An updated assessment of near-surface temperature change from 1850: the HadCRUT5 dataset

C. P. Morice[1], J. J. Kennedy[1], N. A. Rayner[1], J. P. Winn[1], E. Hogan[1], R. E. Killick[1], R. J. H. Dunn[1], T. J. Osborn[2], P. D. Jones[2] and I. R. Simpson[1]

[1] Met Office Hadley Centre, Exeter, EX1 3PB, UK

[2] Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, NR4 7TJ, UK

Corresponding author: Colin Morice (colin.morice@metoffice.gov.uk)

Key Points:

- We have created a new version of the Met Office Hadley Centre and Climatic Research Unit global surface temperature dataset for 1850-2018.
- The new dataset better represents sparsely observed regions of the globe and incorporates an improved sea-surface temperature dataset.
- This dataset shows increased global average warming since the mid-nineteenth century and in recent years, consistent with other analyses.

Abstract

We present a new version of the Met Office Hadley Centre/Climatic Research Unit global surface temperature dataset, HadCRUT5. HadCRUT5 presents monthly average near-surface temperature anomalies, relative to the 1961-1990 period, on a regular 5° latitude by 5° longitude grid from 1850 to 2018. HadCRUT5 is a combination of sea-surface temperature measurements over the ocean from ships and buoys and near-surface air temperature measurements from weather stations over the land surface. These data have been sourced from updated compilations and the adjustments applied to mitigate the impact of changes in sea-surface temperature measurement methods have been revised. Two variants of HadCRUT5 have been produced for use in different applications. The first represents temperature anomaly data on a grid for locations where measurement data are available. The second, more spatially complete, variant uses a Gaussian process based statistical method to make better use of the available observations, extending temperature anomaly estimates into regions for which the underlying measurements are informative. Each is provided as a 200-member ensemble accompanied by additional uncertainty information. The combination of revised input datasets and statistical analysis results in greater warming of the global average over the course of the whole record. In recent years, increased warming results from an improved representation of Arctic warming and a better understanding of evolving biases in sea-surface temperature measurements from ships. These updates result in greater consistency with other independent global surface temperature datasets, despite their different approaches to dataset construction, and further increase confidence in our understanding of changes seen.

Plain Language Summary

We have produced a new version of a dataset that measures changes of near-surface temperature across the globe from 1850 to 2018, called HadCRUT5. We have included an improved dataset of sea-surface temperature, which better accounts for the effects of changes through time in how measurement were made from ships and buoys at sea. We have also included an expanded compilation of measurements made at weather stations on land.

There are two variations of HadCRUT5, produced for different uses. The first, the "HadCRUT5 non-infilled dataset", maps temperature changes on a grid for locations close to where we have measurements. The second, the "HadCRUT5 analysis", extends our estimates to locations further from the available measurements using a statistical technique that makes use of the spatial connectedness of temperature patterns. This improves the representation of less well observed regions in estimates of global, hemispheric and regional temperature change.

Together, these updates and improvements reveal a slightly greater rise in near-surface temperature since the nineteenth century, especially in the Northern Hemisphere, which is more consistent with other datasets. This increases our confidence in our understanding of global surface temperature changes since the mid-nineteenth century.

1 Introduction

Observational evidence plays an essential role in our understanding of the climate, the causes of the observed changes and distance travelled along predicted future trajectories. Compilations of near-surface temperature measurements, as traditionally measured over land in shielded enclosures and at sea by ships and buoys, as well as multi-decadal temperature records derived from these compilations, are a core repository of information underpinning our understanding of a changing climate. Here we present an update to one such assessment, the Met Office Hadley Centre/Climatic Research Unit HadCRUT dataset (version HadCRUT.5.0.0.0, referred to hereafter as HadCRUT5), incorporating additional measurements, improved understanding of non-climatic effects associated with an ever-changing measurement network, and updated gridding methods.

Global near-surface temperature analyses, based on a combination of air temperature observations over land with sea-surface temperature (SST) observations, are among the longest instrumental records of climate change and variability. They are routinely used in assessments of the state of the climate (e.g. Blunden & Arndt, 2019). They underpin our understanding of multi-decadal to centennial changes and the causes of those changes (e.g. Hartmann et al., 2013) and are a key metric against which climate change policy decisions are made and progress against international agreements is measured (e.g. Allen et al., 2018).

Analyses of multi-decadal temperature changes based on instrumental evidence are subject to uncertainty. Assessments of uncertainty and the influence of non-climatic factors on observations are necessary to understand the evolution of near-surface temperature throughout the instrumental period. Known sources of uncertainty include spatial and temporal sampling of the globe (Jones et al., 1997; Brohan et al., 2006), changes in measurement practice and instrumentation (Parker 1994; Kent et al., 2017), siting of observing stations and the effects of changes in their nearby environment (Parker 2006; Menne et al., 2018), and basic measurement error.

Since the release of the predecessor of the dataset presented here, HadCRUT4 (Morice et al., 2012), new analyses of near-surface temperature have been undertaken, and with them understanding has improved of deficiencies in the observing network and in analysis methods. This has led to updates to analyses with long pedigrees (Zhang et al., 2019; Lenssen et al., 2019), the arrival of new and independent analyses (Rohde et al., 2013a; 2013b; Rohde & Hausfather, 2020; Yun et al., 2019), and related studies (Ilyas et al., 2017; Benestad et al., 2019; Kadow et al., 2020).

Efforts to consolidate archives of instrumental air temperature series under the auspices of the International Surface Temperature Initiative (ISTI; Rennie et al., 2014) have greatly increased the availability of meteorological station series. The resulting ISTI databank underpins the updated GHCNv4 air temperature data set (Menne et al., 2018) and regional subsets of station series from the ISTI databank have been selectively included in updates to the CRUTEM4 and CRUTEM5 datasets (Jones et al., 2012; Osborn et al., 2020). These improved data holdings have increased observational coverage of regions that were previously poorly represented, including the rapidly warming high northern latitudes.

Rohde et al. (2013a; 2013b) introduced a new land air temperature analysis developed independently of pre-existing studies. This analysis included a new method for bias-adjusting station records, a process that is commonly known as homogenization, and combined estimation of homogenization adjustments with an independently developed spatial analysis method. The study has since been extended to include analysis of HadSST3 sea-surface temperatures (Kennedy et al., 2011a; 2011b) to produce a merged land-sea data product (Rohde & Hausfather, 2020).

A key uncertainty for estimating long-term change is that associated with corrections for systematic errors in sea-surface temperature measurements. Comparisons of long historical SST data sets (Kent et al., 2017) showed that there were differences between SST data sets which were larger than the estimated uncertainties. A comparison to modern "instrumentally homogeneous" data sets by Hausfather et al. (2017), found that HadSST3 (Kennedy et al., 2011a; 2011b) and COBE-SST-2 (Hirahara et al. 2014) underestimated recent warming. Cowtan et al. (2018) compared SST products to coastal weather stations highlighting discrepancies between temperature trends in land and ocean data sets. Carella et al. (2018) used characteristic daily-cycles in SST measurements to infer how the measurements were made and showed that previous assumptions under-estimated the prevalence of engine-room measurements.

Freeman et al. (2017) compiled release 3.0 of the International Comprehensive Ocean Atmosphere Data Set (ICOADS) including newly digitized data. Two long-term historical SST analyses, HadSST and ERSST, which are based on ICOADS, have been updated using this new release. ERSST has gone through two updates – version 4 (Huang et al., 2016) and 5 (Huang et al., 2017) – which extended bias adjustments to the whole SST record, implemented improvements to the analysis, and quantified uncertainty. HadSST.4.0.0.0 (Kennedy et al., 2019) revisited the bias adjustments applied to the data, using oceanographic measurements to understand and reduce some of the key uncertainties in HadSST3.

Recent updates to instrumental near-surface temperature data products have brought improvements in their assessment of uncertainty, and in provision of uncertainty information for use in onward analyses. Ensemble uncertainty assessments have become commonplace in air temperature datasets (Morice et al., 2012; Menne et al., 2018) and sea-surface temperature datasets (Kennedy et al., 2011b; Huang et al., 2016; Huang et al., 2019; Kennedy et al., 2019).

The NOAAGlobalTemp version 5 analysis (Zhang et al., 2019; Huang et al., 2019) updates previous NOAA analyses (Smith et al., 2008) by bringing together updates to underpinning data holdings over land (Menne et al., 2018) and merges the expanded land data holdings of GHCNv4 with the updated ERSSTv5 data set. An ensemble uncertainty assessment is included (Huang et al., 2019), sampling the uncertainty in parametric choices in the SST adjustments procedure, the station series homogenization algorithm (Menne et al., 2018) and the spatial analysis method used.

The NASA Goddard Institute for Space Studies GISTEMPv4 analysis (Lenssen et al., 2019) introduces an updated uncertainty assessment, applying the GISTEMP spatial analysis methods to the 100-member GHCNv4 ensemble of homogenized station series and basing SST uncertainty assessments on the ERSSTv4 ensemble. Additional uncertainty associated with the production of spatial analyses from incomplete station data is assessed by sub-sampling reanalysis fields from a selection of modern reanalyses.

Coverage of instrumental records of near-surface temperature changes is characterized by often sparse and non-uniform sampling of the globe. Assessments of uncertainty in global and regional average temperature changes have found that sparse data coverage is the most prominent source of uncertainty over monthly to decadal timescales (Brohan et al., 2006; Morice et al., 2012), outweighing uncertainty arising from changes in observing methods. Recent studies have also shown that poor representation of some regions, notably the rapidly warming high northern latitudes, may have contributed to an underestimation of globally averaged temperature changes in recent years (Cowtan and Way, 2014; Karl et al., 2015).

While efforts have been made to increase data coverage in the CRUTEM4 and now CRUTEM5 data set through inclusion of additional meteorological station data in less well-observed regions (Jones et al., 2012; Osborn et al., 2020) and marine data holdings expanded to include recently digitized marine reports (Freeman et al. 2017), statistical analysis methods were not used in HadCRUT4 or its underpinning land and marine datasets to infer temperature changes in regions where measurements are not available. An independent application of local statistical interpolation methods to HadCRUT4, in a study by Cowtan and Way (2014), found that statistically infilled reconstructions showed recent warming over high latitude regions that is not proportionately represented in global mean temperatures calculated from the non-infilled HadCRUT4 data set. The study also included an analysis that used satellite-based upper air temperature estimates as a proxy for near-surface temperature variability in the gaps in data coverage in HadCRUT4, which also showed warming in these high latitude regions. This high-latitude signal contributed to an increase in the assessed rate of change of global average temperatures since the beginning of the 21st century.

Unlike HadCRUT4, other existing near-surface temperature datasets utilize statistical analysis methods to infer spatial fields from scattered observations. Analysis methods based on spatial covariance structure, known variously as optimal interpolation (e.g. as used in Reynolds & Smith, 1994), kriging (e.g. as used in Cowtan & Way, 2014), Gaussian process regression (Rasmussen & Williams, 2006) and variants thereof, have a long history of use, particularly in analyses of sea-surface temperatures (Reynolds et al., 2002; Reynolds & Smith, 1994; Donlon et al. 2012). These methods use knowledge of the covariance structure of spatial fields to infer field values as weighted averages of observations in locations with strong covariation. Typically, weighting is based on a statistical model in which nearby locations are expected to covary strongly and distant locations weakly. Methods of this form are a core part of the Rohde & Hausfather (2020) analysis and of the analysis of Cowtan and Way (2014). The GISTEMP data set also uses a distance-weighted average that, while similarly applying a weighted average of local observations, does not make use of a covariance model and so does not classify as a kriging type analysis.

A second form of spatial analysis methods that are commonly applied in instrumental climate analyses, reduced space methods, decompose spatial temperature variability into a finite, typically orthogonal, set of spatial patterns of variability (Kaplan et al., 1997). These patterns are generally, but not necessarily, global in extent. Spatial reconstructions are then formed as a weighted sum of these patterns. The Empirical Orthogonal Teleconnection (Smith et al., 2008) method employed within the NOAAGlobalTemp v5 analysis falls within this category of reduced space algorithms, employing a finite set of locally defined spatial patterns that are fit to the available data.

A recent assessment of the use of neural networks to estimate missing values in the HadCRUT4 dataset (Kadow et al., 2020) expands the ensemble of methods used to reconstruct global temperatures. Derived global temperature series show good agreement with prior studies using more traditional methods.

Traditionally, surface temperature data sets have combined air temperatures over land with sea-surface temperatures over the ocean, rather than the more natural choice of air temperatures over the ocean. SST measurements are currently far more numerous than marine air temperature (MAT) measurements because SST can be readily measured by automatic sensors on drifting buoys as well as being retrieved from satellite measurements of radiances, while observational sampling of MAT has been in recent decline (Berry & Kent., 2017). There are significant biases in daytime marine air temperature observations (Berry et al., 2004). Night-time measurements have therefore been used to develop observational records of marine air temperature changes (Kent et al. 2013), with up-to-date independent assessments of historical night-time MAT becoming available only recently (Junod & Christy 2020, Cornes et al., 2020). Anomalies in MAT and SST have been expected to be similar over long space and time scales due to the strong physical link between the two. However, Cowtan et al. (2015) showed that MAT and SST changes simulated in coupled climate models differ, with MAT warming slightly faster than SST, affecting comparisons of observed and modelled global temperature change if care is not taken to ensure an “apples to apples” comparison. They also found that decisions about how to handle marginal sea-ice areas could affect the estimated changes, depending on the use of SST or MAT. Therefore, while there is good motivation for the use of MAT (Cowtan et al., 2015; Richardson et al., 2016), there are currently challenges relating to the MAT observational network (Berry & Kent, 2017) that provide an observational rationale for the continued use of SST in monitoring global surface temperature variability and change until these challenges are addressed.

Recent developments in satellite retrievals of surface skin temperatures present a new possibility for near-surface temperature monitoring, bringing the potential for detailed spatial information with sustained measurement over a time frame that is now of sufficient length for climate studies. Recent work (Rayner et al., 2020) has explored the potential of combining air temperature information inferred from satellite skin temperatures with traditional *in situ* observations, expanding on the understanding of relationships between satellite-derived skin temperatures and traditional near-surface air temperature observations, and on the stability of these relationships over time that is required to construct merged data products. Alternatively, dynamical reanalyses, that combine numerical weather prediction models with a range of varied observational data sources, are increasingly being used to monitor the climate (e.g. ERA5, Hersbach et al., 2020; JRA-55, Kobayashi et al., 2015; and MERRA-2, Gelaro et al., 2017). These alternative sources of near-surface temperature data provide useful information in locations that are not well represented in traditional near-surface temperature datasets. However, in all cases, understanding of non-climatic effects affecting observations and arising from analysis methods is required when combining observations from multiple sources.

Here, two ensemble surface temperature datasets are presented. The first, the “HadCRUT5 non-infilled dataset”, adopts the gridding and ensemble generation methods of HadCRUT4 (Morice et al., 2012). The second, the “HadCRUT5 analysis”, uses a statistical infilling method to improve the representation of sparsely observed regions. Through application of the statistical infilling method to the HadCRUT5 non-infilled ensemble, the HadCRUT5 analysis ensemble samples the uncertainty in the gridded near-surface temperature data that arises from residual biases in observational data after correction, for example associated with uncertainty in changes in instrumentation and measurement practices at meteorological stations (Brohan et al., 2006; Morice et al., 2012) and changes in sea-surface temperature measurement methods (Kennedy et al., 2019). It also samples the effects of basic measurement uncertainty, uncertainty arising from estimation of gridded temperature fields from a finite number of observations and residual uncertainties associated with individual marine measurement platforms, where information identifying individual platforms is available (Kennedy et al., 2019). Statistical reconstruction uncertainty is also encoded in the HadCRUT5 analysis ensemble, producing an ensemble that samples a greater range of sources of uncertainty than was possible in HadCRUT4. Thus, the new ensemble analysis communicates the major known sources of uncertainty in an easily accessible way.

The remaining sections of this paper are structured as follows. Section 2 describes the data sets used as inputs and for comparison. Section 3 provides an overview of the methods used to construct HadCRUT5. Results are presented in Section 4 with conclusions and discussion in Section 5.

2 Input Datasets

2.1 HadSST.4.0.0.0

Version 4 of the Met Office Hadley Centre Sea-Surface Temperature data set, HadSST.4.0.0.0 (Kennedy et al., 2019), is based on *in situ* measurements of SST from ships and buoys. The ship and buoy measurements are taken from ICOADS release 3.0 (Freeman et al. 2017) from 1850 to 2014 and release 3.0.1 from 2015 to 2018. From 2016 onwards, measurements from drifting buoys are taken from the Copernicus Marine Environment Monitoring Service, as buoy data in ICOADS were incomplete following a change in data-transmission codes in late 2016. Early measurements made using buckets are adjusted using a physically based model of heat lost from water-sampling buckets (Folland and Parker 1995; Rayner et al., 2006). From the 1940s onwards, ship measurements are adjusted based first on comparisons with near-surface oceanographic measurements (Atkinson et al., 2014) and then, from the early 1990s onwards, on comparisons with buoy measurements. The resulting HadSST.4.0.0.0 data set is presented as anomalies relative to 1961-1990 on a 5° latitude by 5° longitude grid and is representative of SST as measured by drifting buoys at an approximate depth of 20 cm.

Overall, the global SST change estimated from HadSST.4.0.0.0 is larger than that estimated from HadSST.3.1.1.0 (and earlier versions). This is due to two factors. First, new estimates of biases associated with measurements made in ships' engine rooms show that these biases have declined since the 1950/60s, artificially reducing the long-term change represented in the underlying data and in earlier versions of HadSST. Second, a greater proportion of measurements during the 1961-1990 period were estimated to have been made in ships' engine rooms. Other changes include: using buoys as a reference data set; producing ensemble members with step changes in the time evolution of the proportions of canvas and wooden buckets in the early 20th century alongside ensemble members which assume a linear transition; estimating the fraction of incorrect metadata using comparisons with oceanographic measurements; and using comparisons with oceanographic measurements to narrow the range of plausible transition dates from canvas buckets to modern rubber buckets (see Kennedy et al. (2019) for a detailed discussion).

Uncertainty in HadSST.4.0.0.0 is split into three main components associated with: pervasive systematic errors; systematic errors from individual ships or buoys; and uncorrelated errors from individual measurements and incomplete grid-box sampling. The pervasive systematic errors, which have complex temporal and spatial correlations, are represented using a 200-member ensemble generated by varying uncertain parameters and choices in the bias adjustment scheme. The systematic errors are represented using covariance matrices that encode the error covariances between grid cells that arise from ships making measurements in multiple grid cells in a month. Finally, uncertainties from uncorrelated errors are provided as gridded fields. Note that these uncertainty components do not span the full range of uncertainty. In particular, structural uncertainty remains (Thorne et al., 2011) and there may be an underestimate in the systematic error component because it does not currently deal explicitly with errors that correlate at the level of national shipping fleets (Chan & Huybers, 2019) or with marine reports that lack ship call signs or other identifying information (Carella et al., 2017).

2.2 CRUTEM.5.0.0.0

Monthly averages of near-surface air temperature measured at weather stations over the land surface for 1850-2018 are obtained from CRUTEM.5.0.0.0 (Osborn et al., 2020, referred to hereafter as CRUTEM5). The CRUTEM station database is a collection of station series obtained from National Meteorological and Hydrological Services (NMHSs) and large collections such as the European Climate Assessment and Dataset (Klein Tank et al., 2002). CRUTEM incorporates corrections that NMHSs apply to their own data to minimize the impact of changes in weather station instrumentation or location on the measurement series. The monthly average temperatures from stations are subjected to quality control, converted to anomalies (differences from their 1961–1990 means) and then averaged into 5° latitude by 5° longitude grid boxes.

CRUTEM5 has improved quality control checks that: (i) improve the flagging of incorrect data during 1941-1990; (ii) reduce the trend towards increased flagging of suspect data outside of the 1941-1990 period; and (iii) reduce the number of genuine extreme values that are erroneously flagged as incorrect, e.g. during coherent extreme events such as summer 2003 in Europe (see Osborn et al. (2020) for details). The station database has been expanded such that the number of those stations with sufficient data to estimate temperature anomalies has grown from 4842 in CRUTEM.4.0.0.0 (as used in Morice et al., 2012) to 7983 in CRUTEM5 (Osborn et al., 2020). Most of the new data acquisitions are in already-sampled regions, so the number of grid-box values is only moderately expanded (by 9%) relative to CRUTEM.4.0.0.0.

