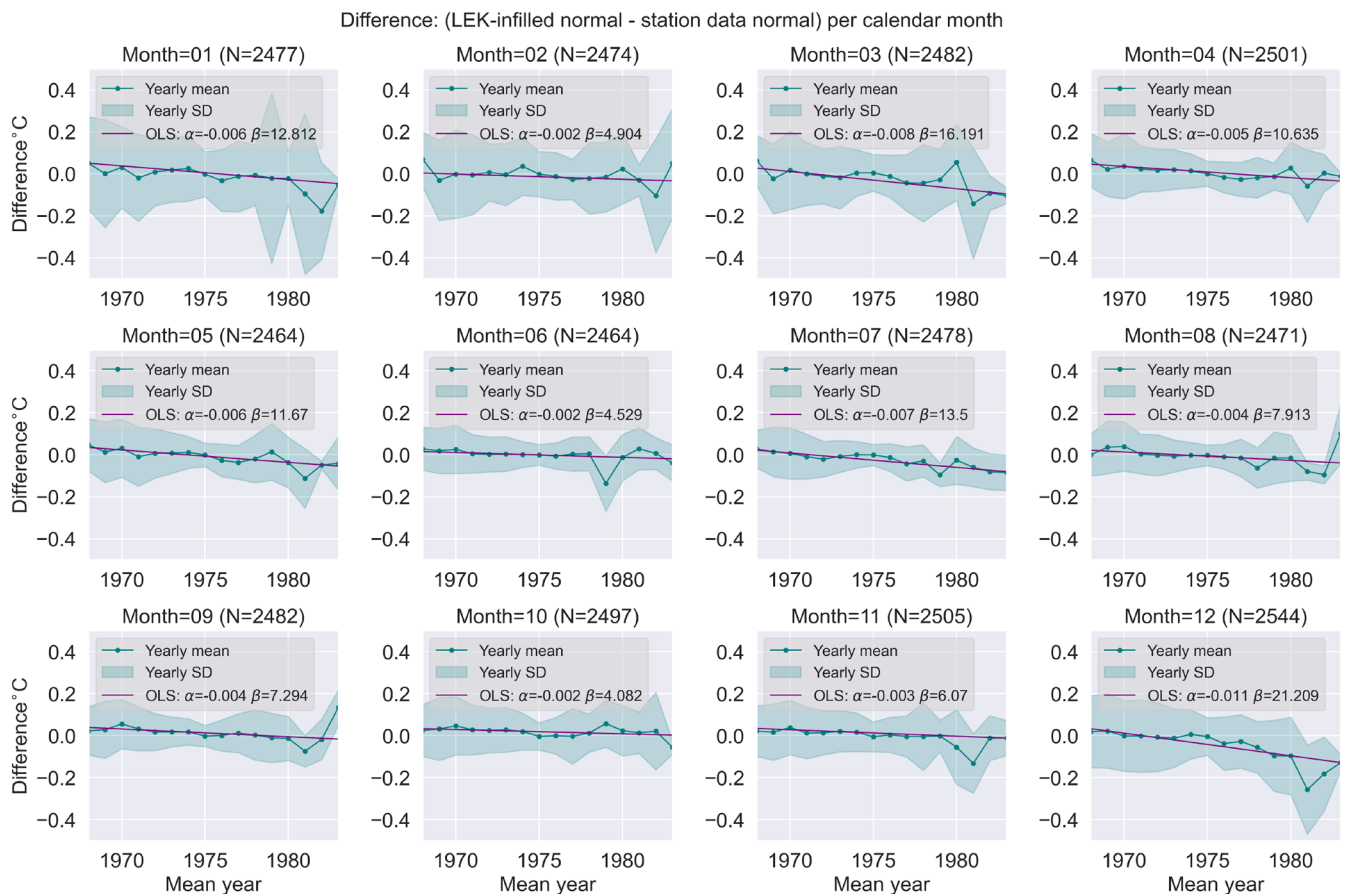


FIGURE 8 | Bias in station normals when calculated from incomplete 1961–1990 data, illustrated by the difference between LEK-based normals (complete data) and normals calculated from co-located (but incomplete) station data. The difference between LEK and station normals is binned by the mean year of the available station data. Differences within each bin are summarised by their mean \pm 1 SD and across bins by the regression lines (solid purple; with slope and intercept annotated). Results are shown separately for each calendar month.

