Improvements in the uncertainty model in the Goddard Institute for Space Studies Surface Temperature (GISTEMP) analysis

Nathan J. L. Lenssen[1,2], Gavin A. Schmidt[1], James E. Hansen[3], Matthew J. Menne[4], Avraham Persin[1,5], Reto Ruedy[1,5], Daniel Zyss[1]

[1]NASA Goddard Institute for Space Studies, New York, New York, USA
[2]Department of Earth and Environmental Sciences, Columbia University, New York, New York, USA
[3]Climate Science, Awareness and Solutions, Columbia University Earth Institute, New York, New York, USA
[4]NOAA National Centers for Environmental Information, Asheville, North Carolina, USA
[5]SciSpace LLC, New York, New York, USA

Key Points:

- A total uncertainty analysis for GISTEMP is presented for the first time.
- Uncertainty in global mean surface temperature is roughly 0.05°C in recent decades increasing to 0.15°C in the 19th Century.
- Annual mean uncertainties are small relative to the long term trend.
- The warmest year on record (so far) was 2016 with 86% confidence

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1029/2018JD029522

Abstract

We outline a new and improved uncertainty analysis for the Goddard Institute of Space Studies (GISS) Surface Temperature product version 4 (GISTEMP v4). Historical spatial variations in surface temperature anomalies are derived from historical weather station data and ocean data from ships, buoys and other sensors. Uncertainties arise from measurement uncertainty, changes in spatial coverage of the station record, and systematic biases due to technology shifts and land cover changes. Previously published uncertainty estimates for GISTEMP included only the effect of incomplete station coverage. Here, we update this term using currently available spatial distributions of source data, state-of-the-art reanalyses and incorporate independently derived estimates for ocean data processing, station homogenization and other structural biases. The resulting 95% uncertainties are near 0.05°C in the global annual mean for the last 50 years, and increase going back further in time reaching 0.15°C in 1880. In addition, we quantify the benefits and inherent uncertainty due to the GISTEMP interpolation and averaging method. We use the total uncertainties to estimate the probability for each record year in the GISTEMP to actually be the true record year (to that date), and conclude with 86% likelihood that 2016 was indeed the hottest year of the instrumental period (so far).

1 Introduction

Attempts to seriously estimate the changes in temperature at the hemispheric and global scale date back at least to *Callendar* [1938] who used 147 land-based weather stations to track near global trends from 1880 to 1935 [*Hawkins and Jones*, 2013]. Subsequent efforts used substantially more data (180 stations in *Mitchell* [1961], 400 stations in *Callendar* [1961], "several hundred" in *Hansen et al.* [1981] etc.), and with a greater global reach. While efforts were made to estimate the uncertainty associated with these products, efforts were more suggestive than comprehensive.

As the data sets have grown in recent years (through digitization and synthesis of previously separate data streams) [*Rennie et al.*, 2014; *Freeman et al.*, 2016; *Thorne et al.*, 2018], and efforts have been made to improve data homogenization, bias corrections and interpolation schemes, the sophistication of the uncertainty models has also grown. Notably, with the introduction of the Hadley Centre SST analysis HadSST3 [*Kennedy et al.*, 2011a,b], Berkeley Earth [*Rohde et al.*, 2013a], and the joint Hadley Centre and University of East Anglia's Climatic Research Unit HadCRUT4 [*Morice et al.*, 2012], Monte Carlo methodologies have been applied to generate observational ensembles that quantify uncertainties more comprehensively than was previously possible.

GISTEMP is a widely-used data product that tracks global climate change over the instrumental era. However, the existing uncertainty analysis currently contains only rough estimates of uncertainty on the land surface air temperature (LSAT) mean and no estimates of the sea surface temperature (SST) or total (land and sea surface combined) global mean. This paper describes a new end-to-end assessment of all the known uncertainties associated with the current GISTEMP analysis (nominally based in the methodology described in *Hansen et al.* [2010], but with changes to data sources as documented on the GISTEMP website and outlined below), denoted version 4. We use independently derived uncertainty models for the land station homogenization [*Menne et al.*, 2010, 2018] and ocean temperature products [*Huang et al.*, 2015, 2017], combined with our own assessment of spatial interpolation and coverage uncertainties, as well as parametric uncertainty in the GISTEMP methodology itself.