The changes in temperature seen in hemispheric or global averages since 1850 are not sensitive to these updates, but some regional differences are apparent. Osborn et al. (2020) describes the effects of updates since CRUTEM.4.0.0.0, and of updates since the more recent CRUTEM.4.6.0.0 (as used in HadCRUT.4.6.0.0), in detail.

An alternative gridding method was explored in Osborn et al. (2020) for CRUTEM5 to address the under-representation of high latitude stations in the standard gridding. This alternative method allows each station to contribute to more than one neighboring 5° latitude by 5° longitude grid box on the same latitude, where the number of grid boxes to which each station can contribute is determined by an inverse cosine latitude weighting. In the current paper, the alternative gridding method is not used because (a) the uncertainty model for the CRUTEM5 grids, as documented in Brohan et al. (2006), only applies to the standard gridding approach (where each station contributes to only one grid box); and (b) the issue of high-latitude sampling is dealt with here by statistical infilling.

HadCRUT5 uses an ensemble version of the CRUTEM5 uncertainty model. The HadCRUT5 non-infilled ensemble grids and accompanying uncertainty grids are produced from the CRUTEM5 station temperature anomaly series, following the methods of Morice et al. (2012), as described in Section 3.2.

2.3 HadISST.2.2.0.0

We use sea ice concentration from the Met Office Hadley Centre sea-Ice and Sea Surface Temperature data set, HadISST.2.2.0.0 (an update to Titchner and Rayner (2014)), on a 1° latitude by 1° longitude grid to determine the presence or absence of sea ice in any individual ocean grid box in each month from 1850 to 2018.

HadISST.2.2.0.0 is updated relative to version 2.1.0.0 in the following ways: (i) reinstatement of a small number of erroneously-removed sea-ice-filled grid boxes after 1978; (ii) an alteration to the adjustments applied to the National Ice Center charts (used to determine the ice edge between 1972 and 1978) correcting a low-bias in the HadISST.2.1.0.0 fields in the Arctic then; and (iii) an improvement in the interpolation applied between two atlas-derived climatologies used to determine ice extents in the Antarctic to produce a smoother transition between them and between 1962 and the start of monthly observations in 1972.

2.4 ERA5

We have used monthly ERA5 analysis 2 m air temperature data from 1979-2018 (Hersbach et al., 2020) for coverage uncertainty estimation and for comparison of global and regional diagnostics. ERA5 was produced using 4D-Var data assimilation in the European Centre for Medium-range Weather Forecasts' (ECMWF) Integrated Forecast System (IFS). We used the (31 km) high resolution realization.

2.5 Other comparison data

Four comparison data sets are used here: NOAAGlobalTemp version 5 (Zhang et al., 2019; Huang et al., 2019), GISTEMP version 4 (Hansen et al., 2010; Lenssen et al., 2019), Berkeley Earth (Rohde & Hausfather, 2020) and Cowtan & Way (Cowtan & Way, 2014).

NOAAGlobalTemp version 5 is based on the Global Historical Climatology Network (GHCN) version 4 land station data set (Menne et al., 2018) and the Extended Reconstruction Sea Surface Temperature (ERSST) data set version 5 (Huang et al., 2017). Station records in GHCN v4 are homogenized using an automated algorithm. SSTs are adjusted using comparisons with marine air temperature and latterly drifting buoys. The data are interpolated using Empirical Orthogonal Teleconnections, providing improved coverage, although coverage does not extend fully into the polar regions.

GISTEMP version 4, like NOAAGlobalTemp v5, is based on a combination of GHCN v4 and ERSST v5. The SST data are interpolated as in NOAAGlobalTemp. Land surface air temperatures are interpolated from station data within a 1200km radius. Extrapolated land surface air temperatures are used over the oceans in sea-ice covered areas. Coverage of the GISTEMP data set is quasi-global in the past twenty years, with good coverage of the poles and other data-sparse regions from interpolated station data.

The Berkeley Earth data set (Rohde & Hausfather, 2020) uses a kriging-based technique to interpolate and homogenize station data. A kriging based technique is also applied to SSTs from the HadSST3 data set to provide coverage over the whole globe. The version of the data set that uses extrapolated land-surface air temperatures over the oceans in sea-ice covered areas is used here.

Cowtan and Way (2014) is based on the HadCRUT4 data set. The land and ocean data are interpolated using kriging. Grid cells that contain data in HadCRUT4 are not modified during interpolation (in contrast to the kriging of HadSST3 data in the Berkeley Earth data set). As with GISTEMP and Berkeley Earth, extrapolated land-surface air temperatures are used over the oceans in sea-ice covered areas.

The Berkeley Earth (1° latitude by 1° longitude resolution) and ERA5 (0.25° latitude by 0.25° longitude resolution) analyses were regridded to 5° latitude by 5° resolution using an area-weighted average of all grid cells falling within a HadCRUT5 5° grid cell. Cowtan and Way and NOAAGlobalTemp were obtained on a 5° grid. The GISTEMP data, which were obtained on a 2° grid, were not regridded.

3 Methods

Two gridded datasets are provided as part of HadCRUT5. The first version of the dataset is produced without statistical infilling, referred to here as the “HadCRUT5 non-infilled dataset”, following the methods of Morice et al. (2012), and is intended for use in applications where statistical infilling is not desired. This is accompanied by a second version of the dataset, hereafter referred to as the “HadCRUT5 analysis”, that is produced using a statistical method to estimate more-complete temperature anomaly fields.

The HadCRUT5 non-infilled dataset and the HadCRUT5 analysis are produced in the following steps. First, an ensemble land-surface air temperature dataset, with accompanying additional uncertainty information, is generated from the CRUTEM5 station data (Section 3.2). The land dataset is then merged with sea-surface temperature anomaly information from HadSST4 through a weighting method based on the land area fraction (Section 3.4) to produce the non-infilled dataset. Next, monthly fields are estimated separately for the land surface air temperature dataset and for HadSST4 using a statistical method to create an ensemble analysis for each (Section 3.3). The separate land and ocean analyses are then merged into a combined land and ocean ensemble analysis using a land-sea weighting scheme that also accounts for sea ice coverage (Section 3.4). Global and regional time series are then computed from the two merged datasets, following the methods of Morice et al. (2012) with updates to the method used to estimate uncertainty associated with incomplete observational coverage described in Section 3.5. Error models for each dataset are described in Section 3.1. Full details of uncertainty propagation for land and ocean merging and global and regional time series are provided in the Supporting Information.

3.1 The HadCRUT5 error models

This section outlines the terms of the error model for grids and time series of the HadCRUT5 non-infilled dataset and the HadCRUT5 analysis. Further details are given in the Supporting Information.

The error models are split into components according to the way that uncertainty information is presented in HadCRUT5. The sources of uncertainty modelled in HadCRUT5 are grouped according to their correlation structure to allow uncertainties to be propagated appropriately into derived diagnostics such as regional average time series.

3.1.1 The HadCRUT5 non-infilled dataset

The error model for the non-infilled dataset describes the estimate of temperature anomaly $\hat{T}(s,t)$ at spatial location s and time t as a sum of the true temperature anomaly $T(s,t)$ and three error terms: a bias term $\varepsilon_b(s,t)$ representing biases with large-scale spatial and temporal structure; a partially correlated error term $\varepsilon_p(s,t)$ for errors with typically local structure; and an uncorrelated error term $\varepsilon_u(s,t)$ describing errors that are independent between spatial and temporal locations. The full error model for non-infilled fields is given by:

$$\hat{T}(s,t) = T(s,t) + \varepsilon_b(s,t) + \varepsilon_p(s,t) + \varepsilon_u(s,t) \qquad (1)$$

This error model for the merged dataset matches the structure of the error model for the land dataset and for HadSST4. For the land dataset, the contributions to the bias term are the land station homogenization error, urbanization and biases from non-standard measurement enclosures. There is no contribution to the partially correlated term and the uncorrelated term models the within grid box measurement and sampling uncertainties (Morice et al., 2012). For HadSST4, the bias term models the effects of residual errors in the adjustments applied to account for changes in measurement methods, the partially correlated term models the effects of residual biases associated with individual observing platforms, and the uncorrelated term models the within grid cell measurement and sampling uncertainties (Kennedy et al., 2019).

The HadCRUT5 non-infilled ensemble samples the uncertainties for the combination $T(s,t) + \varepsilon_b(s,t)$. The uncertainties for partially correlated and uncorrelated errors are not encoded into the non-infilled ensemble. Instead, uncertainty information for partially correlated errors $\varepsilon_p(s,t)$ is provided in spatial error covariance matrices and uncertainties for uncorrelated errors $\varepsilon_u(s,t)$ are provided for each observed grid cell.

The error model for estimates of spatial average time series $\hat{T}(t)$ derived from the gridded data is given as a sum of the true temperature anomaly time series $T(t)$ and four error terms:

$$\hat{T}(t) = T(t) + \varepsilon_b(t) + \varepsilon_p(t) + \varepsilon_u(t) + \varepsilon_c(t) \tag{2}$$

Here $\varepsilon_b(t)$ is the effect of the bias term propagated into the spatial average, $\varepsilon_p(t)$ is the effect of the partially correlated term, $\varepsilon_u(t)$ the effect of the uncorrelated error term. The fourth error term, $\varepsilon_c(t)$, is the error in estimating the spatial average from incomplete spatial coverage, with missing grid cells resulting from limitations in the spatial sampling provided by the observation network. Full details of uncertainty propagation for each of these terms are given in the Supporting Information.

3.1.2 The HadCRUT5 analysis

An overview of the HadCRUT5 analysis is provided in Section 3.4 and a detailed description of methods is provided in Appendix A. The HadCRUT5 analysis error model has fewer terms than that of the non-infilled dataset as the analysis methods combine multiple sources of error into a single analysis error term. The error model for the HadCRUT5 analysis defines the temperature anomaly estimate as the sum of the true temperature $T(s,t)$ and the analysis error $\varepsilon_a(s,t)$:

$$\hat{T}(s,t) = T(s,t) + \varepsilon_a(s,t) \tag{3}$$

The analysis error term combines all errors that are modelled in the Gaussian process analysis, both spatial reconstruction errors and observational errors, as described in Appendix A. The analysis ensemble samples the analysis uncertainty such that each ensemble member is a sample of $T(s,t) + \varepsilon_a(s,t)$.

For the HadCRUT5 analysis, errors in global and regional average time series are derived as a combination of the propagated analysis error and $\varepsilon_a(t)$ and an additional coverage error term $\varepsilon_c(t)$ that represents the error in estimating the spatial average from incomplete analysis grids, noting that this coverage error term differs from that of the non-infilled dataset due to the different spatial coverage of the analysis.

$$\hat{T}(t) = T(t) + \varepsilon_a(t) + \varepsilon_c(t) \tag{4}$$

The propagation of uncertainty associated with these errors is described in the Supporting Information.

3.2 Ensemble land air temperature data set

As in the previous versions of HadCRUT, near-surface air temperature information over land is derived from the CRUTEM data set. As in Morice et al. (2012), an ensemble air temperature data set is produced by sampling from the distributions of known uncertainty in station temperature records. The station data on which the ensemble grids are based has been updated to now use the CRUTEM.5.0.0.0 data set (Osborn et al., 2020).

A detailed description of the land air temperature ensemble sampling method can be found in Morice et al. (2012). The sampling approach is designed so that the effects of known sources of residual systematic error in station anomaly series can be quantified for regional statistics and time series. The ensemble size has been increased to 200 members for HadCRUT5 to match the 200-member HadSST4 ensemble.

The sampling method is as follows. Samples are drawn from the distributions of known uncertainties during the station gridding process. Residual homogenization error and uncertainty in climatology normal information are sampled from distributions described in Brohan et al. (2006) and encoded into realizations of individual station series prior to gridding. The systematic effects of residual urbanization errors (Brohan et al., 2006; Parker, 2010) and non-standard sensor enclosures (Parker, 1994; Folland et al., 2001) are sampled and encoded into the gridded ensemble at a regional level, again following the method of Morice et al. (2012).

Additional uncertainty information for errors that are uncorrelated between grid cells (e.g. from measurement error or incomplete sampling of a grid cell) is not encoded into the land ensemble. Instead, these measurement and sampling-related uncertainties are provided as additional uncertainty information outside of the ensemble, as in Morice et al. (2012).

3.3 Spatial analysis of temperature anomaly fields

HadCRUT5 now includes an ensemble spatial analysis that reconstructs more spatially extensive anomaly fields from the available observational coverage. The purpose of this analysis is to: (1) reduce uncertainty and bias associated with estimation of global and regional climate diagnostics from incomplete and uneven observational sampling of the globe; (2) provide improved estimates of temperature fields in all regions; and (3) provide a method to quantify uncertainty in anomaly patterns.

We adopt a Gaussian process based method for spatial analysis that is closely related to the ordinary kriging approach (Rasmussen & Williams, 2006), and apply the method independently to land air temperature and sea-surface temperature observations before merging the two to produce a global analysis. The method models monthly temperature anomaly fields as realizations of a Gaussian processes with a simple covariance structure, defined as a function of the distance between locations, and an *a priori* unknown mean, and accounts for observational uncertainty. A detailed description of the analysis method is presented in Appendix A.

The Gaussian process method is applied to the 5° latitude by 5° longitude gridded anomaly fields of the land ensemble and the HadSST4 ensemble. The additional observational uncertainty terms that accompany these input ensembles are provided to the Gaussian process estimation as monthly error covariance matrices. The spatial reconstructions are based upon a model of the covariance structure of the 5° latitude by 5° longitude anomaly fields. This covariance structure is modelled using a Matérn covariance function, for which the covariance decays as a function of Euclidian distance between locations. The parameters of the Matérn covariance function are fitted separately for land air temperature and sea-surface temperature anomalies (see Appendix A.2), representing typical variability in each domain.

As a Bayesian method, the approach provides a framework for assessing analysis uncertainties and provides the capability to draw samples from the posterior distribution of the analysis. We generate an ensemble of field estimates through application of the analysis method to each input ensemble member and then drawing samples from the posterior distributions of the Gaussian process estimates. The land and ocean analysis ensembles combine all sources of uncertainty represented in the input gridded datasets whilst respecting the estimated covariance structure of the temperature anomaly field so that each ensemble member is a plausible spatial analysis of the temperature anomaly field.

The analysis has limited capability to reconstruct temperatures at long distances from available observations, as the field estimates are based on a model of local covariance structure. We therefore introduce criteria for excluding regions where there is not a strong observational constraint on the analysis (see Appendix A.4). The masked land air temperature and sea-surface temperature anomaly ensembles are then merged, as described in Section 3.4.

3.4 Blending land air temperatures with sea-surface temperature data

The 200-member ensemble land air temperature data set based on CRUTEM5 and the 200-member HadSST4 are merged as a weighted average of the 5° latitude by 5° longitude land and marine fields. Two versions of the data set are provided: one that uses the spatial analysis method presented in Section 3.2 and one that does not.

3.4.1 Merging non-infilled datasets

For the non-infilled dataset, the land air temperature ensemble and HadSST4 ensemble members are merged following the methods of Morice et al. (2012). The temperature anomaly $T(s,t)$ at location s and time t is defined as the weighted average of the air temperature anomaly $T_L(s,t)$ and sea surface temperature anomaly $T_M(s,t)$, with weights $f(s,t)$:

$$T(s,t) = f(s,t)T_L(s,t) + \big(1 - f(s,t)\big)T_M(s,t) \qquad (5)$$

The weighted average is based on the areal fraction of land and sea in a 5° latitude by 5° longitude grid cell using the same land fraction data set as HadCRUT4, originally derived from the Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA; Donlon et al., 2012) 0.05° land mask information. As in HadCRUT4, land air temperature information receives a minimum weighting of 25% to prevent island stations from receiving near-zero weighting. Where only one of the land air temperature or sea-surface temperature data sets are available, the available data source receives 100% weighting.

Methods for merging the uncertainty fields and measurement error covariance information for land and marine data sets are unchanged from those described in Morice et al. (2012) and are detailed in the Supporting Information.

3.4.2 Merging land and ocean analyses

The land-sea weighting scheme is modified for the HadCRUT5 analysis. Areas of sea ice are treated as if they were land in the weighting (consistent with the approach used by Cowtan & Way (2014)), so that temperature anomalies over sea ice are reconstructed as part of the air temperature analysis rather than the SST analysis.

Sea ice concentrations are obtained from the HadISST.2.2.0.0 data set. Where the ice concentration on the native 1° latitude by 1° longitude HadISST.2.2.0.0 grid exceeds 15%, the threshold value used to define the ice edge in Titchner and Rayner (2014), the area is considered to be ice covered for the purpose of deriving weights. Ice concentrations below 15% are treated as open water. For each HadISST.2.2.0.0 grid cell, a value of one is set if the sea-ice concentration is greater than 15% and zero otherwise. On the 5° latitude by 5° longitude HadCRUT5 grid, the fractional area of water covered by sea ice is then obtained through area-weighted averages of the non-land 1° grid cells of ones and zeroes. This area of ice-covered water is treated as land when deriving weights for land and ocean analyses.

The 25% minimum weighting for land air temperature is retained for any 5° latitude by 5° longitude grid cells that are observed in the non-infilled land air temperature data set so that information from island stations is not lost in the averaging. This minimum weighting is not applied in grid cells that are not directly observed. Reconstructed land air temperatures are not used over 100% sea regions where there are no land stations or sea ice and, similarly, interpolated SST is not used over 100% land regions. This prevents extrapolation of land air temperature far into ocean regions and prevents inland extrapolation of SSTs.

3.5 Estimating uncertainty arising from incomplete coverage

Spatial fields of temperature anomalies in the non-infilled HadCRUT5 data set and the HadCRUT5 analysis are not globally complete. Variability in regions of the world that are not represented in the spatial fields gives rise to uncertainty in global and regional time series. For the non-infilled HadCRUT5, the coverage uncertainty accounts for regions of the globe where insufficient observations are available to compute grid cell average anomalies in the underlying air temperature and SST data sets. For the HadCRUT5 analysis, the coverage uncertainty accounts for the masked regions of the analysis that are not well constrained by observations.

Coverage uncertainty is assessed by sub-sampling globally-complete reanalysis fields to the coverage of HadCRUT5 using the method presented in Brohan et al. (2006) and Morice et al. (2012), which is described in detail in the Supporting Information. The approach is updated here to use the recently-released ERA5 reanalysis (Hersbach et al., 2020) as the globally-complete reference data set, in place of the previously used NCEP/NCAR reanalysis (Kalnay et al., 1996).

Temperature anomalies are computed for the ERA5 monthly 2 m air temperature grids, referenced to the period of ERA5 that overlaps with our climatology period: 1979-1990. Anomalies are then averaged to the 5° latitude by 5° longitude grid used in HadCRUT5 to produce the reference fields for the coverage uncertainty calculations.