TABLE 6 | Stations with normals per continental region.

Region	GloSAT_eba	GloSAT LATsdb
Africa	582	860
Antarctica	31	64
Asia	1777	2213
Europe	2494	3386
Oceania	476	621
North America	2466	2841
South America	609	1149
Total	8435	11,134

4.2 | Location and Format

The GloSAT LATsdb dataset is provided in an open access Zenodo repository for long-term preservation. It is also available via the CRU and Met Office websites to enhance its findability and where additional formats and information may be provided.

Zenodo: <https://doi.org/10.5281/zenodo.14888902>

CRU: <https://crudata.uea.ac.uk/>

Met Office: <https://www.metoffice.gov.uk/hadobs/glosatref/>

The dataset comprises five components:

1. The station database prior to application of exposure bias adjustments and prior to removal of values that failed quality checks
2. The station database after application of exposure bias adjustments but prior to removal of values that failed quality checks
3. The station database after application of exposure bias adjustments and after removal of values that failed quality checks
4. The station climatological normals
5. The station climatological standard deviations

The data files are in plain text format, following the long-standing format and structure used for CRUTEM station data. Additionally, the station data files provided at the Met Office website are in NetCDF4 format.

Readers written in Python to read in the station database, climatological normals and SD files are available at: <https://github.com/patternizer/glosat-py>. The open source code for