The analysis was performed in the open source language R [*R Core Team*, 2016] and the data, code, and intermediate steps needed to generate all figures in this report are available on the GISTEMP website (`https://data.giss.nasa.gov/gistemp/uncertainty`).

2 Overview of Surface Temperature Products

All of the most commonly cited surface temperature analyses split up the calculation of global anomaly fields into separate LSAT and SST anomaly analyses. These independent LSAT and SST analyses are combined into a total (LSAT and SST) global surface temperature index from which spatially averaged global and regional time series can be computed (note this is not strictly equal to the true surface air temperature anomaly [*Cowtan et al.*, 2015]). Likewise, the uncertainty analyses for the LSAT and SST are performed separately, then combined into total global uncertainty.

Semi-operational surface temperature analyses have been available since the first products by NASA/GISS and joint work from the Hadley Centre and Climatic Research Unit in the UK in the late 1970s. There are now multiple updated and peer-reviewed surface temperature products available, notably produced by NASA/GISS (GISTEMP), NOAA National Centers for Environmental Information (NCEI) with the Merged Land–Ocean Surface Temperature Analysis (MLOST), the Hadley Centre/Climatic Research Unit (HadCRUT), an analysis from the Japanese Meteorological Agency (JMA) [*Ishihara*, 2006] and a reanalysis-based product from ECMWF. These analyses use considerably different methods for the calculation of historical global and regional mean time series, but broadly agree on the trends and interannual variations in the global annual mean time series (fig. 1), though they differ at more regional scales as a function of data coverage and interpolation method [*Rao et al.*, 2018]. However, interpreting the comparisons across surface temperature products has to be nuanced since the raw data and intermediate product sources are often shared and not completely independent. Of the six major products that are currently being updated in real-time, GISTEMP was notable in not having rigorous confidence intervals on the global and regional mean time series.

The treatment of missing land surface data is a major distinction between products. Since monthly temperature anomalies are strongly correlated in space, spatial interpolation methods can be used to infill sections of missing data. However, smoothing due to interpolation obscures spatial variability as grid box estimates are some weighted combination of many stations. HadCRUT4 performs the least interpolation. If a $5^\circ \times 5^\circ$ grid box does not have any station data, this grid box is reported as missing [*Morice et al.*, 2012]. The HadCRUT method has the major advantage of clarity in that every grid box is the simple average of the station anomaly values contained in the grid box, but suffers in coverage, particularly in the critical Arctic region. At the other extreme, GISTEMP performs the most interpolation by giving stations a 1,200 km radius of influence, regardless of latitude [*Hansen et al.*, 2010]. The interpolation allows for infilling during the data-poor early years (pre-1960), but makes it more complex to determine how stations contribute to grid box values. We expand on the GISTEMP method in the following section. The NOAA method performs an intermediate amount of interpolation by aggregating a $5^\circ \times 5^\circ$ grid up to a $15^\circ \times 15^\circ$ grid before modeling the fine-scale variability using an empirical orthogonal function teleconnection analysis as described in Appendix A of *Smith and Reynolds* [2005]. The JMA method is similar to that of HadCRUT4. Comparisons to reanalyses products suggest that the interpolated products have less overall bias compared to the true global mean [*Simmons et al.*, 2016] because the missing data areas are predicted (and seen) to be changing more than the global mean.

Recently, the Berkeley Earth group [*Rohde et al.*, 2013a] and *Cowtan and Way* [2014] have released more statistically sophisticated products that confirm the observed warming in the NASA, NOAA, and HadCRUT products and provide a more natural uncertainty quantification. Berkeley Earth used an additive Kriging model for the LSAT analysis to estimate interpolated LSAT fields rigorously. Cowtan and Way took this approach a step further and used methods to interpolate both SST and LSAT fields used in HadCRUT. The results of Cowtan and Way suggest that the inclusion of interpolation is necessary to capture the global effect of the higher rate of warming in the Arctic.