4 Results

4.1 Effects of updated data and methods in HadCRUT5

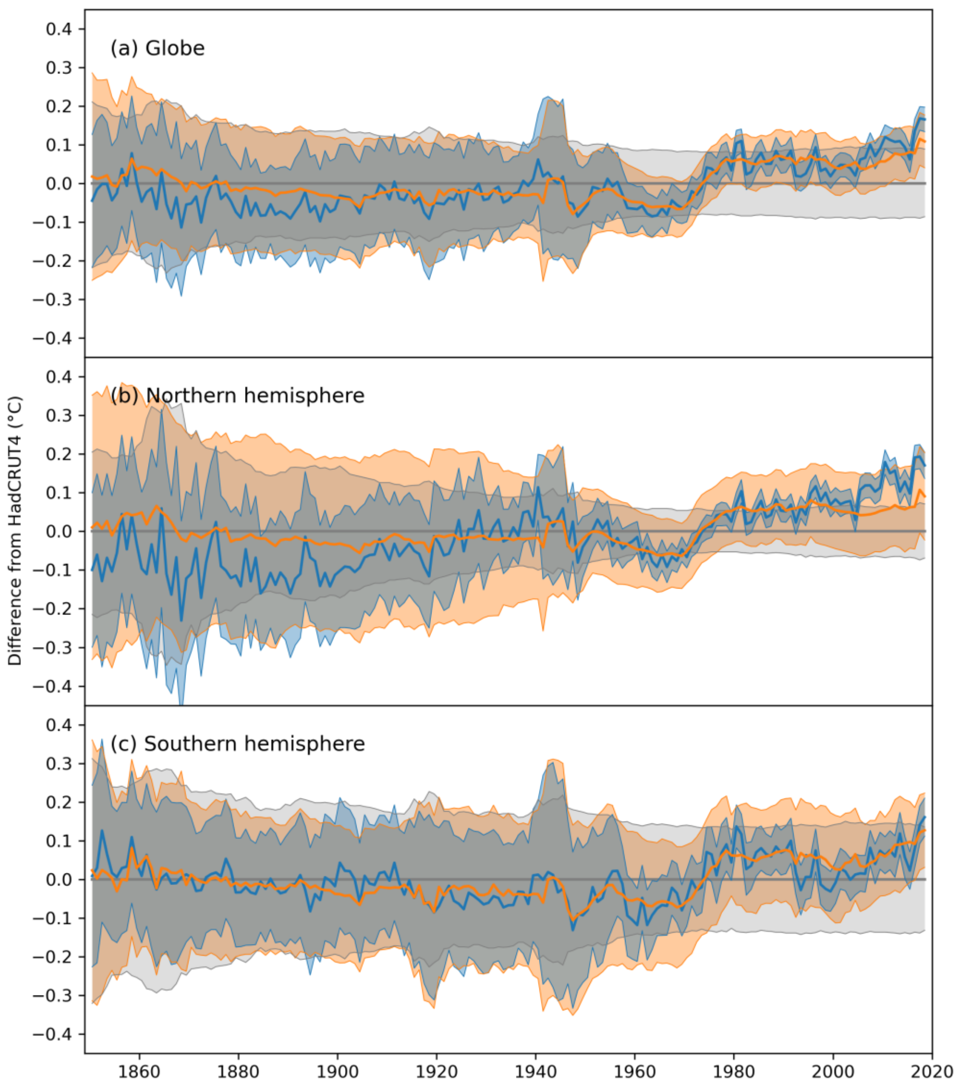


Figure 1. Annual average difference between HadCRUT.5.0.0.0 and HadCRUT.4.6.0.0 (°C), 1850-2018. **(a)** Globe, **(b)** Northern Hemisphere and **(c)** Southern Hemisphere. Orange: non-infilled HadCRUT5. Blue: HadCRUT5 analysis. Solid lines: ensemble mean (HadCRUT.5.0.0.0) or median (HadCRUT.4.6.0.0). Orange/blue shading: 95% confidence interval determined by the ensemble spread and coverage uncertainty (the blue shading for the HadCRUT5 analysis lies mostly within the orange shading, where it appears as a darker grey due to the overlap). Light grey shading: 95% confidence interval on HadCRUT.4.6.0.0. Global means have been calculated by averaging hemispheric anomaly series for northern and southern hemispheres with equal weighting given to each hemisphere.

Differences in global and hemispheric mean time series between HadCRUT4 (version HadCRUT.4.6.0.0) and the HadCRUT5 non-infilled data set and HadCRUT5 analysis are shown in Figure 1. The differences between the non-infilled HadCRUT5 and HadCRUT4 primarily arise from updates to the SST observational bias assessment in HadSST4. The updated bias corrections result in slightly cooler anomalies globally and in each hemisphere from the 1880s to 1970s. Anomalies are warmer from the 1980s onwards.

The most obvious difference is the relative warming of HadCRUT5 between around 1970 and 1980. This arises from improved estimates of biases in measurements made in ship engine rooms at that time. Engine room measurements were biased warm in the 1960s with the warm bias dropping over time, first between 1970 and 1980 and then again between the early 2000s and present. There are also changes around the Second World War, where changes to the assumptions made in HadSST4 about how measurements were taken shifted the mean and broadened the uncertainty range, reflecting the lack of knowledge of biases during this difficult period (Kennedy et al., 2019).

Northern hemisphere uncertainty estimates for the non-infilled HadCRUT5 are slightly wider that those of Morice et al. (2012). This results from a combination of the changes in the SST bias adjustment model and the adoption of ERA5 as the reference data set for coverage uncertainty calculations (Section 3.5). This change of reference data set typically gives wider uncertainty estimates in the northern hemisphere for similar observational coverage. The reverse is true in the southern hemisphere, with similar or slightly smaller coverage uncertainty estimates for the non-infilled HadCRUT5. This reflects differences in regional variability in sparsely observed regions between reanalysis products.

Further differences from HadCRUT4 can be seen in the HadCRUT5 analysis. Temperatures in the latter decades of the 19th century are on average cooler than in the non-infilled HadCRUT5 data set in the global and northern hemisphere series. Temperatures in the 21st century are on average warmer than those in the non-infilled HadCRUT5, primarily due to estimation of additional areas of warm anomalies in high latitude regions in the northern hemisphere, including use of air temperature anomalies over sea ice inferred from land stations. Rebalancing the representation of land and marine regions also affects average temperatures throughout the record. This is consistent with previous studies that adopt local interpolation methods (Cowtan & Way, 2014; Karl et al., 2015; Lenssen et al., 2019). Together these features result in greater warming throughout the 20th and 21st centuries in the HadCRUT5 analysis than is indicated by the non-infilled data set. However, for any given year, the effect of the reconstruction may be to produce either a warmer or cooler annual average and is dependent on variability in reconstructed regions that were not well represented in HadCRUT4 (see also Figure 5 (b) and (d)). Global and northern hemisphere HadCRUT5 analysis series fall outside the upper 95% uncertainty limit of HadCRUT4 in the 21st century but rarely depart from the uncertainty range of the HadCRUT5 non-infilled dataset, which includes the updated HadSST4 bias adjustments and has wider northern hemisphere coverage uncertainty ranges.

The uncertainty range for the HadCRUT5 analysis is narrower than that for the non-infilled data set, as the infilling effectively reduces the coverage uncertainty by filling gaps in the data and accounting for the non-uniform distribution of observations. The effect of this can be clearly seen in the Southern Hemisphere (Figure 1) where the narrowing of the uncertainty range before the 1950s is much less than after the 1950s, when routine monitoring on the Antarctic continent started, and coverage of the HadCRUT5 analysis thereafter approaches 100%.

As discussed in Section 3.1, the error model structure for the non-infilled HadCRUT5 data set is the same as in Morice et al. (2012), with observational bias adjustment uncertainties encoded into the ensemble and separate measurement and sampling uncertainty information provided and propagated into the uncertainty ranges on the hemispheric and global averages shown in Figure 1. The approach adopted for the HadCRUT5 analysis differs in including the effects of measurement and sampling uncertainties in the ensemble while also sampling from the spatial analysis uncertainty. Examples of HadCRUT5 analysis ensemble members are shown in Figure 2.

There is little change in the HadCRUT5 analysis ensemble spread for global or hemispheric averages from the 1970s onwards, reflecting the spread in the underlying SST ensemble and the relatively stable spatial sampling during this period. The ensemble spread in the global average in the 1940s is similar to that prior to the 1870s, though in the 1940s, this spread arises predominantly from uncertainty in the SST biases, whereas prior to the 1870s, the spread is largely due to uncertainty in the spatial field estimates due to limited observational sampling of the globe.

There is coherent spatial structure is the deviations of ensemble member fields from the ensemble mean. This results from uncertainty in the spatial analysis and its estimation from uncertain observations. Some ensemble members may be cool while others are warm in regions where uncertainty is high (for example see differences between ensemble members in Antarctica in Figure 2). The additional coverage uncertainty arising from masked regions is a relatively smaller component of the total uncertainty as a result of the increased coverage in the HadCRUT5 analysis fields and the inclusion of reconstruction uncertainty within the ensemble. On multi-annual timescales, the uncertainty in observational bias adjustments becomes prominent. This is reflected in persistently warm or cool departures from the ensemble mean in global and regional diagnostics over many years for individual ensemble members (for example see ensemble series in Figure 2).

Non-infilled HadCRUT5 ensemble members are shown in Figure 3, matching those shown for the HadCRUT5 analysis in Figure 2. HadCRUT5 analysis fields have greater spatial extent than the non-infilled dataset and are also smoother as a result of measurement and sampling uncertainties being taken into account within the analysis framework. In regions of few, scattered observations, infilled analysis fields have much greater extent but also show diversity in reconstructed anomaly patterns, reflecting uncertainty in the reconstruction in these sparsely observed regions.

Uncertainty ranges for the global average temperature series in Figure 3 show the ensemble spread in relation to the full uncertainty range, accounting for all quantified sources of uncertainty. While the HadCRUT5 analysis and non-infilled data set quantify uncertainty from the same error sources, the HadCRUT5 analysis encodes a greater portion of the uncertainty into the ensemble, whereas the non-infilled ensemble only samples uncertainties that are most important over multi-decadal time scales. The ensemble for the non-infilled HadCRUT5 data set samples the uncertainty associated with observational bias adjustments, with structure that is relevant to multi-decadal climate assessments. Unlike the HadCRUT5 analysis, measurement and sampling uncertainties that are relevant at shorter time scales are not encoded into the ensemble and are instead provided as auxiliary information. Uncertainty from incomplete global coverage of the observing network is a greater portion of the total uncertainty for the non-infilled data set. In contrast, for the HadCRUT5 analysis, the uncertainty from incomplete global coverage is divided between the analysis ensemble spread in reconstructed regions and a smaller coverage uncertainty term relating to regions that are masked.

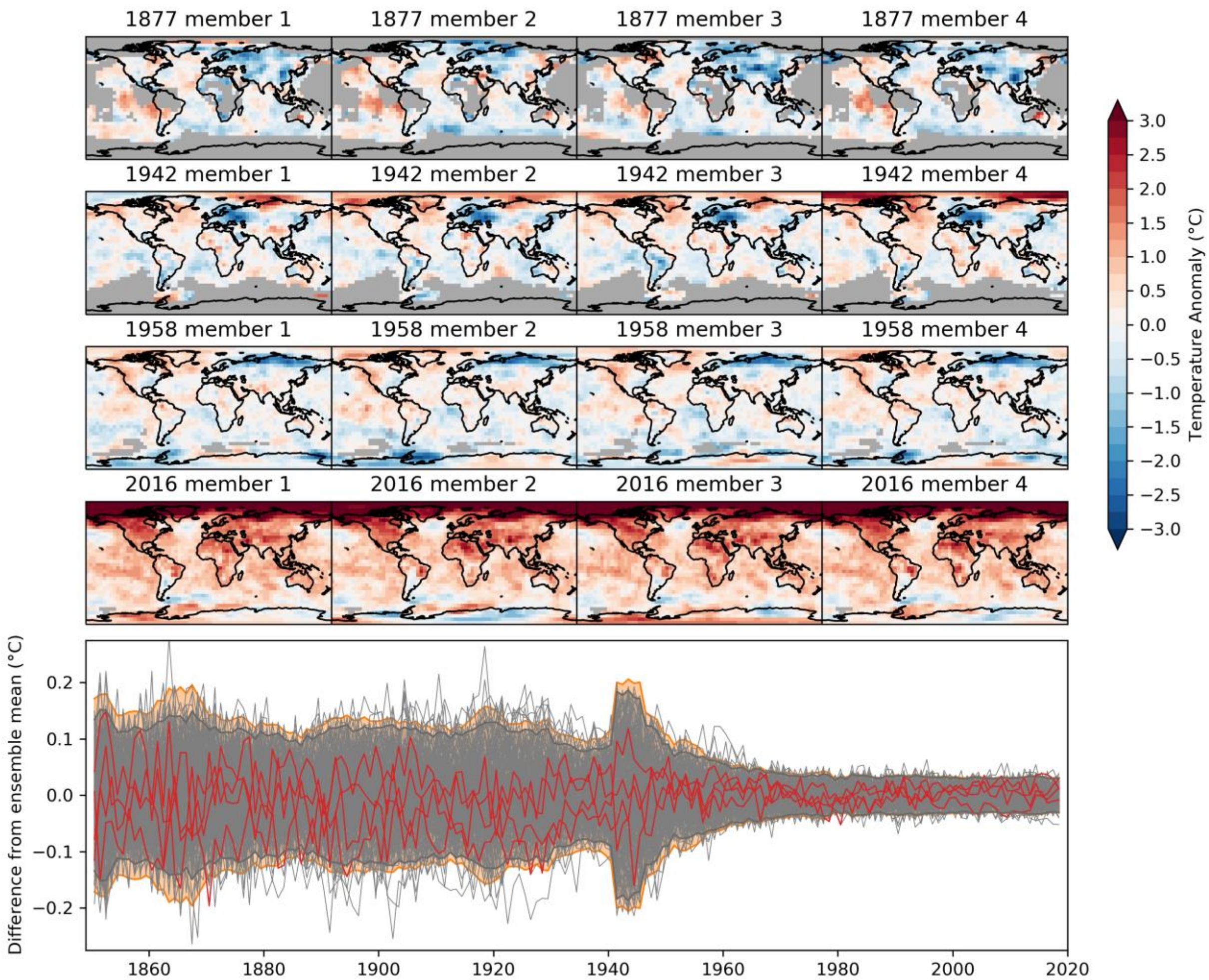


Figure 2. HadCRUT5 analysis ensemble members. Upper panel: annual average temperature anomaly (°C, relative to 1961-90) for 1877, 1942, 1958 and 2016 in four example ensemble members. Lower panel: ensemble spread in global mean (°C), 1850-2018. The difference between each ensemble member and the ensemble mean is shown by the grey lines, with the first four ensemble members (corresponding to the maps above) highlighted in red. Grey shading: 95% confidence interval determined by the ensemble spread. Orange: full uncertainty range adding the additional coverage uncertainty term. Global means have been calculated by averaging anomalies for northern and southern hemispheres for each ensemble member. Maps require six months of data within a year for a grid cell average to be plotted.

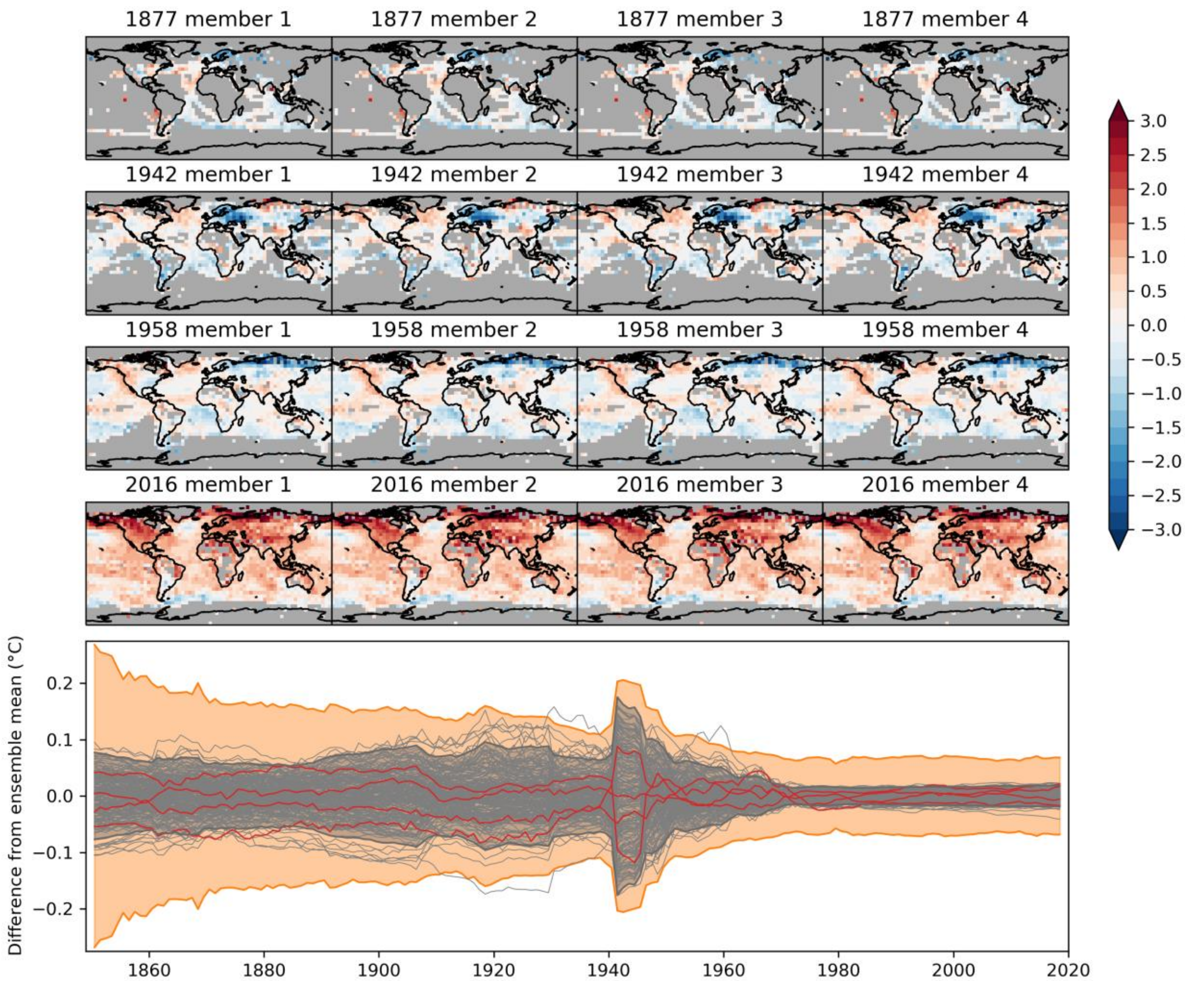


Figure 3. As Figure 2, but for the HadCRUT5 non-infilled dataset. Upper panel: annual average temperature anomaly (°C, relative to 1961-90) for 1877, 1942, 1958 and 2016 in four example ensemble members. Lower panel: ensemble spread in global mean (°C), 1850-2018. The difference between each ensemble member and the ensemble mean is shown by the grey lines, with the first four ensemble members (corresponding to the maps above) highlighted in red. Grey shading: 95% confidence interval determined by the non-infilled ensemble spread. Orange: full uncertainty range including additional measurement and sampling uncertainty terms, that are not sampled by the non-infilled ensemble, and the coverage uncertainty term. Global means have been calculated by averaging anomalies for northern and southern hemispheres for each ensemble member. Maps require six months of data within a year for a grid cell average to be plotted.

4.2 Global, hemispheric and regional series

Annual global and hemispheric average temperature anomaly series for HadCRUT5 are shown in Figure 4, along with the fraction of regional data coverage represented in the non-infilled dataset and the HadCRUT5 analysis.

Areal data coverage in the HadCRUT5 analysis grids first reaches 90% in the 1900s, with two subsequent drops in coverage in the late 1910s and early 1940s associated with the two world wars. Northern hemisphere coverage exceeds 99% in the early 1920s and reaches 100% in the mid-1950s. Uncertainty in southern hemisphere temperatures is greatest in the period prior to the establishment of a sustained Antarctic monitoring network in the 1950s (see also Figure 5 (a)), after which global coverage exceeds 97% in the 1960s. The spatial extent of the observing network in the southern hemisphere is also a prominent contribution to uncertainty in global average series prior to the 1950s. Global coverage of the analysis fields is typically not complete even in modern years due to an absence of sustained observation in the southern South Pacific, and the nearby Southern Ocean and Antarctic.

Southern Hemisphere anomalies are cooler in the HadCRUT5 analysis in the 1990s from around 1992, particularly in 30-60S (Figure 5 (b)). The observing network is less dense in these regions, with regular shipping covering only the equatorward half of the latitude band, leading to differences between non-infilled HadCRUT5 and the HadCRUT5 analysis. Variability in the regional time series (Figure 5) is smaller in the early record in the HadCRUT5 analysis than the non-infilled dataset, particularly in the high latitude regions as a result of reduced uncertainty from spatial sampling in the HadCRUT5 analysis.