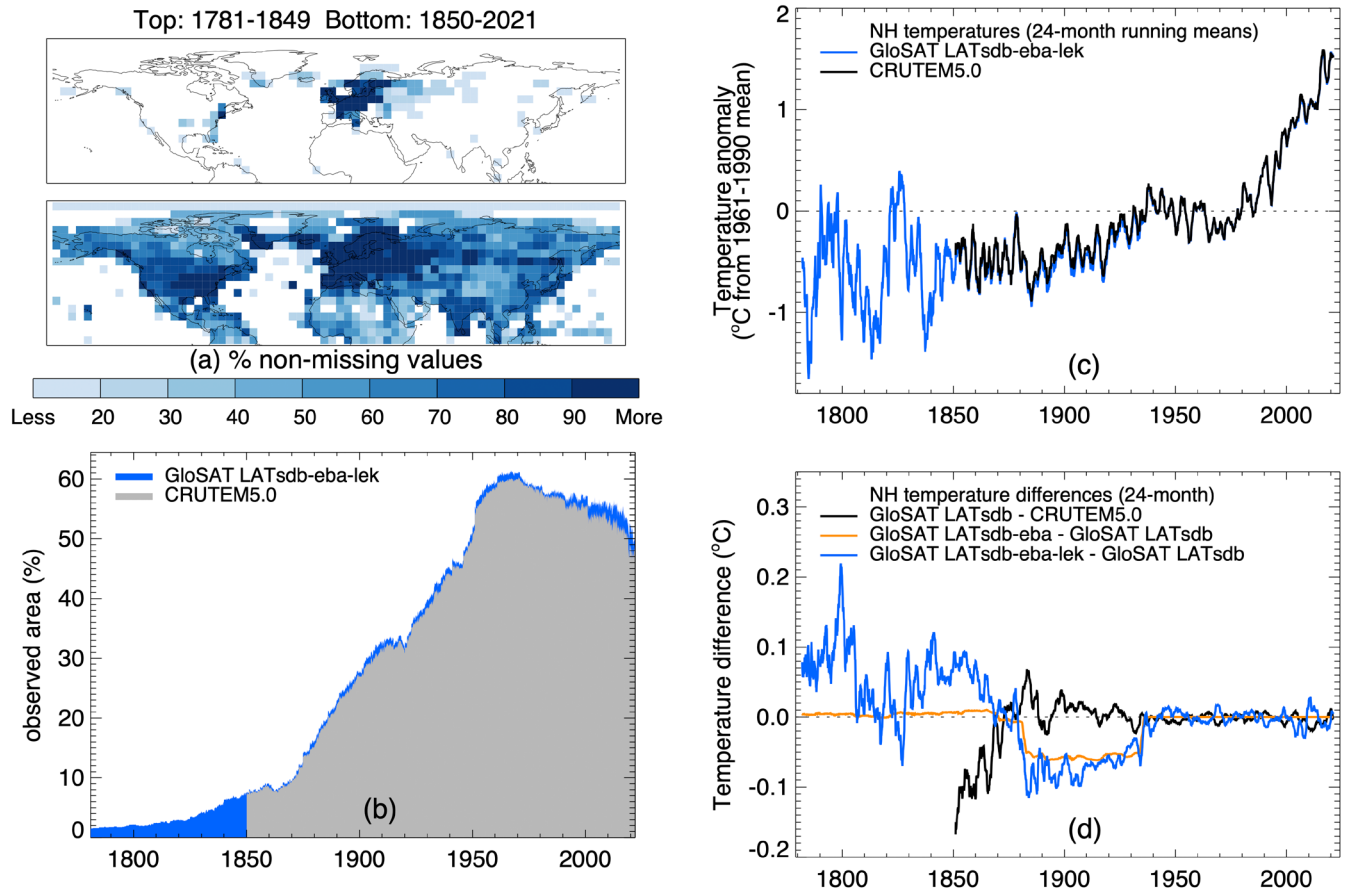


FIGURE 9 | A comparison of Northern Hemisphere temperature coverage and anomalies to show the impact of LEK normals and exposure bias adjustment. The station temperature data were processed using the CRUTEM5 method with the alternative gridding option (Osborn et al. 2021). (a) Temporal completeness of the final GloSAT LATsdb during the 1781–1849 and 1850–2021 periods. (b) Time evolving spatial completeness of the gridded dataset based on either the CRUTEM5.0 station database (grey shading) or the final GloSAT LATsdb including LEK-based normals (blue shading). (c) Hemispheric average temperature anomalies based on the CRUTEM5.0 station database (black) or the final GloSAT LATsdb including LEK-based normals (blue). (d) Difference between hemispheric average temperature anomalies using different pairs of station databases: The extended GloSAT LATsdb prior to exposure bias adjustment and inclusion of LEK-based normals compared with CRUTEM5 (black); the effect of applying exposure bias adjustments (orange); and the effect of also including LEK-based normals (blue).

the local expectation kriging of the station archive is available at: <https://github.com/KCowtan/glosat-homogenisation>. Additionally, the table of data source codes (which includes the broad indicators of homogenisation described in Section 2.3) is available at: <https://github.com/TimOsbornClim/GloSAT-LATsdb>.

4.3 | Data Updates

The GloSAT LATsdb dataset described here spans the year range 1781–2021. It is built on the CRUTEM5 station database and therefore it will benefit from the new data acquisition, updating, and assessment workflow for CRUTEM5 undertaken regularly by CRU and the Met Office. These updates to CRUTEM5 will, in due course, provide updates to GloSAT LATsdb, and the process developed here to estimate station normals means that these routine updates will no longer need to prioritise stations that existed during the 1961–1990 reference period.

5 | Dataset Use

This GloSAT LATsdb dataset is being used, in conjunction with observations of marine air temperature (GloSAT MAT) built from the work of Cornes et al. (2020) and Cropper et al. (2023), to develop a gridded monthly temperature anomaly dataset for the period 1781 to present (Morice et al. 2025). This new land and marine air temperature dataset (named GloSATref) will complement existing global temperature datasets that mostly use sea surface temperature anomalies for their marine component. This data descriptor forms the land component of the reference analysis input data.

Other potential uses for GloSAT LATsdb are for evaluation of climate change at individual stations, to complement analyses that use gridded datasets and which might not capture fine details that require station-level data. For example, exploring elevation-dependent climate change, further work on changes in thermometer screens/exposure or on the inhomogeneities present in the underlying observations.