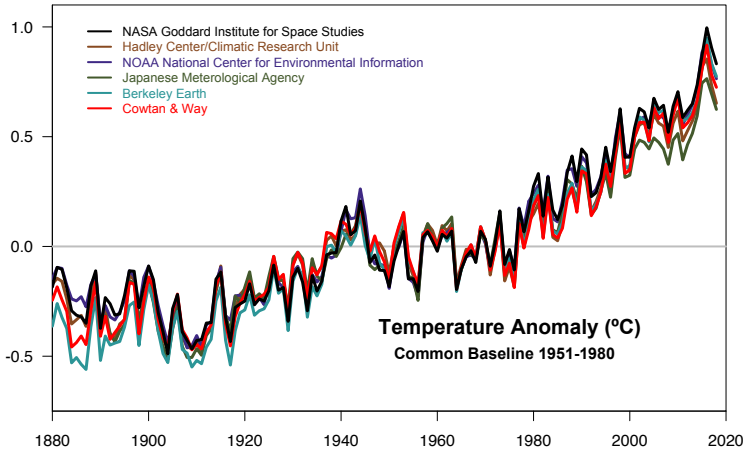


Figure 1. Comparison of six analyses of the annual global surface temperature anomaly through 2018.

3 Operational GISTEMP

The current operational method used in GISTEMP to compute the mean land surface temperature anomaly is an extended version of the process outlined by *Hansen and Lebedeff* [1987]. The analysis contains two major steps: interpolation of individual station data and averaging of interpolated fields. Preliminary to the two core steps, the monthly station data are processed following *Hansen et al.* [2010]. The publicly available code, written in Python, has been updated to modern standards [*Barnes and Jones*, 2011].

GISTEMP uses the equal-area grid developed in *Hansen and Lebedeff* [1987]. The Earth is divided into 80 equal-area boxes arranged in bands of constant latitude. By constraining each box to cover the same area, the bands have unequal numbers of grid-boxes resulting in an irregular grid. There are four bands in each hemisphere representing the polar region, mid-latitudes, subtropics, and tropics which respectively contain 4, 8, 12, and 16 equal area boxes. Therefore the bands account for 10, 20, 30, and 40 percent of the area of the hemisphere. Each of the 80 boxes are divided into 100 equal-area sub-boxes resulting in an equal-area grid of 8000 grid-boxes covering the Earth.

3.1 Interpolation Step

Calculating the values of the sub-boxes in the equal area grid from the station anomaly record is referred to as the interpolation step in this study. For a single sub-box, all stations within a given distance are successively combined starting with the longest record. A new station is averaged in if there is at least a 20-year overlap, and an offset is applied to leave the mean over that common period unchanged individually for each calendar month. The weight W for a station d km away from the sub-box center within a given radius r is determined using a linear radial basis function of the form

$$W_r(d) = \max\left(\frac{r-d}{r}, 0\right) \quad (1)$$

The value of the radius, r = 1,200 km, was estimated based on an investigation of the correlation of the annual mean series of pairs of stations as a function of their spatial separation [*Hansen and Lebedeff*, 1987]; this simple device turned out to be quite similar to the form of the estimated covariance function in the modified Kriging method used by the Berkeley Earth analysis [*Rohde et al.*, 2013b]. If there are no stations within 1200km of a sub-box center, it is given a missing value.

3.2 Averaging Step

The averaging step calculates the regional and global time series from the interpolated sub-box records. In this context, regional refers to hemispheric and the 8 latitudinal bands in the equal area grid. First, an average series is computed for each of the 80 equal area boxes by the method described in the interpolation step section, except that equal weight is given to each equal area sub-box series. The LSAT and SST data are combined when each of the 80 box series are created. In each sub-box, either a pure SST series or a pure LSAT series is selected. SST data are used only for ocean sub-boxes that contain no sea ice and whose center is more than 100 km off the nearest land station. Everywhere else we use the LSAT data.

The averages for the eight latitudinal zonal bands are then computed from the box series weighted by the number of sub-boxes with data. The three extra-tropical bands in each hemisphere are combined in the same way into a single series. These two series and the two tropical series are converted to anomaly series with respect to the 1951–1980 period. Global and hemispheric anomalies are computed as weighted averages of these four band means, weighted by the full area of these bands.

3.3 Changes to Operational GISTEMP 2010–2018

The only difference in methodology since *Hansen et al.* [2010] not caused by changes in the available input data, was combining into single polar boxes the 40 sub-boxes reaching the North and the South poles (starting September 2016). This only insignificantly affected the results, but produced more natural looking images near the poles.