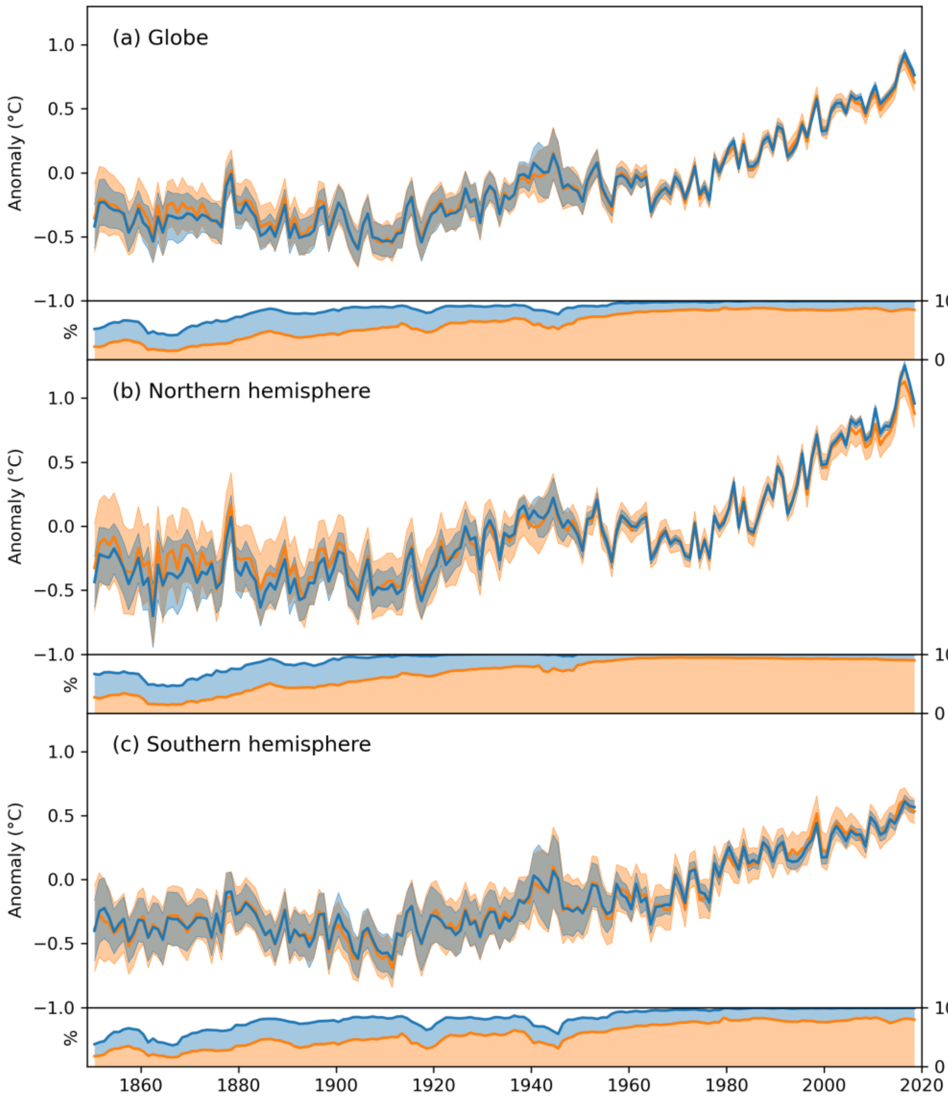


Figure 4. Comparison between the HadCRUT5 analysis and non-infilled data set. **(a)** Globe, **(b)** Northern Hemisphere and **(c)** Southern Hemisphere. Upper panel in each pair: annual average temperature anomaly (°C, relative to 1961-90), 1850-2018. Lower panel in each pair: percentage of area covered by data in each annual average. Orange: non-infilled HadCRUT5 data set. Blue: HadCRUT5 analysis. Solid lines: ensemble mean. Orange/blue shading: 95% confidence interval. Global means have been calculated by averaging anomalies for northern and southern hemispheres.

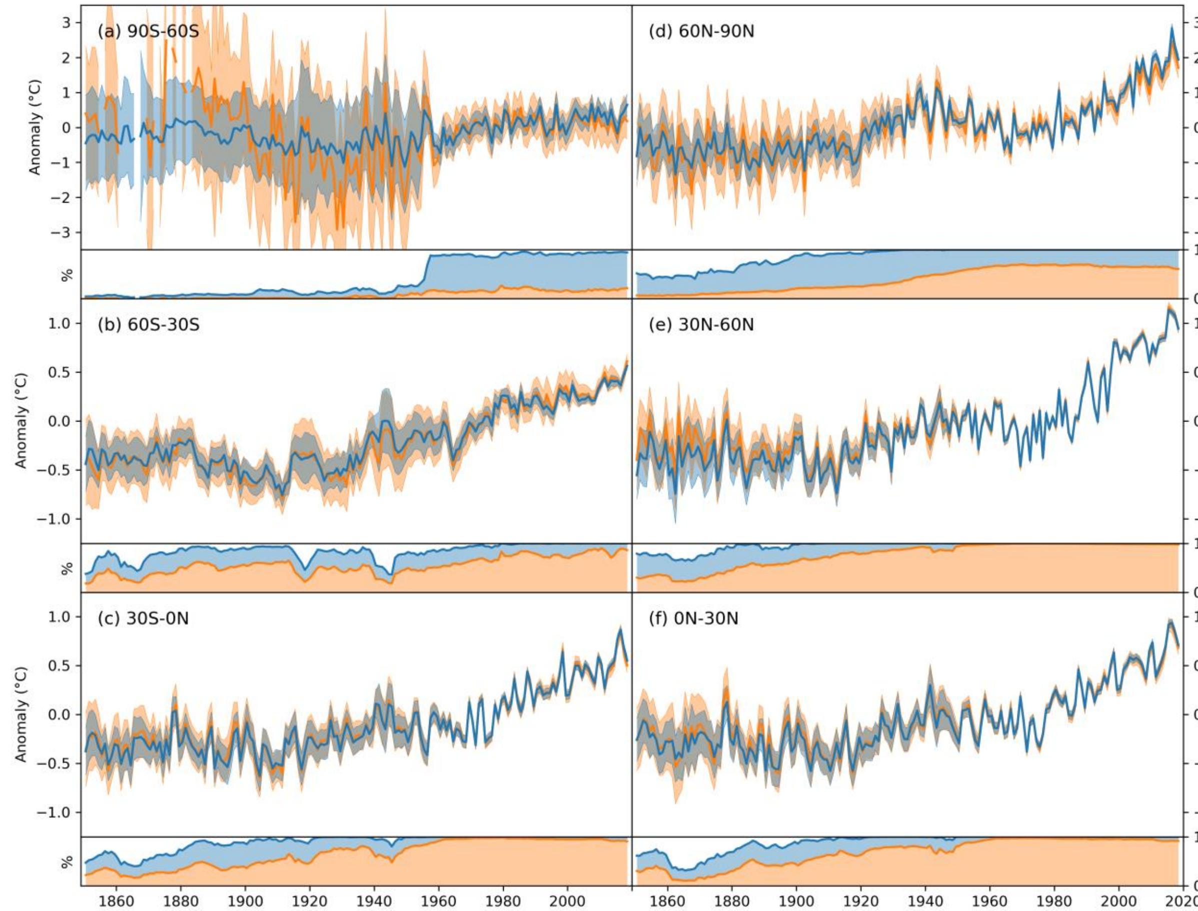


Figure 5. Comparison between the HadCRUT5 analysis and non-infilled data set. **(a)** 90°S-60°S, **(b)** 60°S-30°S, **(c)** 30°S-0°N, **(d)** 60°N-90°N, **(e)** 30°N-60°N and **(f)** 0°N-30°N. Upper panel in each pair: annual average temperature anomaly (°C, relative to 1961-90), 1850-2018. Lower panel in each pair: percentage of area covered by data in each annual average. Orange: non-infilled HadCRUT5 data set. Blue: HadCRUT5 analysis. Solid lines: ensemble mean. Orange/blue shading: 95% confidence interval.

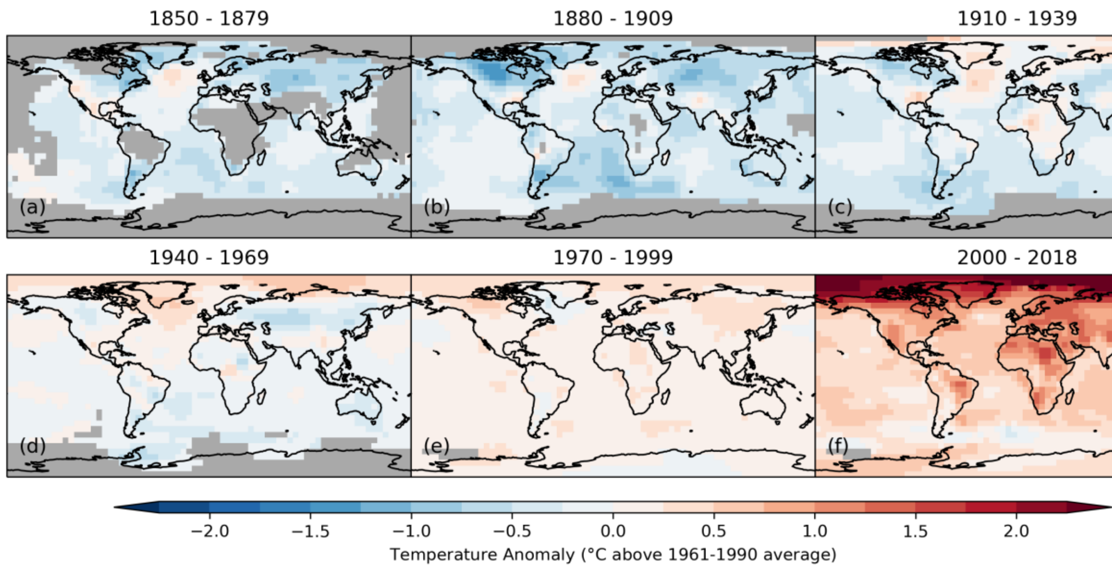


Figure 6. Long-term average temperature anomaly (°C, relative to 1961-90). **(a)** 1850-1879, **(b)** 1880-1909, **(c)** 1910-1939, **(d)** 1940-1969, **(e)** 1970-1999 and **(f)** 2000-2018. Averages require at least one month per quarter, three quarters per year, and 50% of years per multi-year period.

In regions where data are sparse, and hence uncertainty in surface temperature analyses is largest, data that might be used to validate the analyses is also highly limited. Here we have used the ratio of posterior to prior variances to remove regions with weak observational constraint (see Appendix for details). Despite restricting the reconstruction to regions that are locally constrained, there is a marked increase in the area of the globe represented by the HadCRUT5 analysis in comparison to the non-infilled data set (see coverage timeseries in Figure 5 and example monthly fields in Figures S10 to S13 of the Supporting Information).

Figure 6 reveals the patterns of change in successive 30-year periods and the most recent 19 years of the HadCRUT5 analysis. Even in these longer-term averages, there are regions that are particularly warm or cool relative to the global mean. The final panel for 2000-2018 illustrates the greater warming at high northern latitudes and over the land compared to the ocean. The surface waters of the Southern Ocean, in contrast, have warmed more slowly than many other areas. We also see one area of long-term cooling, to the south of Greenland and Iceland (Parker et al., 1994). 1880-1909 was a particularly cool period, with centers of low average anomalies in the South Atlantic, Canada and central Russia.

4.3 Comparisons with other analyses

Average temperature changes over the whole period of record in 30° latitude bands for a range of analyses are shown in Figure 7. These analyses include NOAAGlobalTemp v5 (Huang et al. 2019), NASA GISTEMP v4 (Hansen et al., 2010; Lenssen et al. 2019), the Cowtan & Way analysis (Cowtan & Way, 2014), and the Berkeley Earth analysis (Rohde & Hausfather, 2020). The HadCRUT.5.0.0.0 analysis is also shown.

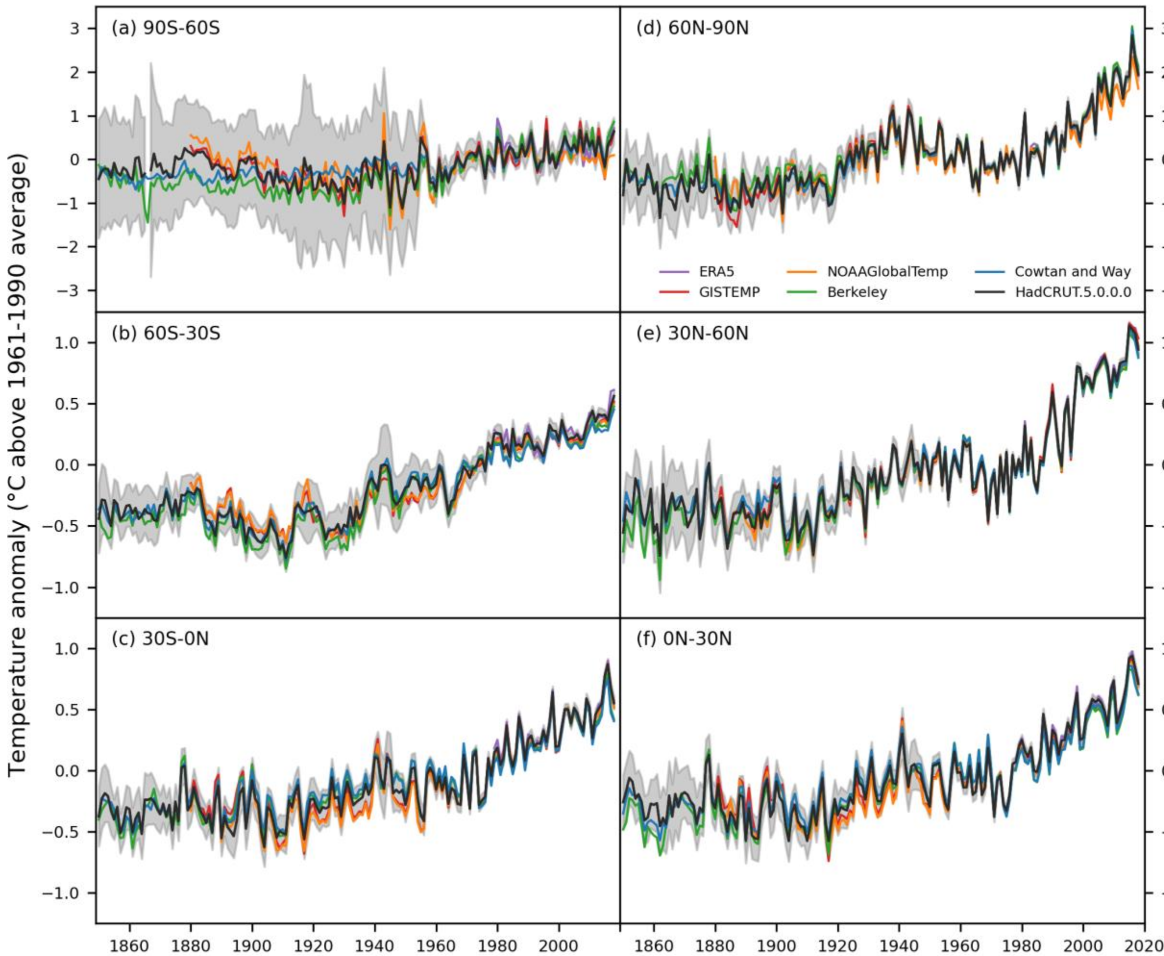


Figure 7. Comparison between long-term near-surface temperature data sets. Annual average temperature anomaly (°C, relative to 1961-90), 1850-2018. **(a)** 90°S-60°S, **(b)** 60°S-30°S, **(c)** 30°S-0°N, **(d)** 60°N-90°N, **(e)** 30°N-60°N and **(f)** 0°N-30°N. Black: HadCRUT5 analysis ensemble mean. Pink: ERA5. Red: GISTEMP. Orange: NOAAGlobalTemp. Green: Berkeley Earth. Blue: Cowtan & Way. Grey shading: 95% confidence interval on the HadCRUT.5.0.0.0 analysis determined by the ensemble spread and coverage uncertainty.

All of the analyses shown use spatial infilling. Cowtan & Way and Berkeley Earth use interpolation methods based on a statistical model of local covariance structure (although within a more complex statistical model of global temperature variation in the Berkeley Earth analysis). NOAAGlobalTemp uses a model of spatially-varying local patterns of temperature variability.

GISTEMP employs a distance-weighted interpolation for land based meteorological station data and uses the same large-scale analysis of sea-surface temperatures used in NOAAGlobalTemp. GISTEMP, Cowtan & Way and Berkeley are each close to globally complete since the 1950s while the NOAAGlobalTemp data set does not extend into data-sparse polar regions.

The analyses are most similar in regions with the densest observational coverage, such as in the northern mid-latitudes (Figure 7 (e)). Where observational coverage is lowest, the analyses become sensitive to assumptions underpinning reconstruction methods. For example, NOAAGlobalTemp reconstructs fields through low-frequency smoothing and a model of dominant spatial patterns of variability, while methods based on local covariance structure may tend toward a field mean in the case of Cowtan & Way, Berkeley, or the HadCRUT5 analysis ensemble mean, or towards the anomalies observed at nearby locations for the GISTEMP land analysis method. The analyses also differ in how regions that are distant from observed locations are included or are masked.

The HadCRUT5 analysis method is closely related to the method used in Cowtan & Way but differs in three key aspects. First, it accounts for the spatial variation in data uncertainty as well as the estimated measurement and sampling error covariances. This is particularly important for the oceans, where less-reliable ship data are combined with more accurate data from drifting and moored buoys. Second, the spatial analysis method is used to make improved temperature estimates at all locations, not just grid cells without data. Third, by using a full covariance model for both the temperature field and the observational uncertainty within a Bayesian analysis framework, it is possible to sample from the posterior of the distribution to generate a consistent ensemble data set that combines all known sources of uncertainty whilst respecting the estimated covariance structure of the temperature anomaly field.

The differences between the HadCRUT5 analysis ensemble mean and Cowtan & Way in the post 1950 period, are largely due to changes in the estimated SST biases. As Berkeley Earth shows similar differences and uses the same SST data set as Cowtan & Way, we can infer that changes in the estimated SST biases are the key difference here as well. The changes in SST bias estimates are larger in the more sparsely observed regions – the tropics and southern hemisphere – where there are fewer ships, so changes in assumptions about observing practice of a few countries can have a proportionately larger effect.

Differences between HadCRUT5 and the ERSST-based data sets, GISTEMP and NOAAGlobalTemp are also largely due to differences in estimated SST biases. In particular, ERSST tends to be cooler than HadSST4 from the early 20th century to the start of the Second World War and from the end of the war to around 1955; this difference is associated with uncertainty in the estimated biases associated with bucket measurements, particularly in the Southern Hemisphere and the tropics. From the 1960s, agreement between HadSST4 and ERSST is better, though there is a notable cooling of ERSST relative to HadSST4 in the early 1990s associated with a relative cooling of marine air temperature compared to SST (see Kennedy et al. 2019). From the late 1990s onwards, both ERSSTv5 and HadSST4 show good relative stability compared to instrumentally homogeneous data sets (Hausfather et al., 2017; Kennedy et al., 2019). Notable structural uncertainty remains in early SST records.

Differences can be seen in the first half of the 20th century between GISTEMP/NOAAGlobalTemp and Cowtan & Way/HadCRUT5 over the latitude band 0°N-30°S with GISTEMP/NOAAGlobalTemp cooler (Figure 7 (c)). Regional differences over land partly result from differences in homogenization and the underlying station data sets. HadCRUT5 uses homogenized station data (from CRUTEM5), as provided by national meteorological services or research projects. Other datasets include automated homogenization algorithms (Huang et al., 2019; Menne et al., 2018; Rohde et al., 2013b). This may result in regional differences between data sets, particularly where the measurement network is less dense and, as a consequence, there is greater uncertainty in homogenization.