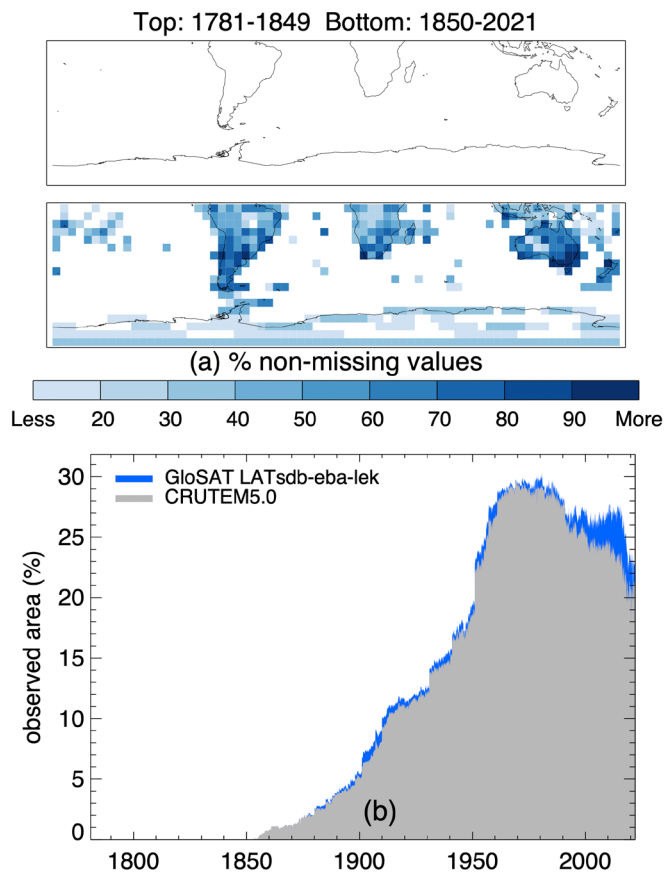


FIGURE 10 | As Figure 9, but for the Southern Hemisphere.