All other changes relate solely to the input data. In 2010, GISTEMP was using GHCN-Monthly version 2 (GHCNv2), the U.S. Historical Climatology Network version 2.0 (USHCN2) and the Scientific Committee on Antarctic Research (SCAR) temperature data over land, with Hadley Centre Sea Ice and Sea Surface Temperature data set (HadISST) and Optimum Interpolation Sea Surface Temperature (OISSTv3) for the ocean. With the upgrade to GHCNv3 in December 2011 (and then to v3.2 in September 2012, and now to v4), the need for USHCN2 was obviated. In GHCNv3 as in GHCNv4, the various data series from different sources for a location, that were available in GHCNv2, are merged into a single series, and the resulting inhomogeneities are resolved in the adjustment procedure. Hence GISTEMP is using the adjusted GHCNv3 and v4 data. Whereas combining different sources at a location and manual corrections are no longer needed, the GISS urban adjustment scheme is still being applied. For the ocean data, the ocean temperature product was replaced with the more homogeneous Extended Reconstructed Sea Surface Temperature (ERSST) v3b in January 2013, which was updated to ERSSTv4 in July 2015, and to ERSSTv5 in August 2017. The impacts over time of these changes are recorded and maintained on the GISTEMP History page `https://data.giss.nasa.gov/gistemp/history`.

Analyses subsequent to *Hansen et al.* [2010] that use GHCNv3 are now being denoted GISTEMP v3. The integration of GHCNv4 into the GISTEMP code in January 2019 is denoted as GISTEMP v4.0; this version does not use the SCAR data except as far as they are part of GHCNv4. Going forward, a more rigorous version numbering scheme will be adopted to better track methodological and input data variations. GISTEMP v3

will nonetheless be maintained for the time being for legacy purposes. The uncertainty analysis presented here is strictly valid for GISTEMP v4.0, but the differences with it applied to v3 are insignificant and primarily arise from differences in GHCN homogenization.

3.4 Prior Uncertainty Estimates

GISTEMP has previously presented uncertainties due to incomplete spatial coverage of the station record [*Hansen and Lebedeff*, 1987]. Most recently, *Hansen et al.* [2010] reported estimates of this uncertainty for three large time periods: 1880–1900, 1900–1950, and 1960–2008. The analysis sub-sampled a long run of the GISS-ER climate model [*Hansen et al.*, 2007] according to the coverage of the station network on the Earth during these three time periods. This model had a $4° \times 5°$ latitude by longitude grid. Global annual land-only means of the sub-sampled model were compared with global annual land-only means using all of the grid-boxes.

Since the global mean calculation in GISTEMP aggregates from small sub-boxes to the 80 equal-area boxes, the coarse model grid approach has considerable value in quantifying the large-scale sampling uncertainty in the approach assuming that the model is capturing sufficient statistical structure of the underlying fine-scale global temperature anomaly field. The uncertainty calculation also roughly captures large-scale spatial and temporal sparsity. An equal-area box that has no data within 1200km is "missing" in the GISTEMP global and regional mean calculation and is on the approximate scale of the model grid. Furthermore, the large grid-box size of that model serves as a rough approximation of the interpolation step of the GISTEMP procedure.

We address a number of deficiencies in the legacy GISTEMP LSAT sampling uncertainty analysis in this study. The first goal is increasing the temporal resolution of uncertainty from around 50 years to decadal estimates of LSAT uncertainty. Further refinements to the annual or even monthly timescale do not make a substantive difference. Second, we aim to better capture the uncertainty in the interpolation step of GISTEMP. The coarse resolution of the previously used model grid does not describe the fine scale behavior of the true temperature anomaly field and does not allow us to replicate the interpolation step. As we detail in the following section, we now use a product with a much finer horizontal grid to replicate the entire GISTEMP global and regional mean calculation. Thus, we are more confident that our calculated uncertainties will reflect the actual analysis method used. Finally, we compare the uncertainties of the GISTEMP band averaging scheme with a simple latitude-weighted mean in the mean land surface temperature uncertainty.