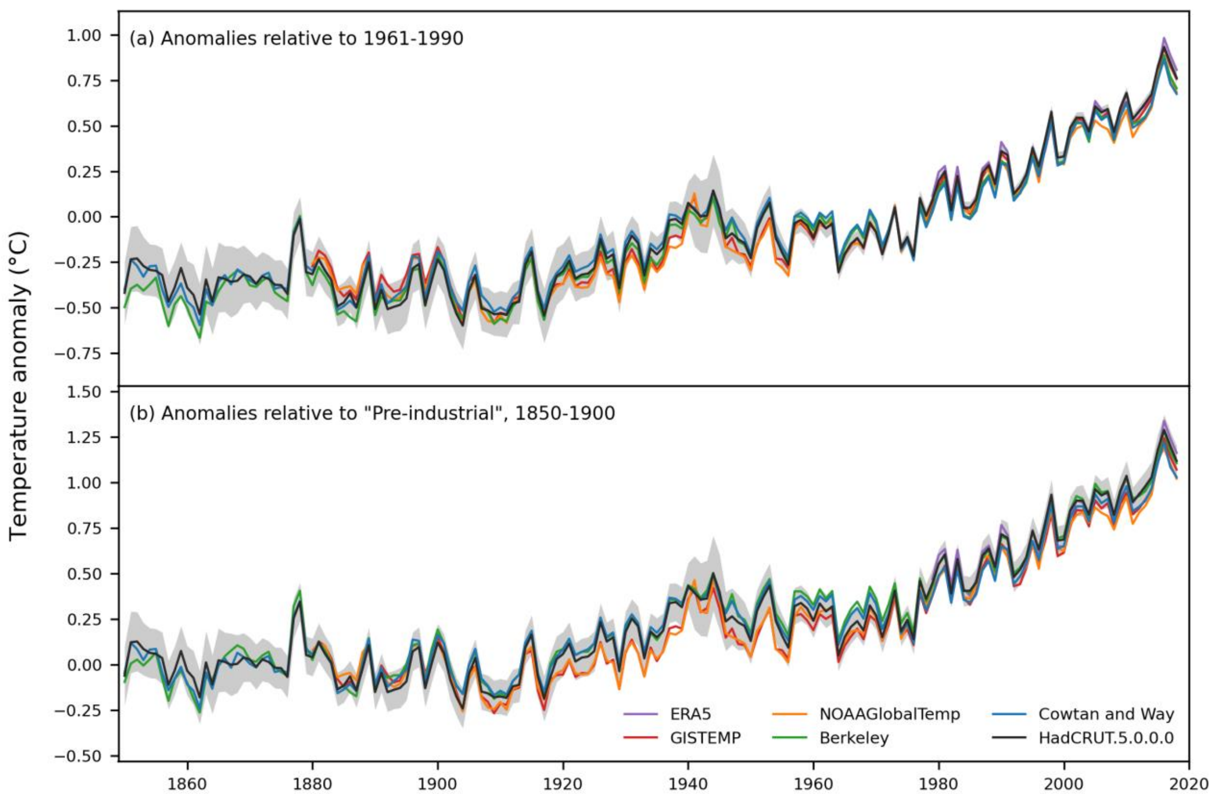


Figure 8. Comparison of annual global average temperature anomaly series (ºC) relative to two baselines: (a) 1961-1990 and (b) 1850-1900, taken as representative of pre-industrial conditions. Black: HadCRUT5 analysis ensemble mean. Pink: ERA5. Red: GISTEMP. Orange: NOAAGlobalTemp. Green: Berkeley Earth. Blue: Cowtan and Way. Grey shading: 95% confidence interval on the HadCRUT5 analysis determined by the ensemble spread only. Global means have been calculated for each data set by averaging anomalies for northern and southern hemispheres. For all datasets except for ERA5, anomaly series are computed by adjusting monthly time series to the appropriate baseline using data available in the anomaly reference period before averaging to annual series. ERA5 timeseries are shifted to match the 1981-2010 average for the HadCRUT5 analysis series, due to insufficient data in the climatology periods to

compute anomalies. Anomaly series and uncertainties provided by the dataset producers using each dataset's native methods are shown in Supporting Information Figure S9.

Temperature changes relative to the average over the late 19th century are shown in Figure 8. The 51-year period 1850-1900 is often considered for practical purposes to be representative of pre-industrial conditions. This approximation of pre-industrial temperatures is consistent with that adopted in IPCC AR5 (Hartmann et al., 2013) and IPCC SR1.5 (Allen et al., 2018), noting that any choice of period is a compromise, with natural variability and forcing playing a role (Hawkins et al., 2017). For analyses that do not extend back to 1850 (NOAAGlobalTemp and GISTEMP), 1880 to 1900 is used as the reference period here. By referencing the time series to this early period, the spread of temperature anomalies later in the series is increased. This increased spread reflects uncertainty in temperatures in the early reference period and not uncertainty in recent temperature changes. On the global mean, the analyses are remarkably consistent with one another despite the differences in their construction.

5 Conclusions

An updated data set of global near-surface temperature change, HadCRUT5, is presented. Updates in the CRUTEM5 dataset have expanded the underlying land station series and provided additional data quality checks. Updates in HadSST4 have brought improved understanding of the evolution of the marine observing network, contributing improved bias adjustments and uncertainty estimates. These are combined both in a non-infilled data set and in a new ensemble statistical analysis that provides a more spatially complete assessment of global and regional changes and uncertainty therein.

The new HadCRUT5 analysis ensemble samples a greater range of the quantified uncertainties than our previous assessment (Morice et al., 2012). Uncertainties arising from systematic errors associated with observational methods, measurement and sampling errors and spatial analysis uncertainty are all encoded into the expanded 200-member ensemble, communicating the major known sources of uncertainty in an easily accessible way.

Time series of globally averaged temperature anomalies show greater 21st century warming for the HadCRUT5 analysis than for the HadCRUT5 non-infilled data set. The increased warming is predominantly associated with improved representation of the rapidly warming but sparsely observed high latitudes of the northern hemisphere. This finding is consistent with other independently-produced statistical analyses of global temperature changes and is also consistent with temperature changes observed in reanalysis data sets that assimilate observational data into a numerical weather prediction model (Kobayashi et al., 2015; Gelaro et al., 2017; Blunden & Arndt, 2019; Hersbach et al., 2020).

The HadCRUT5 analysis indicates that globally averaged temperatures in the second half of the 19th century were on average cooler than estimates based on non-infilled HadCRUT5. This is also consistent with assessments based on other independently produced statistically infilled analyses. Combined with the evidence of increased warming in recent years, the infilled analyses indicate that warming since the 19th century is likely greater than is indicated by HadCRUT4 as a result both of observational sampling in the non-infilled data set and of updates to our understanding of biases in sea-surface temperature measurements resulting from changes in the make-up of the marine observing network.

There is, however, uncertainty in our understanding of 19th century temperatures resulting from limitations in observational sampling, particularly in the southern hemisphere, and uncertainty associated with residual observational biases. Uncertainty remains in the early instrumental record in locations for which observational data are not available to inform the analysis. This is most evident in the Antarctic, the Arctic and regions of the southern hemisphere land, prior to the establishment of permanent observing sites.

Methodological choices in representation of data sparse regions in different data sets lead to differences between global and regional average temperature time series. The impacts of these choices are most evident in regions and at times in which the observational data required to constrain the analysis is limited or unavailable, particularly in regions of the southern hemisphere in the early record. The spread of 19th century temperature analyses produced by different monitoring centers in part reflects the sensitivity to differences in methods used. These methods assume different statistical models for the data; therefore, the differences between analyses are not necessarily captured by the uncertainty estimates of any single method.

The updated analysis methods assist in mitigation of the impacts of low availability of observational data in data sparse regions. We anticipate that an extension, in potential future work, of the analysis covariance model to describe regional variation in variability would further improve the analysis temperature fields and uncertainty estimates. However, digitization of as yet unavailable observations and submission of these to open archives continues to be invaluable to improve regional data coverage and reduce uncertainty further.

The use of marine air temperature observations has recently been proposed to reconcile differences between datasets produced as a blend of SST and air temperature observations and model-based studies using near-surface air temperatures over ocean (Cowtan et al., 2015; Richardson et al., 2016). However, uncertainties in observed long-term changes in marine air temperature and their differences from observed SSTs are important to understand (Kennedy et al. 2019, Chan and Huybers 2019, Chan et al. 2019), and the marine air temperature observing network is less robust than that for SST and is in long-term decline (Berry & Kent, 2017). Challenges also remain in monitoring near-surface temperature changes in the cryosphere, given sparse observational coverage and changes in sea-ice extent, with impacts on downstream assessments (Richardson et al., 2018).

Relative biases in sea-surface temperature measurements arise from differences in measurement methods and instrumentation. Such biases change regionally and over time with gradual as well as abrupt changes in the composition of the observing network or underlying databases. The characteristics of different bias adjustment schemes can be seen in the differences between analyses, broadly grouping data sets into those (GISTEMP, Lenssen et al. (2019) and NOAAGlobalTemp, Huang et al. (2019)) that adopt the ERSST v5 dataset (Huang et al., 2017), those (Cowtan & Way (2014) and Berkeley Earth (Rohde & Hausfather, 2020)) that adopt HadSST3 (Kennedy et al., 2011a and b), and that which uses the improved HadSST4 data set (Kennedy et al., 2019), as is documented here. Differences between bias adjustments applied in each data set are smaller than the assessed adjustments themselves, which result in a net reduction in observed warming compared to unadjusted measurements (Kennedy et al., 2019). Nevertheless, differences in SST bias assessments feature prominently as a source of difference between studies and remain a key uncertainty in assessing long-term change (Kent et al., 2017).

Despite methodological differences, temperature series derived from different analyses are in good agreement, generally lying within the assessed uncertainty range of the HadCRUT5 analysis. Updates in HadCRUT5 bring our estimates of global and hemispheric series closer to those of other recent studies. Remaining differences between estimates are understood to predominantly arise from differences in spatial analysis methods applied and differences in how each analysis accounts for changes in marine observing methods.

Acknowledgments

CPM, JJK, NAR, JPW, EH, REK, RJHD and IRS were supported by the Met Office Hadley Centre Climate Programme funded by BEIS and Defra. TJO and PDJ were supported by UK NERC (grant number NE/N006348/1, SMURPHS). The CRUTEM5 dataset was developed as part of the UK National Centre for Atmospheric Science. We thank three anonymous reviewers for their constructive comments which improved the manuscript.

Data access

The gridded temperature anomalies, the global and hemispheric timeseries and their uncertainty intervals will be available from the Met Office website (https://www.metoffice.gov.uk/hadobs). HadCRUT5 data will be archived for long term preservation and reuse as part of the HadCRUT catalogue at CEDA https://catalogue.ceda.ac.uk/uuid/f7189fabb084452c9818ba41e59ccabd. The CEDA archive of the HadCRUT.5.0.0.0 data can be accessed from https://catalogue.ceda.ac.uk/uuid/b9698c5ecf754b1d981728c37d3a9f02.

ERA5 was obtained from the Copernicus Climate Change Service (C3S) (2017): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS), date of access: 28/11/2019, https://cds.climate.copernicus.eu.

HadISST.2.2.0.0 was accessed on 11/12/2019 from https://www.metoffice.gov.uk/hadobs/hadisst2/.

The HadSST.4.0.0.0 ensemble is available from https://www.metoffice.gov.uk/hadobs/hadsst4/.

CRUTEM5 data will be available from https://www.metoffice.gov.uk/hadobs and the CRUTEM collection at CEDA https://catalogue.ceda.ac.uk/uuid/eeabb5e1ff2140f48e76ea1ffda6bb48. The CEDA archive of the CRUTEM.5.0.0.0 data can be accessed from https://catalogue.ceda.ac.uk/uuid/901f576dacae4e049630ab879d6fb476.

HadCRUT.4.6.0.0 is available from https://www.metoffice.gov.uk/hadobs/hadcrut4/.

GISTEMP version 4 was accessed on 17/11/2019 at 15:45 GMT from https://data.giss.nasa.gov/gistemp/.

NOAAGlobalTemp version 5 was accessed on 15/10/2019 at 07:07 GMT from https://www.ncdc.noaa.gov/noaa-merged-land-ocean-global-surface-temperature-analysis-noaaglobaltemp-v5.

Berkeley Earth was accessed on 17/11/2019 at 16:25 GMT from https://berkeleyearth.org/data-new/.

Cowtan and Way was accessed on 14/10/2019 at 10:40 GMT from https://www-users.york.ac.uk/~kdc3/papers/coverage2013/series.html.

Appendix A: Details of spatial analysis methods

A.1 Modelling the temperature anomaly field as a Gaussian process

Here we describe the methods used to construct the HadCRUT5 analysis. The method described in this section follows the Gaussian process method with explicit basis functions, described in Rasmussen & Williams (2006). The methods for analysis hyperparameter estimation are described in Appendix A.2. Appendix A.3 describes application to the non-infilled land air temperature and sea surface temperature ensemble grids, including methods for sampling analysis uncertainties. Regional masking of the analyses is described in Appendix A.4.

For a monthly temperature anomaly field $\boldsymbol{g}$, we model a vector of gridded temperature anomaly observations $\boldsymbol{y}$ as an additive combination of the true grid cell temperature anomaly values at the observed grid cells, denoted $\boldsymbol{g}_{obs}$, and an observational error term $\boldsymbol{\varepsilon}$:

$$\boldsymbol{y} = \boldsymbol{g}_{obs} + \boldsymbol{\varepsilon} \tag{A1}$$

The temperature anomaly field is decomposed into a regression model for the field mean, described in terms of a matrix of basis functions $\boldsymbol{H}$ with coefficients $\boldsymbol{\beta}$, and a spatially correlated field $\boldsymbol{f}$. The observations are then modelled by this decomposition, notating the basis function and the spatial field values at the observed grid cells as $\boldsymbol{H}_{obs}$ and $\boldsymbol{f}_{obs}$:

$$\boldsymbol{y} = \boldsymbol{f}_{obs} + \boldsymbol{H}_{obs}^T\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{A2}$$

Similarly, we define $\boldsymbol{g}_*$ as the values true temperature anomaly values at a set of prediction grid cells, notating basis functions and the spatial random field values at the prediction grid cells as $\boldsymbol{H}_*$ and $\boldsymbol{f}_*$, so that $\boldsymbol{g}_* = \boldsymbol{f}_* + \boldsymbol{H}_*^T\boldsymbol{\beta}$. In this analysis, $\boldsymbol{H}$ is set as a vector of ones so that the regression model acts as an estimate of a constant field mean for the analyzed month.

The spatial field $\boldsymbol{f}$ is defined in terms of its covariance structure. This covariance structure is parameterized as a function of distance between locations, as is common in Gaussian process or kriging analyses. The covariance $k(s_m, s_n)$ in spatial field values between locations s_m and s_n is defined as:

$$k(s_m, s_n) = \mathrm{cov}(f(s_m), f(s_n)) \tag{A3}$$

which defines the elements of a covariance matrix $\boldsymbol{K}$, with elements $[\boldsymbol{K}]_{mn} = k(s_m, s_n)$. In this analysis, a Matérn covariance function is used to model the covariances $k(s_m, s_n)$. This covariance function is parameterized by a smoothing hyperparameter ν, a range hyperparameter r that controls the rate at which covariance decays with distance between locations, and an amplitude hyperparameter σ. We use a stationary covariance function, with fixed values of the model hyperparameters fitted independently for the land air temperature and sea-surface temperature analyses. Covariances are evaluated as a function of Euclidian distance, rather than great circle distance, to retain the flexibility of Matérn covariance functions for data on the surface of a spherical Earth, avoiding restrictions to the range of smoothing hyperparameter values ν for which Matérn covariances are valid (i.e. to produce positive-definite covariance matrices) when using great circle distances (Gneiting, 2013). For separation distances with sufficiently strong covariance to be physically important, the Euclidian distance is close to the great circle distance.

Values of the field at observed grid cells, $\boldsymbol{f}_{obs}$, are modelled as realizations from $\boldsymbol{f}_{obs} \sim N(\boldsymbol{0}, \boldsymbol{K}_{obs})$ while those at predictions locations, $\boldsymbol{f}_*$, are modelled as $\boldsymbol{f}_* \sim N(\boldsymbol{0}, \boldsymbol{K}_*)$. Cross covariances between observed grid cells and prediction grid cells (i.e. the full output grid) are defined as $\boldsymbol{K}_{cross}$. We define $\boldsymbol{K}_{\boldsymbol{y}}$ as the sum of the covariance $\boldsymbol{K}_{obs}$ and the observational error covariance $\boldsymbol{R}$:

$$\boldsymbol{K}_{\boldsymbol{y}} = \boldsymbol{K}_{obs} + \boldsymbol{R} \tag{A4}$$

The observational error covariance matrices are constructed from the error model terms of the non-infilled datasets. When the analysis method is applied to an ensemble member of the land air temperature ensemble (i.e. the observation vector $\boldsymbol{y}$ contains the grid cell values for an individual land ensemble member for one month), the observational error covariance $\boldsymbol{R}$ contains the additional uncorrelated within-grid-cell measurement and sampling error variances on the leading diagonal with zeros elsewhere. When applied to a sea-surface temperature ensemble member (i.e. $\boldsymbol{y}$ contains the grid cell values for an individual HadSST4 ensemble member), $\boldsymbol{R}$ is constructed from the HadSST4 per-platform uncertainties for the partially correlated error component, provided as full error covariances in HadSST4, with additional uncertainty from uncorrelated measurement and sampling error variances added onto the leading diagonal.

Estimation proceeds following Rasmussen & Williams (2006). The expected value of the anomaly field $\boldsymbol{g}_*$ given the observations $\boldsymbol{y}$ is defined as $\boldsymbol{\mu}_{\boldsymbol{g}_*|\boldsymbol{y}}$ where:

$$\boldsymbol{\mu}_{\boldsymbol{g}_*|\boldsymbol{y}} = \boldsymbol{K}_{cross}^{T}\boldsymbol{K}_{\boldsymbol{y}}^{-1}\boldsymbol{y} + \boldsymbol{F}^{T}\boldsymbol{\mu}_{\boldsymbol{\beta}|\boldsymbol{y}} \tag{A5}$$

and:

$$\boldsymbol{F}_* = \boldsymbol{H}_* - \boldsymbol{H}_{obs}\boldsymbol{K}_{\boldsymbol{y}}^{-1}\boldsymbol{K}_{cross} \tag{A6}$$

Here the terms involving the estimation of regression coefficients $\boldsymbol{\beta}$ (of which we need no prior knowledge) are:

$$\boldsymbol{\mu}_{\boldsymbol{\beta}|\boldsymbol{y}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}|\boldsymbol{y}}\boldsymbol{H}_{obs}\boldsymbol{K}_{\boldsymbol{y}}^{-1}\boldsymbol{y} \tag{A7}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}|\boldsymbol{y}} = \left(\boldsymbol{H}_{obs}\boldsymbol{K}_{\boldsymbol{y}}^{-1}\boldsymbol{H}_{obs}^{T}\right)^{-1} \tag{A8}$$

The posterior covariance $\boldsymbol{\Sigma}_{\boldsymbol{g}_*|\boldsymbol{y}}$ for the Gaussian process prediction is given by:

$$\boldsymbol{\Sigma}_{\boldsymbol{g}_*|\boldsymbol{y}} = \boldsymbol{K}_* - \boldsymbol{K}_{cross}^{T}\boldsymbol{K}_{\boldsymbol{y}}\boldsymbol{K}_{cross} + \boldsymbol{F}^{T}\boldsymbol{\Sigma}_{\boldsymbol{\beta}|\boldsymbol{y}}\boldsymbol{F} \tag{A9}$$

Together, $\boldsymbol{\mu}_{\boldsymbol{g}_*|\boldsymbol{y}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{g}_*|\boldsymbol{y}}$ define the full posterior distribution of the Gaussian process estimate of the gridded temperature anomaly field $\boldsymbol{g}_*$ for all output grid cells, given observations $\boldsymbol{y}$.

A.2 Kernel hyperparameter estimation

The estimation of the amplitude (σ) and decorrelation range (r) parameters of our spatial model is based on application of the maximum marginal likelihood method that is described in Rasmussen & Williams (2006). Here, the kernel hyperparameters $\boldsymbol{\theta} = (\sigma, r)$ are fit through numerical optimization to find the parameters that maximize the marginal log likelihood function, rearranged here as:

$$\log p(\boldsymbol{y}|\boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{K}_{y}^{-1}\boldsymbol{y} + \frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{\beta}|\boldsymbol{y}}^{\boldsymbol{T}}\boldsymbol{\Sigma}_{\boldsymbol{\beta}|\boldsymbol{y}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}|\boldsymbol{y}} - \frac{1}{2}\log\left|\boldsymbol{K}_{y}\right| + \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\boldsymbol{\beta}|\boldsymbol{y}}\right| - \frac{N-J}{2}\log(2\pi) \tag{A10}$$

Here, N is the number of observed grid cells in $\boldsymbol{y}$ and J is the number of covariates included in the regression portion of the analysis model. We include a single covariate for the analysis field mean, hence $J = 1$ in our application.