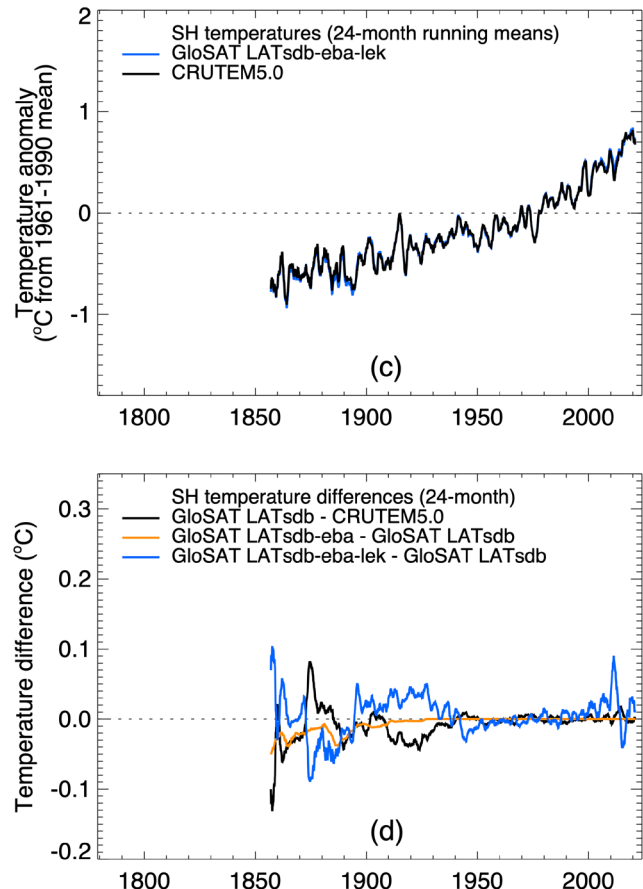
6 | Summary

This data descriptor paper explains the processes and data sources used to construct a new database of monthly mean temperatures recorded at weather stations around the world. While it is firmly based on the existing CRUTEM5 station database, a number of new aspects are achieved: (1) A 69-year extension back to begin in 1781. (2) Acquisition of 1233 station records not previously included in the CRUTEM5 compilation, and improvements to the homogeneity of many more stations based on homogenisation exercises undertaken by the source National Meteorological Services. (3) Application of adjustments for changes in thermometer screens at mid-latitude stations in both hemispheres from the recent work of Wallis et al. (2024). (4) Development and evaluation of a method, using local expectation kriging, to estimate the 1961–1990 reference period averages ('normals') for stations which either did not have estimated normals or to improve those normals that were previously estimated from incomplete data.

As discussed in Section 5, this station database is being used to create gridded temperature datasets; it will be updated as new data become available, and it is available for researchers to use for station-based analyses from local to global scales.

Acknowledgements

The research presented in this paper was funded by the UK National Environment Research Council (NERC) GloSAT project (NE/



S015582/1 and NE/S015566/1). We are grateful to the Centre for Environmental Data Analysis computational facilities that provided the group workspace for code and data storage and high-performance computing via the JASMIN platform. We acknowledge weather station observers and their national meteorological services for kindly making available their time series measurements of surface temperature. The team acknowledges gratefully the other members of the GloSAT project, including the scientific leadership of Liz Kent and the project management by Amani Becker (both at the National Oceanography Centre, UK), and two reviewers whose careful evaluation helped to improve the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Zenodo at <https://doi.org/10.5281/zenodo.14888902>.

References

- Arguez, A., and R. S. Vose. 2011. "The Definition of the Standard WMO Climate Normal: The Key to Deriving Alternative Climate Normals." *Bulletin of the American Meteorological Society* 92, no. 6: 699–704. <https://doi.org/10.1175/2010BAMS2955.1>.
- Barber, D. 2012. *Bayesian Reasoning and Machine Learning*. Cambridge University Press.
- Beck, H. E., N. E. Zimmermann, T. R. McVicar, N. Vergopolan, A. Berg, and E. F. Wood. 2018. "Present and Future Köppen-Geiger Climate