The previously reported GISTEMP uncertainties do not include parametric uncertainties due to homogenization of the station record or uncertainties associated with the SST reconstruction. By adding in the homogenization uncertainty from the GHCN dataset and propagating the uncertainty from the ERSSTv5 dataset through the GISTEMP procedure, we obtain a holistic estimate of the full uncertainty in the GISTEMP product.

4 Sources of Uncertainty

4.1 Statistical Formulation of Uncertainty

Before outlining the sources of uncertainty in the land and ocean reconstructions, it is useful to step back and discuss the underlying statistics in general terms. Letting $\mu(t)$ be the true (latent) global anomaly for a year t, we view the calculated (observed) annual mean temperature anomaly $A(t)$ as

$$A(t) = \mu(t) + \epsilon(t). \tag{2}$$

In this formulation, $\epsilon(t)$ is a random variable that represents the total uncertainty in our estimate of the annual mean temperature anomaly. Assuming that our estimation procedure is unbiased (an assumption we will revisit in the discussion of our results), the expected value $\mathbb{E}[\epsilon(t)] = 0$ for all years t. The uncertainty in our calculation of the global mean is then defined as

$$\mathcal{E}(t) = \mathrm{Var}(\epsilon(t)). \tag{3}$$

Our analysis breaks down the uncertainty into two components: the uncertainty in the global mean due to uncertainties in the land calculation $\epsilon_L(t)$ and uncertainty in the global mean due to uncertainties in the sea surface calculation $\epsilon_S(t)$. We decompose our total uncertainty as

$$\epsilon(t) = \epsilon_L(t) + \epsilon_S(t). \tag{4}$$

If these uncertainties are independent, the calculation of the uncertainty is the sum of the individual variances

$$\mathcal{E}(t) = \mathrm{Var}(\epsilon(t)) = \mathrm{Var}(\epsilon_L(t)) + \mathrm{Var}(\epsilon_S(t)) \tag{5}$$

We proceed on the assumption that the land and ocean uncertainties are independent. However, there is potentially correlation between the uncertainty due to the land calculation and the uncertainty due to the ocean calculation. In addition to correlation between the land and ocean uncertainties, we also expect some amount of correlation in time, particularly at the monthly time scale. Not accounting for positive correlation of uncertainties in time will lead to underestimation of the uncertainty. To reduce the impact of this autocorrelation, we look at the annual mean temperature anomalies which exhibit much lower autocorrelation.

4.2 Land Surface Temperature Uncertainty

Quantifying the uncertainties that arise from using the land station record to calculate regional and global land-only mean temperatures has been an active field for many years. In particular, NOAA [*Vose et al.*, 2012] and HadCRUT [*Morice et al.*, 2012] groups have developed sophisticated uncertainty models for this portion of the analysis. It is generally assumed that there are three major independent sources of uncertainty in the land record that add uncertainty to global temperature calculations: station uncertainty, bias uncertainty, and sampling uncertainty. We will outline these three briefly (though see *Brohan et al.* [2006] for a detailed discussion). As with the operational GISTEMP, we define the land surface as any grid-box that is classified as land or sea ice.

4.2.1 Station Uncertainty

Station uncertainty encompasses the systematic and random uncertainties that occur in the record of a single station and include measurement uncertainties, transcription errors, and uncertainties introduced by station record adjustments and missed adjustments in post-processing. The random uncertainties can be significant for a single station, but comprise a very small amount of the global LSAT uncertainty to the extent that they are independent and randomly distributed. Their impact is reduced when looking at the average of thousands of stations.

The major source of station uncertainty is due to systematic, artificial changes in the mean of station time series due to changes in observational methodologies. These station records need to be homogenized, or corrected to better reflect the evolution of temperature. The homogenization process is a difficult, but necessary statistical problem that corrects for important issues albeit with significant uncertainty for both global and local temperature estimates.