The hyperparameters are fit to monthly ‘best estimate’ gridded temperature anomaly fields separately for land air temperatures and sea-surface temperatures. Observational uncertainties are derived from the HadCRUT5 land ensemble uncertainty model (described in Morice et al., 2012) and HadSST4 uncertainty model (Kennedy et al., 2019), as described below.

As we fit hyperparameters to best estimates of the non-filled grids, we include an additional uncertainty component in the observational error covariance to represent the observational bias uncertainty that is encoded into the land ensemble and the HadSST4 ensemble. Hence, when fitting hyperparameters, an extended observational error covariance $\boldsymbol{R}'$ is substituted for $\boldsymbol{R}$ where $\boldsymbol{R}' = \boldsymbol{R} + \boldsymbol{\Sigma}_{ensemble}$ and $\boldsymbol{\Sigma}_{ensemble}$ is an error covariance matrix that is empirically derived from the ensemble. The ensemble-derived error covariance matrices are only used when fitting hyperparameters for the best estimate fields. They are not included in the observational error covariance term when fitting the analysis fields for individual ensemble members in Appendix A.3.

For land hyperparameter estimation, the monthly observation vector $\boldsymbol{y}$ is constructed from a CRUTEM5 best estimate field. The observational error covariance $\boldsymbol{R}$ is constructed from the uncorrelated measurement and sampling uncertainty grids, from the Brohan et al. (2006) error model, while $\boldsymbol{\Sigma}_{ensemble}$ is computed from the HadCRUT5 land ensemble. For marine hyperparameter estimation, the observation vector $\boldsymbol{y}$ is constructed from a HadSST4 ensemble median field. The observational error covariance matrices $\boldsymbol{R}$ are constructed by combining HadSST4 uncorrelated measurement and sampling uncertainties with the HadSST4 ‘micro bias’ error covariance matrices and $\boldsymbol{\Sigma}_{ensemble}$ is computed from the HadSST4 ensemble.

Hyperparameter estimates are computed for each of the 360 monthly fields in the 1961 to 1990 climatology period, during which the observational sampling is near global in extent. The hyperparameters used in the analysis are taken as the average of the hyperparameters fitted in the 360 monthly optimizations, with scale parameters rounded to the nearest 0.05 ˚C and range parameters rounded to the nearest 50 km. The resulting amplitude parameter σ and range parameter r for the land air temperature analysis are $\sigma = 1.2$˚C and $r = 1300$ km. For the sea surface temperature analysis, the fitted parameters are $\sigma = 0.6$˚C and $r = 1300$ km. The smoothing parameter was fixed at $\nu = 1.5$. This model represents typical land and marine temperature anomaly variability. The model does not include regional and seasonal variations in these parameters, nonetheless where there is a sufficient observational constraint the method can reproduce appropriate regional and seasonal variability in the analysis anomaly fields. Additional information on the monthly hyperparameter fits can be found in the Supporting Information.

A.3 Ensemble analysis

The HadCRUT5 ensemble land and marine analyses are constructed by applying Gaussian process regression to each ensemble member of the non-infilled land and marine data sets. Uncertainty is further explored by encoding analysis uncertainty into the ensemble, sampling from the Gaussian process posterior distribution through a process called conditional simulation (Chilès & Delfiner, 2012).

We denote a vector of observed grid cell temperature anomalies for a non-infilled ensemble member as $\boldsymbol{y}_d$, with the subscript d indexing the ensemble member. We then apply the Gaussian process analysis method to compute the expected value of the temperature anomaly field $\boldsymbol{\mu}_{\boldsymbol{g}_{*d}|\boldsymbol{y}_d}$ for the ensemble member, substituting $\boldsymbol{y}_d$ and $\boldsymbol{\mu}_{\boldsymbol{g}_{*d}|\boldsymbol{y}_d}$ into Equation A5. We then proceed to sample the analysis uncertainty through conditional simulation, as described below.

For each ensemble member, we draw a random sample from the joint prior distribution of the anomaly field at observed and prediction locations, setting the regression coefficient for each sample to an arbitrary value of $\boldsymbol{\beta}' = \boldsymbol{0}$. This sampling distribution is defined as:

$$\begin{bmatrix} \boldsymbol{g}'_{obs} \\ \boldsymbol{g}'_{*} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{H}_{obs}^T \boldsymbol{0} \\ \boldsymbol{H}_{*}^T \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}_{obs} & \boldsymbol{K}_{cross}^T \\ \boldsymbol{K}_{cross} & \boldsymbol{K}_{*} \end{bmatrix} \right) \qquad \text{(A11)}$$

This provides samples of the anomaly field, according to the Gaussian process model on the full output grid, drawn as $\boldsymbol{g}'_* = \boldsymbol{f}'_* + \boldsymbol{H}_*^T\boldsymbol{0}$, and at the observed locations $\boldsymbol{g}'_{obs} = \boldsymbol{f}'_{obs} + \boldsymbol{H}_{obs}^T\boldsymbol{0}$, with the correct covariance structure between observed and output grid locations.

We then generate pseudo-observations $\boldsymbol{y}'$ of the simulated temperature field by sampling from the observational error model $\boldsymbol{\varepsilon}' \sim N(\boldsymbol{0}, \boldsymbol{R})$. The simulated observation is then defined as:

$$\boldsymbol{y}' = \boldsymbol{f}'_{obs} + \boldsymbol{H}_{obs}^T\boldsymbol{0} + \boldsymbol{\varepsilon}' \qquad \text{(A12)}$$

Simulations of reconstruction error are based on application of the Gaussian process estimation to the simulated anomaly fields and simulated (pseudo) observations. The difference between the simulated field sample $\boldsymbol{g}'_*$ and the estimate based on the simulated pseudo observations $\boldsymbol{\mu}_{\boldsymbol{g}'_*|\boldsymbol{y}'}$ is a sample of the reconstruction error according to the Gaussian process model. This difference,

$\boldsymbol{e}' = \boldsymbol{\mu}_{\boldsymbol{g}'_*|\boldsymbol{y}'} - \boldsymbol{g}'_*$, is a sample from the posterior distribution of the Gaussian process regression, i.e. $\boldsymbol{e}' \sim N(\boldsymbol{\mu}_{\boldsymbol{g}'_*|\boldsymbol{y}'}, \boldsymbol{\Sigma}_{\boldsymbol{g}'_*|\boldsymbol{y}'})$.

For an ensemble member indexed by d with observation vector $\boldsymbol{y}_d$, the analysis values $\boldsymbol{g}_{*d}$ are computed as the sum of the Gaussian process estimate $\boldsymbol{\mu}_{\boldsymbol{g}_{*d}|\boldsymbol{y}_d}$, based on the real observations $\boldsymbol{y}_d$, and a simulated reconstruction error sample $\boldsymbol{e}'_d$:

$$\boldsymbol{g}_{*d} = \boldsymbol{\mu}_{\boldsymbol{g}_{*d}|\boldsymbol{y}_d} + \boldsymbol{e}'_d \qquad \text{(A13)}$$

The resulting ensemble encodes both the bias terms in the underlying observational ensemble and the reconstruction error for the Gaussian process.

The applied Gaussian process estimation is purely spatial and so does not provide information on temporally-correlated reconstruction error. To mitigate this, we modify the above sampling method to encode temporal correlation into the conditional simulation process. The simulated spatial fields $\boldsymbol{g}'_*$ and $\boldsymbol{g}'_{obs}$ are sampled such that they are fully correlated throughout a year, i.e. the same spatial field is used for each sample within a year. This provides a conservative upper bound on uncertainty in annual averages derived from the ensemble.

Known temporal correlations in observational measurement and sampling errors, which are not represented in the non-infilled land and marine ensembles, are similarly encoded into the observational error samples $\boldsymbol{\varepsilon}'_d$ when generating pseudo-observations. This strategy is applied for the residual SST micro biases that are represented in the HadSST4 observational error covariance matrices. These are encoded using the same random draw for all months in a year when sampling. This allows uncertainty in annual averages to be computed under a conservative assumption of full temporal correlation of SST micro biases within a year, as defined by the HadSST4 uncertainty model (Kennedy et al., 2019). Other measurement and sampling uncertainties, associated with temporally uncorrelated errors, are sampled independently for each month. No additional temporal correlation is encoded into the ensemble for land air temperatures as there is no temporal correlation in the measurement and sampling error terms for CRUTEM5 (although the analyzed land ensemble does already sample time correlated observational errors from residual station biases, which are distinct from the measurement and sampling uncertainty terms discussed here).

Although knowledge of temporal correlation in errors is not used to improve the estimated anomaly fields, the result of the sampling process is to enable an upper bound on uncertainty in annual averages to be obtained directly from the ensemble.

A.4 *Observational constraint mask*

Despite the application of spatial reconstruction, there are regions of the world in which the available observational coverage, particularly in the early part of the record, is such that a reliable reconstruction is not possible. In regions where local observations are not available, the analysis ensemble mean reverts towards the regression model estimate of the mean temperature anomaly, inferred from observed regions, while the ensemble spread tends towards that described by the Gaussian process prior distribution.

Consequently, regions where the constraint from local observations is poor are removed from the analysis. The reconstruction in these regions is highly sensitive to the prior covariance model and the estimated regression term $\boldsymbol{H}_*^T\boldsymbol{\mu}_{\boldsymbol{\beta}|\boldsymbol{y}}$, for which the coefficient estimate may be biased towards observed regions. This has been found to be the case in test analyses of climate model simulations in which global average temperature estimates have been found to be biased towards northern hemisphere temperatures during periods with sparse southern hemisphere coverage.

The criteria used to mask regions, defined in terms of a threshold α, is based on the ratio of posterior and prior variance of the local Gaussian process estimate, omitting the global regression term which has an improper prior, with regions of the analysis masked where the following inequality is satisfied:

$$\boldsymbol{1} - \frac{diag\left(\boldsymbol{K}_* - \boldsymbol{K}_{cross}^T\boldsymbol{K}_{\boldsymbol{y}}\boldsymbol{K}_{cross}\right)}{diag(\boldsymbol{K}_*)} < \alpha \qquad \text{(A14)}$$

The left-hand side of Equation A14 is bounded between zero and one and we use a threshold of $\alpha = 0.25$ to provide a balance between retaining regions with useful information content and masking those regions that have a weak observational constraint. Global and hemispheric average temperature series for varying α are provided in the Supporting Information and indicate that these diagnostics are insensitive to the choice of α values in the range 0.1 to 0.5.

References

Allen, M. R., O. P. Dube, W. Solecki, F. Aragón-Durand, W. Cramer, S. Humphreys, M. Kainuma, J. Kala, N. Mahowald, Y. Mulugetta, R. Perez, M. Wairiu, and K. Zickfeld (2018), Framing and Context. In: Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty [Masson-Delmotte, V., P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield (eds.)].

Atkinson, C. P., N. A. Rayner, J. J. Kennedy and S. A. Good (2014), An integrated database of ocean temperature and salinity observations, Journal of Geophysical Research: Oceans, 119, 7139– 7163, doi:10.1002/2014JC010053

Benestad, R. E., H. B. Erlandsen, A. Mezghani and K. M. Parding (2019), Geographical distribution of thermometers gives the appearance of lower historical global warming, Geophysical Research Letters. 46. https://doi.org/10.1029/2019GL083474

Berry, D. I., E. C. Kent, and P. K. Taylor (2004), An Analytical Model of Heating Errors in Marine Air Temperatures from Ships, J. Atmos. Oceanic Technol., 21, 1198–1215, doi:10.1175/1520-0426(2004)021<1198:AAMOHE>2.0.CO;2

Berry, D. I. and Kent, E. C. (2017), Assessing the health of the in situ global surface marine climate observing system. Int. J. Climatol., 37: 2248-2259. doi:10.1002/joc.4914

Blunden, J. and D. S. Arndt, Eds. (2019), State of the Climate in 2018. Bull. Amer. Meteor. Soc., 100 (9), Si–S305, doi:10.1175/2019BAMSStateoftheClimate.1.

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 dataset from 1850, J. Geophys. Res, 111, D12106, doi:10.1029/2005JD006548.

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

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

Chan, D. and P. Huybers (2019), Systematic Differences in Bucket Sea Surface Temperature Measurements among Nations Identified Using a Linear-Mixed-Effect Method, J. Climate, 32, 2569–2589, doi:10.1175/JCLI-D-18-0562.1

Chilès, J-P and P. Delfiner (2012). Geostatistics: Modeling Spatial Uncertainty. Wiley Series In Probability and Statistics. 10.1002/9781118136188.

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

Cowtan, K. and R. G. Way (2014), Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends, Q.J.R. Meteorol. Soc., 140, 1935–1944. doi:10.1002/qj.2297

Cowtan, K., Z. Hausfather, E. Hawkins, P. Jacobs, M. E. Mann, S. K. Miller, B. A. Steinman, M. B. Stolpe and R. G. Way (2015), Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures, Geophys. Res. Lett., 42, 6526– 6534, doi:10.1002/2015GL064888.

Cowtan, K., R. Rohde, Z. Hausfather (2018), Evaluating biases in sea surface temperature records using coastal weather stations, Q J R Meteorol Soc., 144, 670–681, doi:10.1002/qj.3235

Donlon, C. J., M. Martin, J. Stark, J. Roberts-Jones, E. Fiedler and W. Wimmer (2012), The operational sea surface temperature and sea ice analysis (OSTIA) system, Remote Sensing of Environment, 116, 140-158, doi:10.1016/j.rse.2010.10.017

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

Folland, C. K., N. A. Rayner, S. J. Brown, T. M. Smith, S. S. P. Shen, D. E. Parker, I. Macadam, P. D. Jones, R. N. Jones, N. Nichols and D. M. H. Sexton (2001), Global temperature change and its uncertainties since 1861, Geophysical Research Letters, 28(13), 2621-2624

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

Gelaro, R., W. McCarty, M.J. Suárez, R. Todling, A. Molod, L. Takacs, C.A. Randles, A. Darmenov, M. G. Bosilovich, R. Reichle, K. Wargan, L. Coy, R. Cullather, C. Draper, S. Akella, V. Buchard, A. Conaty, A.M. da Silva, W. Gu, G. Kim, R. Koster, R. Lucchesi, D. Merkova, J. E. Nielsen, G. Partyka, S. Pawson, W. Putman, M. Rienecker, S. D. Schubert, M. Sienkiewicz, and B. Zhao, 2017: The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). J. Climate, 30, 5419–5454, https://doi.org/10.1175/JCLI-D-16-0758.1

Gneiting, T. (2013). Strictly and non-strictly positive definite functions on spheres, Bernoulli, 19, 1327–1349, doi:10.3150/12-BEJSP06

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

Hartmann, D. L., A. M. G. Klein Tank, M. Rusticucci, L.V. Alexander, S. Brönnimann, Y. Charabi, F. J. Dentener, E. J. Dlugokencky, D. R. Easterling, A. Kaplan, B. J. Soden, P. W. Thorne, M. Wild and P. M. Zhai (2013), Observations: Atmosphere and Surface. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

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

Hawkins, E. et al., (2017), Estimating changes in global temperature since the pre-industrial period. Bulletin of the American Meteorological Society, BAMS–D–16–0007.1, doi:10.1175/bams-d-16-0007.1.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S. and Thépaut, J.-N. (2020), The ERA5 Global Reanalysis. Q J R Meteorol

Soc. Accepted Author Manuscript. doi:10.1002/qj.3803 Hirahara, S., M. Ishii, and Y. Fukuda (2014), Centennial-scale sea surface temperature analysis and its uncertainty, J. Climate, 27, 57–75, doi:10.1175/JCLI -D-12-00837.1,

Huang, B., P. W. Thorne, T. M. Smith, W. Liu, J. H. Lawrimore, V. F. Banzon, H. Zhang, T. C. Peterson, M. J. Menne (2016), Further exploring and quantifying uncertainties for Extended Reconstructed Sea Surface Temperature (ERSST) version 4 (v4). J. Climate, doi:10.1175/JCLI-D-15-0430.1.

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

Huang, B., M. J. Menne, T. Boyer, E. Freeman, B. E. Gleason, J. H. Lawrimore, C. Liu, J. J. Rennie, C. J. Schreck, F. Sun, R. Vose, C. N. Williams, X. Yin, and H. Zhang (2019), Uncertainty estimates for sea surface temperature and land surface air temperature in NOAAGlobalTemp version 5, J. Climate, 0, doi:10.1175/JCLI-D-19-0395.1

Ilyas, M., C. M. Brierley and S. Guillas (2017), Uncertainty in regional temperatures inferred from sparse global observations: Application to a probabilistic classification of El Niño, Geophys. Res. Lett., 44, 9068– 9074, doi:10.1002/2017GL074596.

Jones, P. D., D. H. Lister, T. J. Osborn, C. Harpham, M. Salmon, and C. P. Morice (2012), Hemispheric and large-scale land surface air temperature variations: An extensive revision and an update to 2010, J. Geophys. Res., 117, D05127, doi:10.1029/2011JD017139

Jones, P. D., T. J. Osborn, and K. R. Briffa (1997), Estimating Sampling Errors in Large-Scale Temperature Averages. J. Climate, 10, 2548–2568, https://doi.org/10.1175/1520-0442(1997)010<2548:ESEILS>2.0.CO;2

Junod, R. A. and J. R. Christy (2020), A new compilation of globally gridded night-time marine air temperatures: The UAHNMATv1 dataset. Int J Climatol, 40, 2609– 2623. https://doi.org/10.1002/joc.6354

Kadow, C., Hall, D. M. & Ulbrich, U. (2020) Artificial intelligence reconstructs missing climate information. Nat. Geosci., https://doi.org/10.1038/s41561-020-0582-5

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K.C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph (1996), The NCEP/NCAR 40-Year Reanalysis Project. Bull. Amer. Meteor. Soc., 77, 437–472, https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2

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

Kaplan, A., Y. Kushnir, M. A. Cane, and M. B. Blumenthal (1997), Reduced space optimal analysis for historical data sets: 136 years of Atlantic sea surface temperatures, J. Geophys. Res., 102(C13), 27835– 27860, doi:10.1029/97JC01734.

Kennedy J. J., N. A. Rayner, R. O. Smith, M. Saunby and D. E. Parker (2011a), Reassessing biases and other uncertainties in sea-surface temperature observations since 1850 part 1: measurement and sampling errors, J. Geophys. Res., 116, D14103, doi:10.1029/2010JD015218

Kennedy J. J., N. A. Rayner, R. O. Smith, M. Saunby and D. E. Parker (2011b), Reassessing biases and other uncertainties in sea-surface temperature observations since 1850 part 2: biases and homogenisation, J. Geophys. Res., 116, D14104, doi:10.1029/2010JD015220

Kennedy, J. J., N. A. Rayner, C. P. Atkinson, and R. E. Killick (2019), An ensemble data set of sea-surface temperature change from 1850: the Met Office Hadley Centre HadSST.4.0.0.0 data set, Journal of Geophysical Research: Atmospheres, 124, doi:10.1029/2018JD029867

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

Kent, E. C., J. J. Kennedy, T. M. Smith, S. Hirahara, B. Huang, A. Kaplan, D. E. Parker, C. P. Atkinson, D. I. Berry, G. Carella, Y. Fukuda, M. Ishii, P. D. Jones, F. Lindgren, C. J. Merchant, S. Morak-Bozzo, N. A. Rayner, V. Venema, S. Yasui, and H. Zhang (2017), A Call for New Approaches to Quantifying Biases in Observations of Sea Surface Temperature, Bull. Amer. Meteor. Soc., 98, 1601–1616, doi:10.1175/BAMS-D-15-00251.1.