- Classification Maps at 1-km Resolution." *Scientific Data* 5, no. 1: 180214. <https://doi.org/10.1038/sdata.2018.214>.
- Bishop, C. 2006. *Pattern Recognition and Machine Learning*. Springer.
- Briffa, K. R., T. M. Melvin, T. J. Osborn, et al. 2013. "Reassessing the Evidence for Tree-Growth and Inferred Temperature Change During the Common Era in Yamalia, Northwest Siberia." *Quaternary Science Reviews* 72: 83–107. <https://doi.org/10.1016/j.quascirev.2013.04.008>.
- Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones. 2006. "Uncertainty Estimates in Regional and Global Observed Temperature Changes: A New Data Set From 1850." *Journal of Geophysical Research: Atmospheres* 111, no. D12: 2005JD006548. <https://doi.org/10.1029/2005jd006548>.
- Calvert, B. T. T. 2024. "Improving Global Temperature Datasets to Better Account for Non-Uniform Warming." *Quarterly Journal of the Royal Meteorological Society* 150: 3672–3702. <https://doi.org/10.1002/qj.4791>.
- Cornes, R. C., E. Kent, D. Berry, and J. J. Kennedy. 2020. "CLASSnmat: A Global Night Marine Air Temperature Data Set, 1880–2019." *Geoscience Data Journal* 7, no. 2: 170–184. <https://doi.org/10.1002/gdj3.100>.
- Cowtan, K., P. Jacobs, P. Thorne, and R. Wilkinson. 2018. "Statistical Analysis of Coverage Error in Simple Global Temperature Estimators." *Dynamics and Statistics of the Climate System* 3, no. 1: dzy003. <https://doi.org/10.1093/climsys/dzy003>.
- Cressie, N. 1990. "The Origins of Kriging." *Mathematical Geology* 22: 239–252.
- Cropper, T. E., D. I. Berry, R. C. Cornes, and E. C. Kent. 2023. "Quantifying Daytime Heating Biases in Marine Air Temperature Observations From Ships." *Journal of Atmospheric and Oceanic Technology* 40, no. 4: 427–438. <https://doi.org/10.1175/jtech-d-22-0080.1>.
- Everitt, B. S., S. Landau, M. Leese, and D. Stahl. 2011. *Cluster Analysis*. 5th ed. Wiley.
- Gillespie, I. M., L. Haimberger, G. P. Compo, and P. W. Thorne. 2021. "Assessing Potential of Sparse-Input Reanalyses for Centennial-Scale Land Surface Air Temperature Homogenisation." *International Journal of Climatology* 41, no. S1: E3000–E3020. <https://doi.org/10.1002/joc.6898>.
- Guttman, N. 1989. "Statistical Descriptors of Climate." *Bulletin of the American Meteorological Society* 70: 602–607. https://journals.ametsoc.org/view/journals/bams/70/6/1520-0477_1989_070_0602_sdock_2_0_co_2.xml.
- Hawkins, E., P. Ortega, E. Suckling, et al. 2017. "Estimating Changes in Global Temperature Since the Preindustrial Period." *Bulletin of the American Meteorological Society* 98, no. 9: 1841–1856. <https://doi.org/10.1175/BAMS-D-16-0007.1>.
- Huang, B., M. J. Menne, T. Boyer, et al. 2020. "Uncertainty Estimates for Sea Surface Temperature and Land Surface Air Temperature in NOAA GlobalTemp Version 5." *Journal of Climate* 33, no. 4: 1351–1379. <https://doi.org/10.1175/jcli-d-19-0395.1>.
- Hulme, M. 2021. *Climates Multiple: Three Baselines, Two Tolerances, One Normal*. Academia Letters. <https://doi.org/10.20935/AL102>.
- Jones, P. D., and A. Moberg. 2003. "Hemispheric and Large-Scale Surface Air Temperature Variations: An Extensive Revision and an Update to 2001." *Journal of Climate* 16, no. 2: 206–223. [https://doi.org/10.1175/1520-0442\(2003\)016<0206:halssa>2.0.co;2](https://doi.org/10.1175/1520-0442(2003)016<0206:halssa>2.0.co;2).
- Jones, P. D., T. J. Osborn, and K. R. Briffa. 1997. "Estimating Sampling Errors in Large-Scale Temperature Averages." *Journal of Climate* 10: 2548–2568. https://journals.ametsoc.org/view/journals/clim/10/10/1520-0442_1997_010_2548_eseils_2.0.co_2.xml.
- King, J. C., J. Turner, S. Colwell, et al. 2021. "Inhomogeneity of the Surface Air Temperature Record From Halley, Antarctica." *Journal of Climate* 34: 4771–4783. <https://doi.org/10.1175/JCLI-D-20-0748.1>.