4.2.2 Bias Uncertainty

Bias uncertainty refers to the biases in a single station record due to non-climatic sources. Thermometer exposure change bias [*Parker*, 1994] refers to biases introduced to the station record by the evolution of temperature measurement techniques, such as the switch to Stevenson screens in the 19th Century or the change to Max-Min Temperature Sensor (MMTS) automated recorders in recent decades in the United States [*Menne et al.*, 2009]. Urban biases are not due to systematic biases in the instrumentation, but rather due the local warming effect of urban centers through land surface changes, reductions in evapotranspiration, and local heat sources. These urban biases are corrected for in our global temperature study since the goal is to understand the changes in the global climate system, not the localized effect of urban heat islands. An urban bias correction was added to GISTEMP in 1998 [*Hansen et al.*, 1999]; it confirmed that its impact on global temperature anomalies is small. As shown in *Hansen et al.* [2010], the effect of the urban adjustment on global temperature change is on the order of 0.01°C.

4.2.3 Sampling Uncertainty

Sampling uncertainty is an umbrella term for uncertainties introduced into global and regional annual means by incomplete spatial and temporal coverage. Whereas the station uncertainties are observed to mostly cancel out in modern-era global annual means, as many of the uncertainties are independent from station to station, the sampling uncertainties remain significant. Understanding the sampling uncertainty of GISTEMP is crucial because, unlike HadCRUT, GISTEMP extrapolates out the anomaly field into regions without station data. Quantifying the sampling uncertainty will provide a measure of confidence in the extrapolation. Since reduction in bias in the global mean due to interpolation comes with an uncertainty variance increase, we need to ensure that the interpolation does not drastically inflate the sampling uncertainty.

Quantifying the sampling uncertainty is critical to providing uncertainties for the mean temperatures for two reasons. First, the HadCRUT analysis has shown that the sampling uncertainty is a significant component of the uncertainty in the global annual means in the modern instrumental era [*Morice et al.*, 2012]. Second, updating the sampling uncertainty model provides transparent continuity in the GISTEMP analysis for numerous researchers that rely on the data product for their own analyses. As we will detail in the following section, GISTEMP has historically made only rough estimates of the sampling uncertainty. Our update here provides a transition from the original GISTEMP uncertainty model towards a more modern statistical approach.

4.3 Sea Surface Temperature Uncertainty

The current production versions of GISTEMP use the ERSSTv5 product provided by NOAA/NCEI [*Huang et al.*, 2017] for ocean temperatures. ERSSTv5 uses the same underlying method [*Huang et al.*, 2015] and uncertainty quantification method [*Liu et al.*, 2015; *Huang et al.*, 2016] as the previous generation ERSSTv4. The major upgrade in v5 is a more sophisticated parameter tuning, resulting in more realistic spatiotemporal patterns in the reconstructed SST fields. In addition, v5 incorporates new data sources from the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) 3.0 [*Freeman et al.*, 2016] and the Argo float network of near-surface readings.

The uncertainty calculation in ERSSTv4/v5 breaks down the ocean uncertainty into two independent components: parametric uncertainty and reconstruction uncertainty [*Liu et al.*, 2015; *Huang et al.*, 2016]. Parametric uncertainty quantifies the internal statistical variability of the ERSST procedure and is defined by the standard deviation of a perturbed parameter ensemble. The ensemble has been constructed such that the parametric uncertainty contains both the bias and sampling uncertainty [*Huang et al.*, 2016]. Recon-

struction uncertainty represents the information lost in using a finite number of empirical orthogonal teleconnection (EOT) functions to model the high-frequency component. Reconstruction uncertainty can be large at small spatial scales, but averages out to nearly zero at global scales as seen in Figure 2c of *Huang et al.* [2016]. Since we are concerned with global and hemispheric mean uncertainty in this study, it is reasonable to ignore the reconstruction uncertainty and focus only on the parametric uncertainty.

5 Update to GISTEMP's Uncertainty Analysis: Methods

5.1 Updated Land Surface Temperature Uncertainty Methodology

5.1.1 Data Sources

GHCN: The primary data source for LSAT data in GISTEMP v4.0 is the GHCN product from NOAA/NCEI. As mentioned in our discussion of the updates to operational GISTEMP in Section 3.3, we have replaced the combined GHCNv3 and SCAR with GHCNv4 as of January 2019. Thus, we perform our LSAT uncertainty analysis using GHCNv4, but comment briefly on how the results apply to GISTEMP v3. GHCNv4 contains significantly more stations than GHCNv3/SCAR, though many of the additional time-series are short. In general, the added stations in GHCNv4 do not significantly alter spatial coverage after interpolation and so will not effect the spatial uncertainty significantly, though it does slightly reduce some homogenization uncertainty. We compared the number of grid boxes in the Modern-Era Retrospective Analysis for Research and Applications (MERRA) model that contained a station with decadal coverage within the 1,200 km interpolation radius of influence (Figure 2) and find nearly no difference between versions. The increased quantity of stations will likely be most useful for more localized analyses.