Klein Tank, A. M. G., J. B. Wijngaard, G.P.Können, R. Böhm, G. Demarée, A. Gocheva, M. Mileta, S. Pashiardis, L. Hejkrlik, C. Kern-Hansen, R. Heino, P. Bessemoulin, G. Müller-Westermeier, M. Tzanakou, S. Szalai, T. Pálsdóttir, D. Fitzgerald, S. Rubin, M. Capaldo, M., Maugeri, A. Leitass, A. Bukantis, R. Aberfeld, A. F. V. van Engelen, E. Forland, M. Mietus, F. Coelho, C. Mares, V. Razuvaev, E. Nieplova, T. Cegnar, J. Antonio López, B. Dahlström, A. Moberg, W. Kirchhofer, A. Ceylan, O. Pachaliuk, L. V. Alexander and P. Petrovic (2002), Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. Int. J. Climatol., 22: 1441-1453. doi:10.1002/joc.773

Kobayashi, S., et al. (2015), The JRA-55 Reanalysis: General Specifications and Basic Characteristics, Journal of the Meteorological Society of Japan. Ser. II, 93(1), 5-48, doi:10.2151/jmsj.2015-001.

Lenssen, N., G. Schmidt, J. Hansen, M. Menne, A. Persin, R. Ruedy, and D. Zyss (2019), Improvements in the GISTEMP uncertainty model, J. Geophys. Res. Atmos., 124, 12, 6307-6326, doi:10.1029/2018JD029522.

Menne, M. J., C. N. Williams, B. E. Gleason, J. J. Rennie, and J. H. Lawrimore (2018), The Global Historical Climatology Network Monthly Temperature Dataset, Version 4, J. Climate, 31, 9835–9854, doi:10.1175/JCLI-D-18-0094.1.

Osborn, T. J., P. D. Jones, D. H. Lister, C. P. Morice, I. R. Simpson and I. C. Harris (2020), Land surface air temperature variations across the globe updated to 2019: the CRUTEM5 dataset, Submitted to J. Geophys. Res.

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

Parker, D.E. (1994), Effects of changing exposure of thermometers at land stations. Int. J. Climatol., 14: 1-31. doi:10.1002/joc.3370140102

Parker, D. E., Jones, P. D., Folland, C. K., and Bevan, A. (1994), Interdecadal changes of surface temperature since the late nineteenth century, J. Geophys. Res., 99 (D7), 14373– 14399, doi:10.1029/94JD00548

Parker, D.E. (2006), A Demonstration That Large-Scale Warming Is Not Urban. J. Climate, 19, 2882–2895, https://doi.org/10.1175/JCLI3730.1

Parker, D. E. (2010), Urban heat island effects on estimates of observed climate change, Wiley Interdisciplinary Reviews: Climate Change, 1(1), 123-133

Rasmussen, C. E. and C. K. I. Williams (2006), Gaussian Processes for Machine Learning, the MIT Press, 2006, ISBN 026218253X.

Rayner, N. A., P. Brohan, D. E. Parker, C. K. Folland, J. J. Kennedy, M. Vanicek, T. J. Ansell and S. F. B. Tett (2006), Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: the HadSST2 data set, Journal of Climate, 19, 3, 446– 469, doi:10.1175/JCLI3637.1.

Rayner, N. A., R. Auchmann, J. Bessembinder, S. Brönnimann, Y. Brugnara, F. Capponi, L. Carrea, E. M. A. Dodd, D. Ghent, E. Good, J. J. Kennedy, E. C. Kent, R. E. Killick, P. van der Linden, F. Lindgren, K. S. Madsen, C. J. Merchant, H. R. Mitchelson, C. P. Morice, P. Nielsen-Englyst, P. F. Ortiz, J. J. Remedios, G. van der Schrier, A. A. Squintu, A. Stephens, P. W. Thorne, R. T. Tonboe, T. Trent, K. L. Veal, A. M. Waterfall, K. Winfield, J. P. Winn, R. I. Woolway (2020), The EUSTACE project: delivering global daily information on surface air temperature, ,Bull. Amer. Met. Soc, https://doi.org/10.1175/BAMS-D-19-0095.1.

Rennie, J. J., J. H. Lawrimore, B. E. Gleason, P. W. Thorne, C. P. Morice, M. J. Menne, C. N. Williams, .N., W. G. de Almeida, J. Christy, M. Flannery, M. Ishihara, K. Kamiguchi, A. M. G. Klein-Tank, A. Mhanda, D. H. Lister, V. Razuvaev, M. Renom, M. Rusticucci, J. Tandy, S. J. Worley, V. Venema, W. Angel, M. Brunet, B. Dattore, H. Diamond, M. A. Lazzara, F. Le Blancq, J. Luterbacher, H. Mächel, J. Revadekar, R. S. Vose, and X. Yin (2014), The international surface temperature initiative global land surface databank: monthly temperature data release description and methods. Geosci. Data J., 1: 75-102. doi:10.1002/gdj3.8

Reynolds, R. W. and T. M. Smith (1994), Improved Global Sea Surface Temperature Analyses Using Optimum Interpolation. J. Climate, 7, 929–948, https://doi.org/10.1175/1520-0442(1994)007<0929:IGSSTA>2.0.CO;2

Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang (2002), An improved in situ and satellite SST analysis for climate. J. Climate, 15, 1609-1625.

Richardson, M., Cowtan, K., Hawkins, E., & Stolpe, M. B. (2016), Reconciled climate response estimates from climate models and the energy budget of Earth. Nature Climate Change, 6(10), 931-935. https://doi.org/10.1038/nclimate3066

Richardson, M., K. Cowtan, and R.J. Millar (2018), Global temperature definition affects achievement of long-term climate goals. Environmental Research Letters, 13(5), 054004, doi:10.1088/1748-9326/aab305.

Rohde, R., R. A. Muller, R. Jacobsen, E. Muller, S. Perlmutter, A. Rosenfeld, J. Wurtele, D. Groom and C. Wickham (2013a), A New Estimate of the Average Earth Surface Land Temperature Spanning 1753 to 2011, Geoinfor Geostat: An Overview, 1:1, doi:10.4172/gigs.1000101.

Rohde R., R. M. Muller, R. Jacobsen, S. Perlmutter, A. Rosenfeld, J. Wurtele, J. Curry, C. Wickham and S. Mosher (2013b), Berkeley Earth Temperature Averaging Process, Geoinfor Geostat: An Overview, 1:2, doi:10.4172/gigs.1000103.

Rohde, R. A. and Hausfather, Z. (2020), The Berkeley Earth Land/Ocean Temperature Record, Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2019-259, in review

Smith, T. M., R. W. Reynolds, T. C. Peterson, and J. Lawrimore (2008), Improvements to NOAA's historical merged land–ocean surface temperatures analysis (1880–2006), Journal of Climate, 21, 2283–2296, doi:10.1175/2007JCLI2100.1.

Thorne, P. W., J. R. Lanzante, T. C. Peterson, D. J. Seidel and K. P. Shine (2011), Tropospheric temperature trends: history of an ongoing controversy. WIREs Clim Change, 2: 66-88. doi:10.1002/wcc.80

Titchner, H. A., and N. A. Rayner (2014), The Met Office Hadley Centre sea ice and sea surface temperature data set, version 2: 1. Sea ice concentrations, J. Geophys. Res. Atmos., 119, 2864-2889, doi: 10.1002/2013JD020316.

Yun, X., Huang, B., Cheng, J., Xu, W., Qiao, S., and Li, Q. (2019), A new merge of global surface temperature datasets since the start of the 20th century, Earth Syst. Sci. Data, 11, 1629–1643, https://doi.org/10.5194/essd-11-1629-2019.

Zhang, H-M, J. H. Lawrimore, B. Huang, M. J. Menne, X. Yin, A. Sȧnchez-Lugo, B. E. Gleason, R. Vose, D. Arndt, J. J. Rennie, and C. N. Williams. (2019), Updated Temperature Data Give a Sharper View of Climate Trends. *Eos,* 100, https://doi.org/10.1029/2019EO128229

Figure 1.

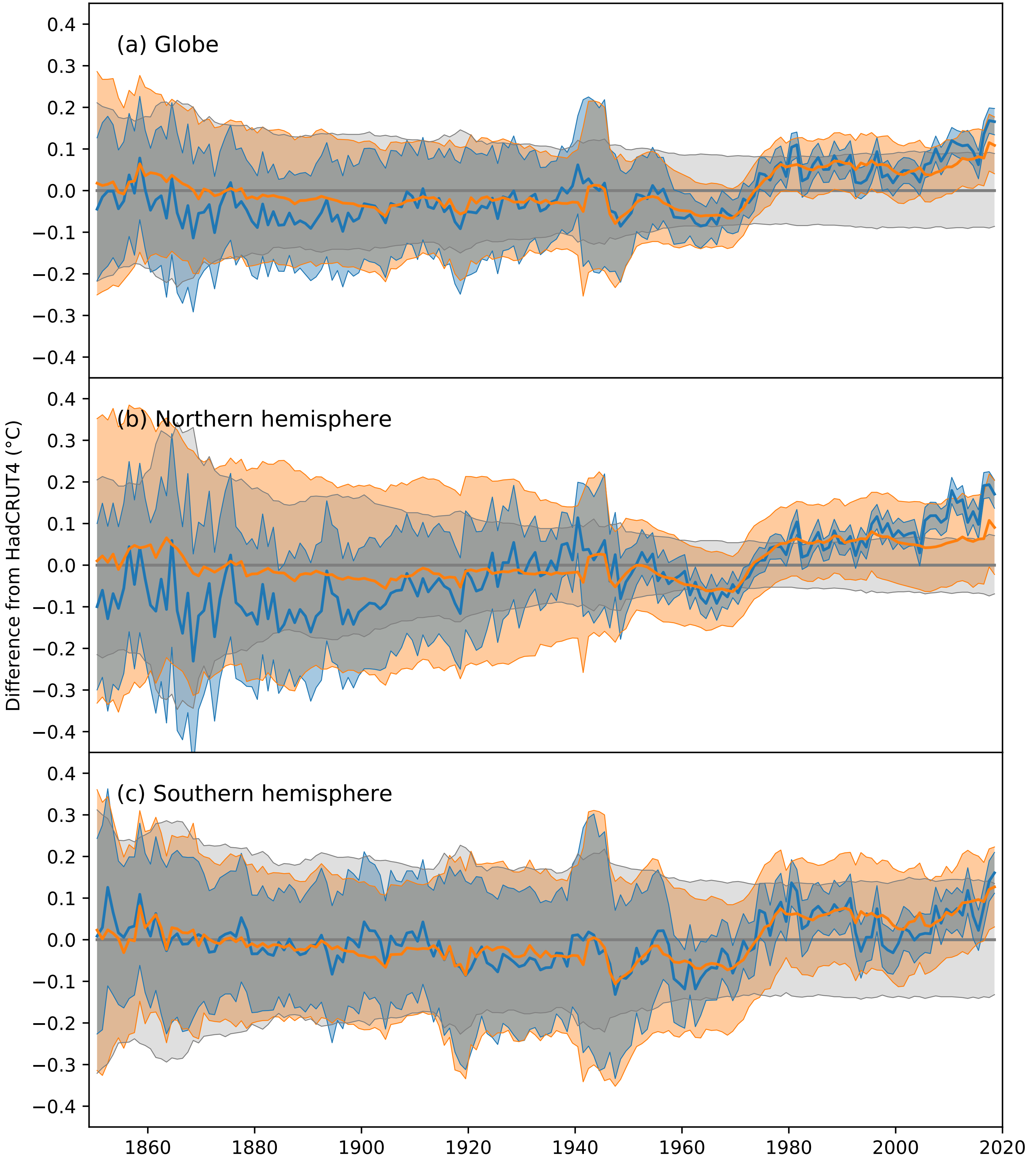
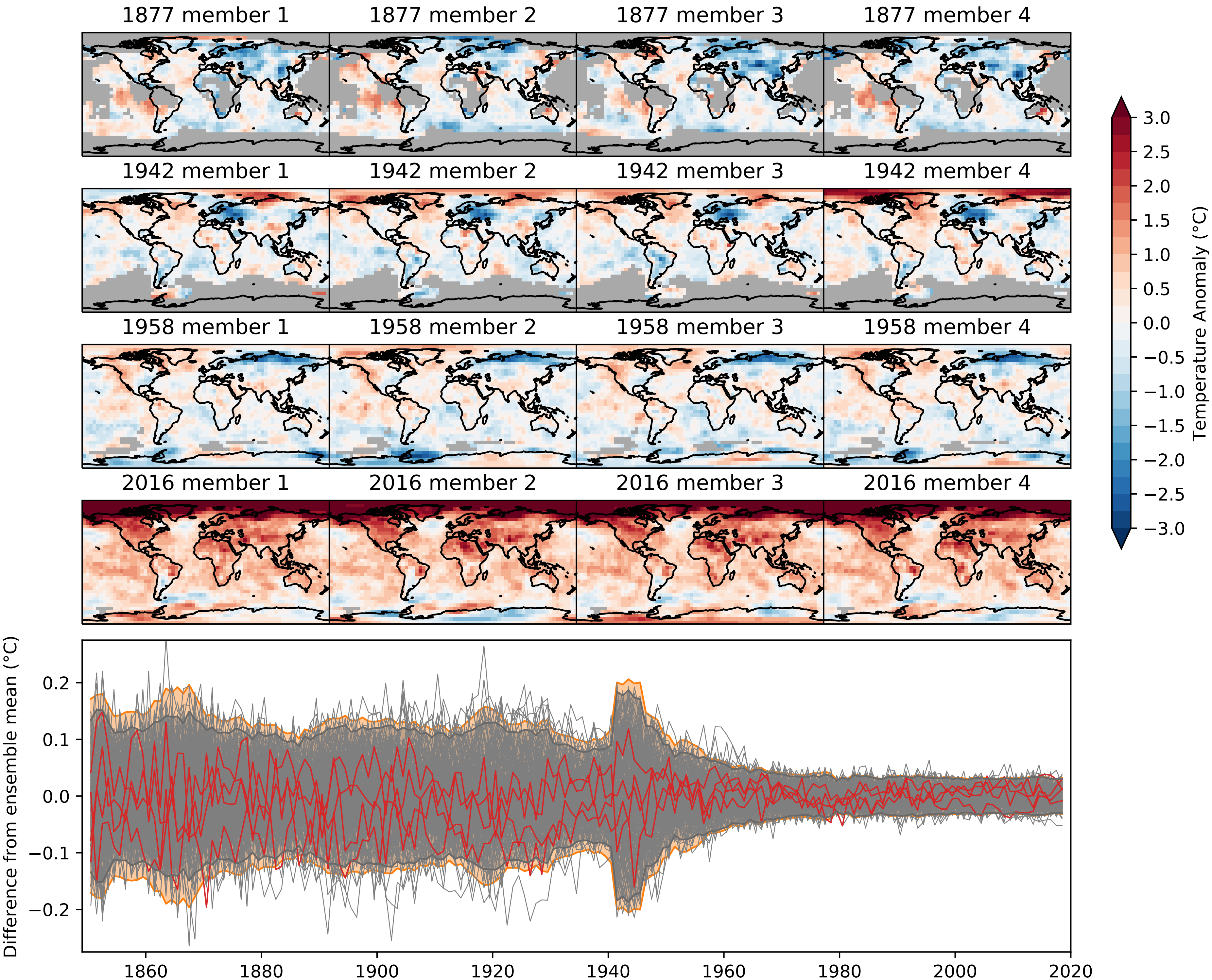
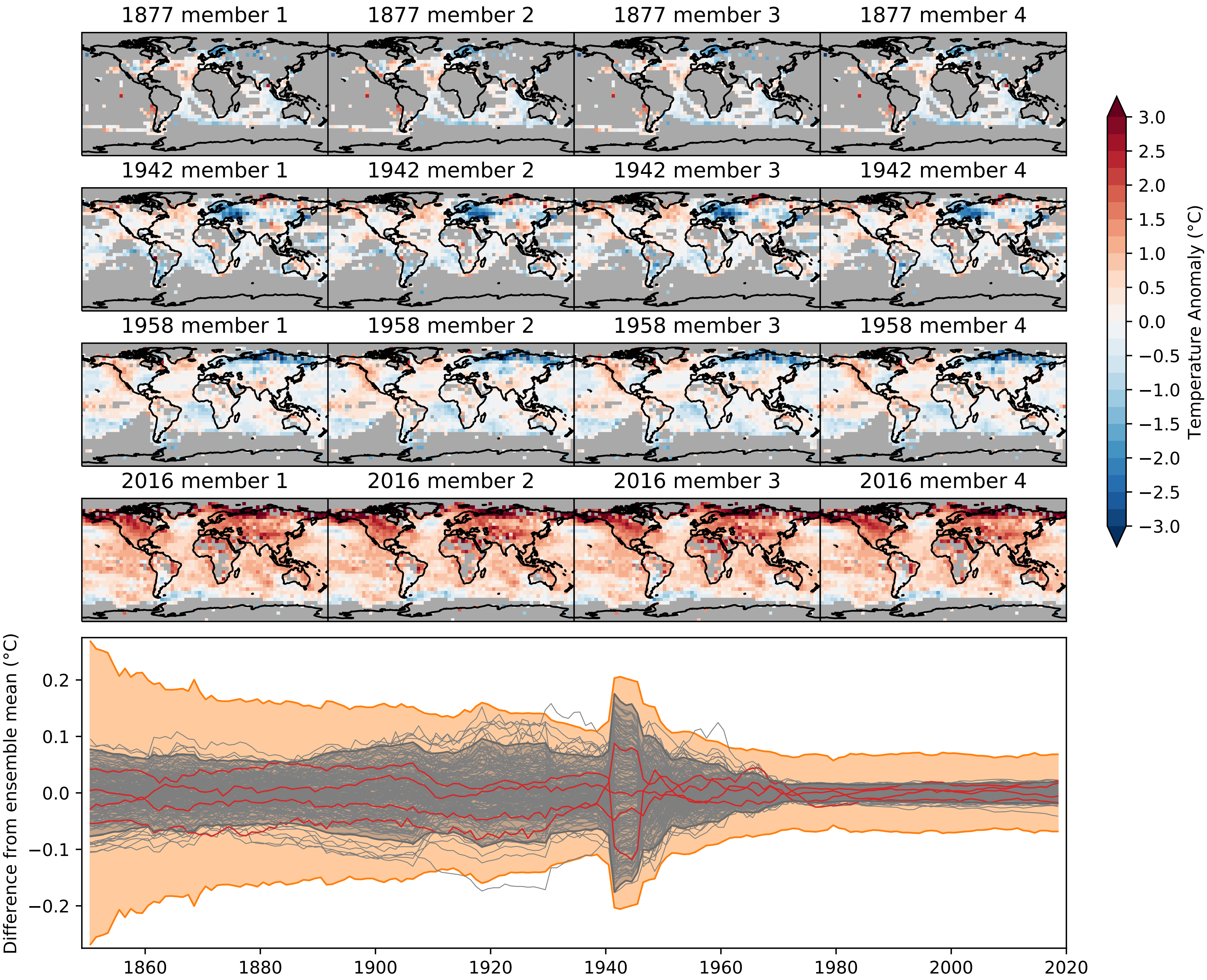

(a) Globe
(b) Northern hemisphere
(c) Southern hemisphere
Difference from HadCRUT4 (°C)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1860
1880
1900
1920
1940
1960
1980
2000
2020

Figure 2.



1877 member 1
1877 member 2
1877 member 3
1877 member 4
1942 member 1
1942 member 2
1942 member 3
1942 member 4
1958 member 1
1958 member 2
1958 member 3
1958 member 4
2016 member 1
2016 member 2
2016 member 3
2016 member 4
Temperature Anomaly (°C)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
Difference from ensemble mean (°C)
0.2
0.1
0.0
−0.1
−0.2
1860
1880
1900
1920
1940
1960
1980
2000
2020

Figure 3.



1877 member 1
1877 member 2
1877 member 3
1877 member 4
1942 member 1
1942 member 2
1942 member 3
1942 member 4
1958 member 1
1958 member 2
1958 member 3
1958 member 4
2016 member 1
2016 member 2
2016 member 3
2016 member 4
Temperature Anomaly (°C)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
Difference from ensemble mean (°C)
0.2
0.1
0.0
−0.1
−0.2
1860
1880
1900
1920
1940
1960
1980
2000
2020

Figure 4.