- Mahony, C. R., T. Wang, A. Hamann, and A. J. Cannon. 2022. "A Global Climate Model Ensemble for Downscaled Monthly Climate Normals Over North America." *International Journal of Climatology* 42, no. 11: 5871–5891. <https://doi.org/10.1002/joc.7566>.
- Matthews, T., L. B. Perry, I. Koch, et al. 2020. "Going to Extremes: Installing the World's Highest Weather Stations on Mount Everest." *Bulletin of the American Meteorological Society* 101, no. 11: E1870–E1890. <https://doi.org/10.1175/BAMS-D-19-0198.1>.
- Morice, C. P., D. I. Berry, R. C. Cornes, et al. 2025. "An Observational Record of Global Gridded Near Surface Air Temperature Change Over Land and Ocean From 1781." *Earth System Science Data*. <https://doi.org/10.5194/essd-2024-500>.
- Morice, C. P., J. J. Kennedy, N. A. Rayner, et al. 2021. "An Updated Assessment of Near-Surface Temperature Change From 1850: The HadCRUT5 Data Set." *Journal of Geophysical Research: Atmospheres* 126, no. 3: e2019JD032361. <https://doi.org/10.1029/2019jd032361>.
- New, M., M. Hulme, and P. Jones. 1999. "Representing Twentieth-Century Space-Time Climate Variability. Part I: Development of a 1961–90 Mean Monthly Terrestrial Climatology." *Journal of Climate* 12: 829–856. [https://doi.org/10.1175/1520-0442\(1999\)012<0829:RTCSTC>2.0.CO;2](https://doi.org/10.1175/1520-0442(1999)012<0829:RTCSTC>2.0.CO;2).
- Osborn, T. J., and P. D. Jones. 2014. "The CRUTEM4 Land-Surface Air Temperature Data Set: Construction, Previous Versions and Dissemination via Google Earth." *Earth System Science Data* 6, no. 1: 61–68. <https://doi.org/10.5194/essd-6-61-2014>.
- Osborn, T. J., P. D. Jones, D. H. Lister, et al. 2021. "Land Surface Air Temperature Variations Across the Globe Updated to 2019: The CRUTEM5 Data Set." *Journal of Geophysical Research: Atmospheres* 126, no. 2: e2019JD032352. <https://doi.org/10.1029/2019jd032352>.
- Pang, Y., Y. Wang, X. Lai, S. Zhang, P. Liang, and X. Song. 2023. "Enhanced Kriging Leave-One-Out Cross-Validation in Improving Model Estimation and Optimization." *Computer Methods in Applied Mechanics and Engineering* 414: 116194. <https://doi.org/10.1016/j.cma.2023.116194>.
- Parker, D. E. 1994. "Effects of Changing Exposure of Thermometers at Land Stations." *International Journal of Climatology* 14, no. 1: 1–31. <https://doi.org/10.1002/joc.3370140102>.
- Perry, M., and D. Hollis. 2005. "The Development of a New Set of Long-Term Climate Averages for the UK." *International Journal of Climatology* 25, no. 8: 1023–1039. <https://doi.org/10.1002/joc.1160>.
- Przybylak, R., and P. Wyszynski. 2020. "Air Temperature Changes in the Arctic in the Period 1951–2015 in the Light of Observational and Reanalysis Data." *Theoretical and Applied Climatology* 139, no. 1–2: 75–94. <https://doi.org/10.1007/s00704-019-02952-3>.
- Rohde, R., R. Muller, R. Jacobsen, S. Perlmutter, and S. Mosher. 2013. "Berkeley Earth Temperature Averaging Process." *Geoinformatics & Geostatistics: An Overview* 1, no. 2: 1–13. <https://doi.org/10.4172/2327-4581.1000103>.
- Rohde, R. A., and Z. Hausfather. 2020. "The Berkeley Earth Land/Ocean Temperature Record." *Earth System Science Data* 12, no. 4: 3469–3479. <https://doi.org/10.5194/essd-12-3469-2020>.
- Stehr, N., and H. Von Storch, eds. 2000. *Eduard Brückner – The Sources and Consequences of Climate Change and Climate Variability in Historical Times*. Springer Netherlands. <https://doi.org/10.1007/978-94-015-9612-1>.
- van der Schrier, G., and P. D. Jones. 2008. "Daily Temperature and Pressure Series for Salem, Massachusetts (1786–1829)." *Climatic Change* 87, no. 3–4: 499–515. <https://doi.org/10.1007/s10584-007-9292-x>.
- Wackernagel, H. 2003. "Ordinary Kriging." In *Multivariate Geostatistics*, 79–88. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-05294-5_11.