Reanalyses: We use the three distinct reanalysis products as globally complete 'ground truth' temperature fields to quantify the contribution of the incomplete spatial and temporal coverage of the station record to the uncertainty in the global temperature anomaly. They are the fifth generation European Centre for Medium-range Weather Forecasting (ECMWF) atmospheric reanalysis (ERA5), the JMA JRA-55 analysis (hereafter JRA), and MERRA-2 (hereafter MERRA). Since the legacy approach inherently aggregates spatially due to large grid-box size it cannot utilize GISTEMP's interpolation method for the uncertainty analysis. In this study, we take a similar methodological approach using a high-resolution reanalysis product in place of the climate model output. The rough idea is the same: total coverage global means are compared with realistic (reduced) coverage global means and the uncertainty is described by summary statistics. The finer spatial resolution of the reanalyses than the climate model used previously allows us to treat single grid-box temperature anomaly values as station anomalies. The combination of improved spatial resolution and an analysis more closely mirroring the production GISTEMP procedure will give us more robust calculations of the coverage uncertainty for the global mean.

The primary reanalysis used in our study is ERA5 from 1979–2018 [*Copernicus Climate Change Service (C3S)*, 2017]. We average the 2m temperature to the $0.5° \times 0.625°$ MERRA grid to facilitate comparison and speed up computation. We find no significant changes to our results when verified on the native 31km grid. We choose ERA5 as the primary reanalysis since it best replicates the observed global mean over its record. Furthermore, we find that ERA5 and JRA55 produce generally consistent results while results found using MERRA often deviate.

MERRA provides monthly temperature means for the entire Earth from 1980–2018 at a $0.5° \times 0.625°$ resolution [*Gelaro et al.*, 2017]. The addition of the MERRA reanalysis is also due to the observational data sources used in assimilation. Since our goal is to determine the uncertainty that arises from the incomplete coverage of the GHCN station record, it is ideal to use a reanalysis that does not incorporate any GHCN information. Over land, MERRA only assimilates surface temperature data from the surface reading of

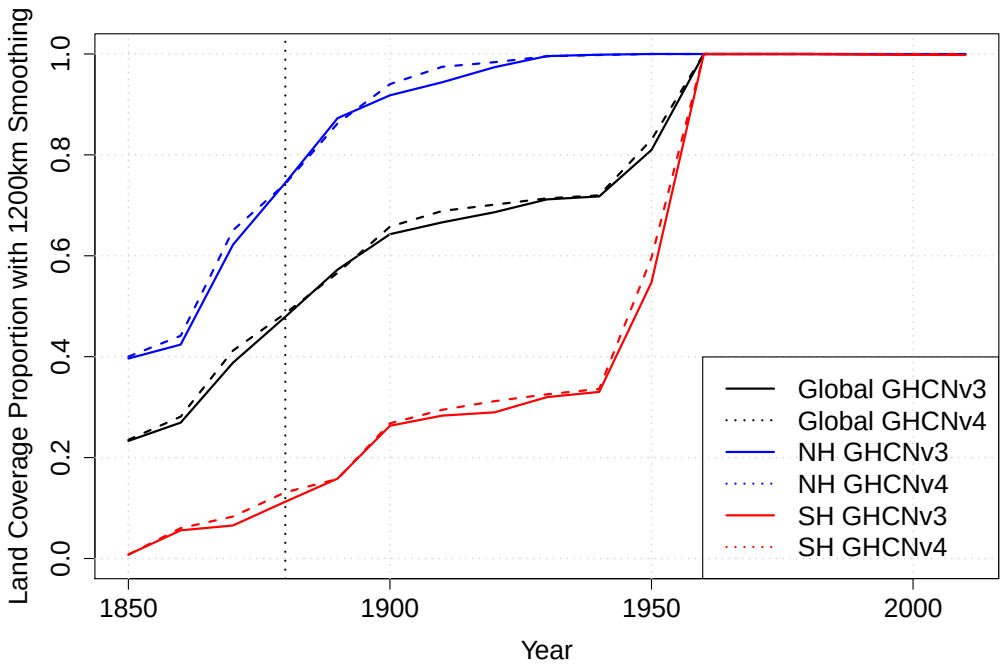


Figure 2. A comparison of the decadal land area coverage proportion in GHCNv3 and v4. A location is said to be covered if it is within 1200 km of a station with decadal coverage and will be included in the production GISTEMP analysis.