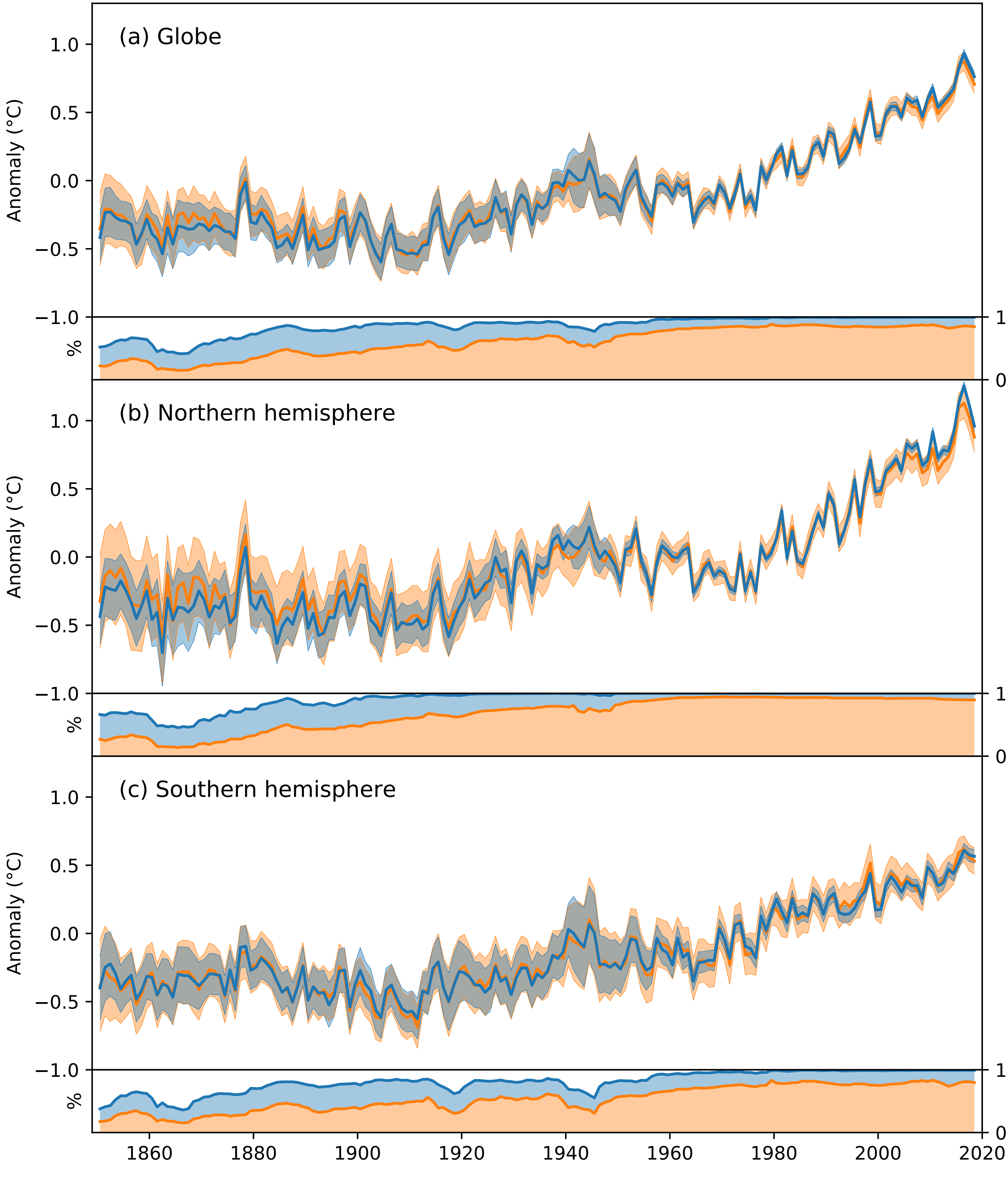
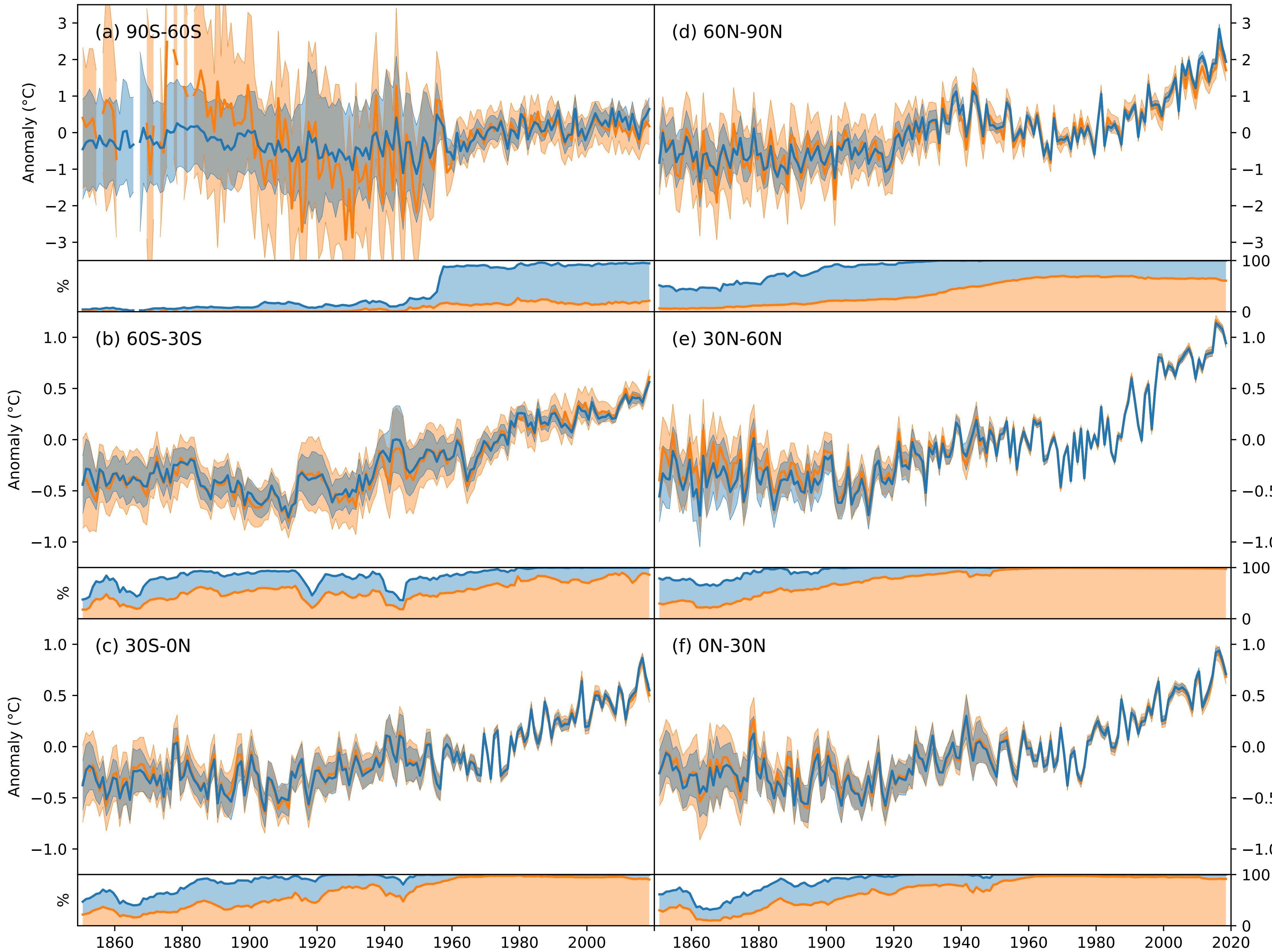
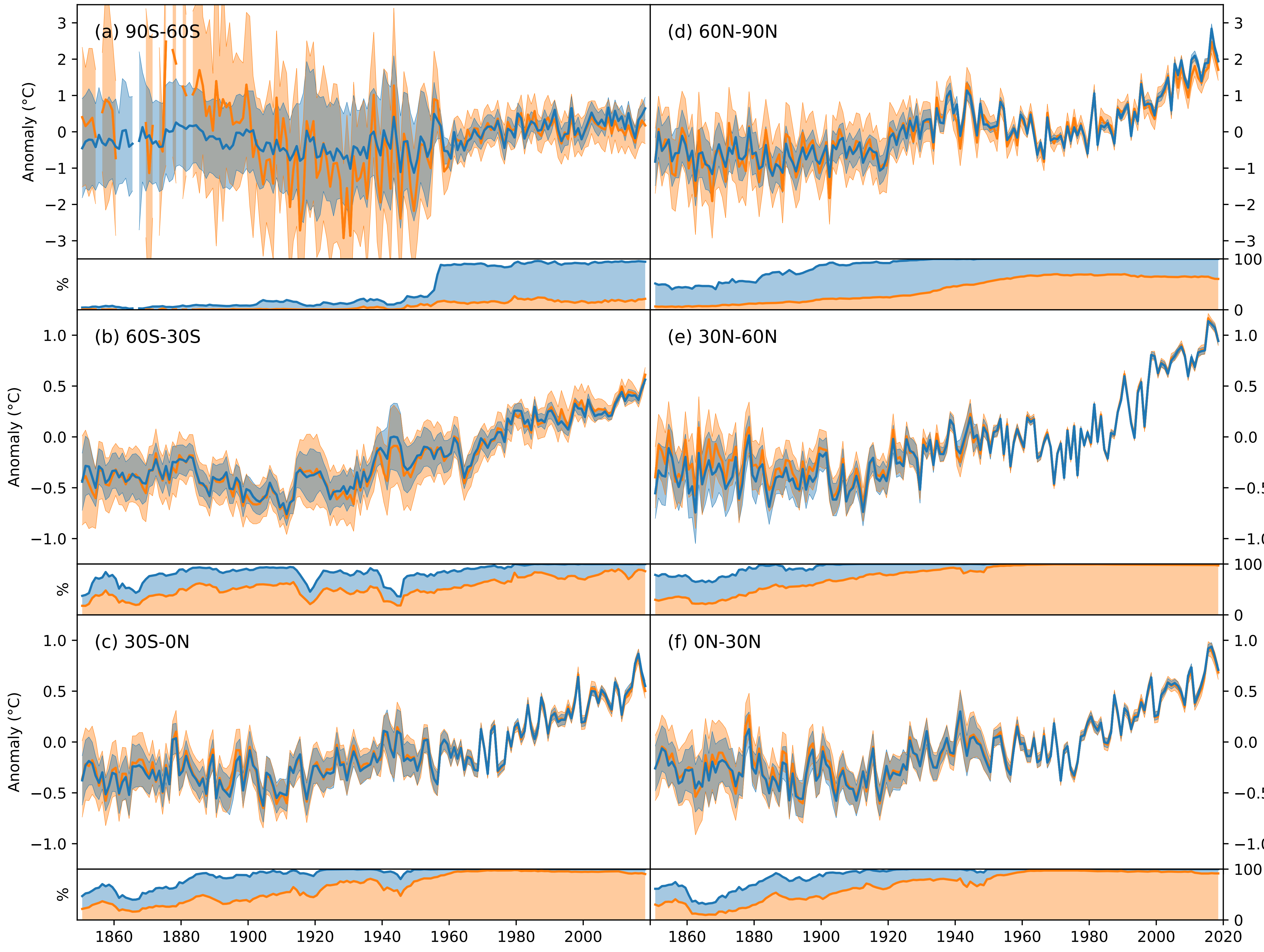

(a) Globe
(b) Northern hemisphere
(c) Southern hemisphere
Anomaly (°C)
%
1.0
0.5
0.0
−0.5
−1.0
100
0
1860
1880
1900
1920
1940
1960
1980
2000
2020

Figure 5.



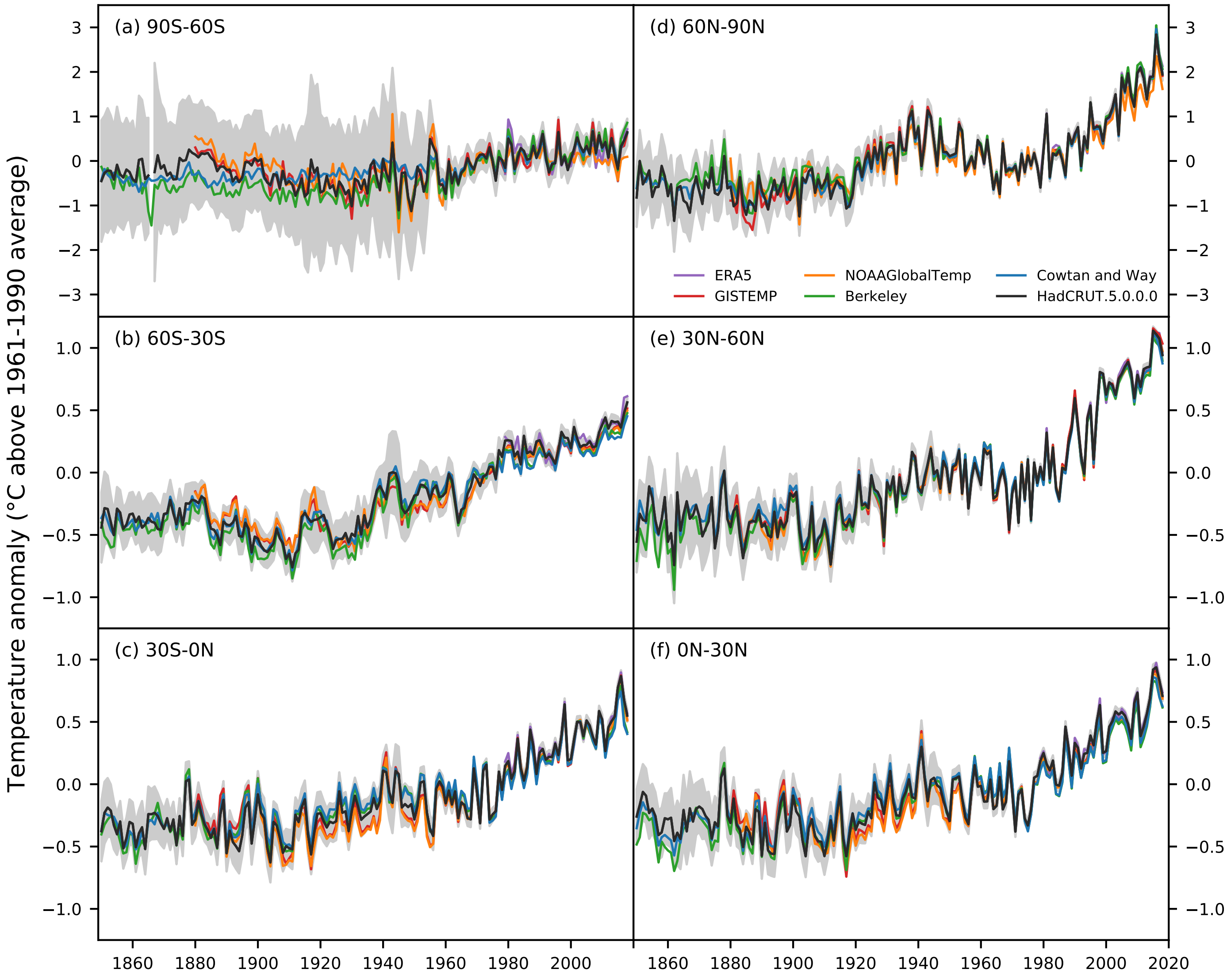
(a) 90S-60S
(b) 60S-30S
(c) 30S-0N
(d) 60N-90N
(e) 30N-60N
(f) 0N-30N
Anomaly (°C)
%

Figure 6.



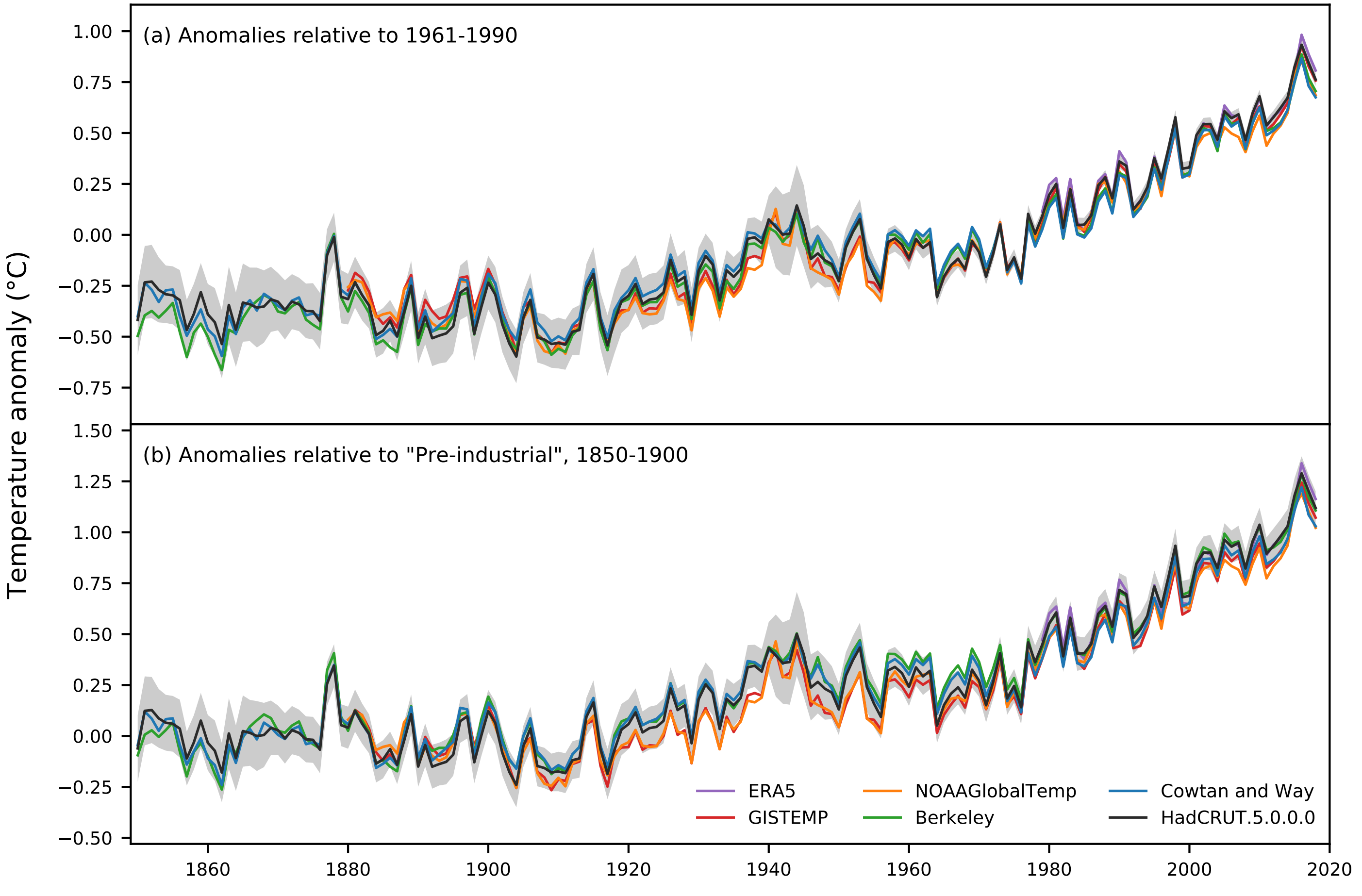
(a) 90S-60S
(b) 60S-30S
(c) 30S-0N
(d) 60N-90N
(e) 30N-60N
(f) 0N-30N
Anomaly (°C)
%
1860
1880
1900
1920
1940
1960
1980
2000
2020

Figure 7.



(a) 90S-60S
(b) 60S-30S
(c) 30S-0N
(d) 60N-90N
(e) 30N-60N
(f) 0N-30N
ERA5
GISTEMP
NOAAGlobalTemp
Berkeley
Cowtan and Way
HadCRUT.5.0.0.0
Temperature anomaly (°C above 1961-1990 average)
1860
1880
1900
1920
1940
1960
1980
2000
2020

Figure 8.



(a) Anomalies relative to 1961-1990
(b) Anomalies relative to "Pre-industrial", 1850-1900
Temperature anomaly (°C)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
1860
1880
1900
1920
1940
1960
1980
2000
2020
ERA5
GISTEMP
NOAAGlobalTemp
Berkeley
Cowtan and Way
HadCRUT.5.0.0.0