Wallis, E. J., T. J. Osborn, M. Taylor, P. D. Jones, M. Joshi, and E. Hawkins. 2024. "Quantifying Exposure Biases in Early Instrumental Land Surface Air Temperature Observations." *International Journal of Climatology* 44, no. 5: 1611–1635. <https://doi.org/10.1002/joc.8401>.

Wang, Y., X. Zhang, W. Ning, et al. 2023. "The AntAWS Dataset: A Compilation of Antarctic Automatic Weather Station Observations." *Earth System Science Data* 15, no. 1: 411–429. <https://doi.org/10.5194/essd-15-411-2023>.

Way, R. G., and P. P. Bonnaventure. 2015. "Testing a Reanalysis-Based Infilling Method for Areas With Sparse Discontinuous Air Temperature Data in Northeastern Canada." *Atmospheric Science Letters* 16, no. 3: 398–407. <https://doi.org/10.1002/asl2.574>.

WMO. 1971. *Climatological Normals (CLINO) for CLIMAT and CLIMAT SHIP Stations for the Period 1931–1960*. World Meteorological Organization. WMO 117. <https://library.wmo.int/idurl/4/37368>.

WMO. 1989. *Calculation of Monthly and Annual 30-Year Standard Normals*. World Meteorological Organization. WMO/TD-341. <https://library.wmo.int/idurl/4/43912>.

WMO. 1996. "Climatological Normals (CLINO) for the Period 1961–1990." <https://community.wmo.int/en/activity-areas/climate-services/climate-products-and-initiatives/wmo-climatological-normals>.

WMO. 2007. *The Role of Climatological Normals in a Changing Climate*. World Meteorological Organization. WMO/TD-1377. <https://library.wmo.int/idurl/4/52499>.

WMO. 2017. *WMO Guidelines on the Calculation of Climate Normals*. World Meteorological Organization. WMO-1203. <https://library.wmo.int/records/item/55797-wmo-guidelines-on-the-calculation-of-climate-normals>.

Appendix A

Climatological Normals

It is useful to briefly review how the concept and definition of normals have evolved. The concept first appeared in 1840, followed in 1872 by an international agreement to compile mean values over a uniform period (Guttman 1989). The first widely used climatological normals were developed during the early 20th century (see Arguez and Vose 2011; Hulme 2021) for the designated periods of 1881–1915 and 1916–1950. Examples of their use can be seen in a number of publications from the early 20th century (e.g., the annual volumes of the publication British Rainfall, <https://www.metoffice.gov.uk/research/library-and-archive/archive-hidden-treasures/british-rainfall>).

The original purpose of normals was to provide an expectation of what each weather/climate station should get for monthly mean temperature and monthly total precipitation. This could then be used for planning purposes, particularly in the water and agricultural sectors. The number of variables for which normals were calculated was extended during the 20th century to include maximum and minimum temperature, sunshine amounts, mean sea level pressure, and a humidity measure such as vapour pressure or relative humidity (New et al. 1999).

It is believed that the original 35-year periods resulted from work by Eduard Brückner [see Stehr and Von Storch (2000) for a modern translation], who believed such cycles existed in varves in eastern European lakes. Later, the International Meteorological Organization (IMO) in 1937 (see Hulme 2021) designated 30-year periods, the first of which was 1901–30. Subsequent periods were introduced later by the World Meteorological Organization (WMO): 1931–1960, 1961–1990, and 1991–2020, all non-overlapping with each other. Publications with the normals for the 1931–1960 and 1961–1990 periods were produced (WMO 1971, 1996, respectively).

During the 1980s and 1990s, it began to be realised that the original purpose of normals was less useful due to climatic warming. Arguez and Vose (2011) noted they would be out of date 10–15 years after their final year. They made a number of suggestions following discussions within the Commission for Climatology within WMO (1989, 2007). Some of

the suggestions related to the overlapping periods which countries were asked by WMO to produce (1941–1970, 1951–1980, 1971–2000 and 1981–2010). These in-between periods were not designated official periods, which are still the non-overlapping 30-year periods, and compilations of their results were not formally published. Calculation of normals is carried out by the National Meteorological Services (NMS) and guidance for NMS has been provided over the years by the WMO. The most recent such guidance is WMO (2017).

For the purpose of assessing long-term climate change, the period 1961–1990 is recommended and will continue to be used (WMO 2017). However, another period ('pre-industrial') has been enshrined in the Paris Agreement on Climate Change in 2015. This agreement did not specify a period (see the discussion of possibilities in Hawkins et al. 2017), but the second half of the 19th century (1850–1900) has since been used by IPCC and in many other studies.