the radiosonde network, ensuring that we are fitting our statistical model for GHCN sampling uncertainty over an independent data source [*McCarty et al.*, 2016]. We also verify all results with the JRA55 reanalysis over 1979–2013 [*Kobayashi et al.*, 2015] to provide clarity when the results from MERRA and ERA5 disagree.

5.1.2 LSAT Sampling Uncertainty Method

A grid-box is determined to be land for the purpose of our study if its land area proportion is greater than 0% on the MERRA grid, approximately replicating the 100 km influence of land stations onto ocean grid cells in operational GISTEMP. As in operational GISTEMP, we determine sea ice extent for each month by the maximal extent of sea ice in the MERRA reanalysis. Grid-boxes that are not classified as land are classified as ocean with uncertainty quantified by the SST uncertainty analysis.

Monthly temperature anomalies are computed for the entire reanalysis grid for each of the 12 months by removing the single month mean for each grid-box time series. The full monthly temperature anomaly fields are used to calculate the baseline global and zonal annual means. We use a modified version of the GISTEMP averaging step with the same zonal bands and 80 equal area grid boxes and replace the sub-boxes with the reanalysis grid. The baseline global mean represents the true global anomaly $\mu(t)$ which will be compared with the means calculated with reduced coverage $A(t)$.

The spatial sub-sampling of the anomaly field is determined at a decadal temporal resolution. A station has temporal coverage in a decade if it has coverage for at least 5 of the 10 years. To have coverage for a year it must have coverage for at least 3 seasons

which requires at least two months in the season. A grid box is said to have coverage in a decade if it contains at least one station with coverage as defined above. Using these definitions, we create 14 decadal coverage masks on the grid, one for each decade from the 1880s to the 2010s. That is, we have a constant mask that describes the coverage of the observing network for each decade.

Reduced coverage global annual means, $A_k(t)$ are calculated for each of the 14 decadal time periods using a modified GISTEMP procedure. In the notation $A_k(t)$, k represents the decade used and t represents the year in the reanalysis record. The interpolation step is performed on the grid using a radius of 1200km. Then the averaging step is performed as described in the baseline global mean calculation with the sub-boxes taken to be the area-weighted gridboxes. Thus, the baseline global mean is an annual time series indexed by t spanning 1980–2017. There are $k = 1, \ldots, 14$ reduced coverage global means for the 14 decades of the study, each annual time series spanning from 1980–2017.

As the sampling uncertainty in ocean regions is quantified as part of the SST uncertainty analysis for ERSSTv5, we only include land area in the LSAT sampling uncertainty. The global and reduced coverage global land means are taken over land and sea ice regions following the GISTEMP procedure. Sea ice regions are defined using MERRA as the maximum extent of ice for each month over the reanalysis record.

We calculate $\mathcal{E}_L(t)$, the variance of the sampling uncertainty in GISTEMP, by the sample variance of the difference between the baseline global mean and each of the mask means. Rearranging equation 2, we define the difference series $D_k(t)$ for decade k as

$$D_k(t) \equiv \mu(t) - A_k(t) = \epsilon_k(t). \tag{6}$$

Then the uncertainty is $\mathrm{Var}(D_k(t))$. Note that this method assumes that our method of calculating the global mean does not have any systematic bias.

5.1.3 LSAT Sampling Extensions

Our sampling uncertainty analysis allows us to investigate other properties of the GISTEMP LSAT method. We describe three experiments addressed in our study. First, we challenge the assumption that the land surface mean temperature estimate is an unbiased estimate. Then, we calculate the minimum achievable sampling uncertainty due to the GISTEMP interpolation assuming full global station coverage. Finally, we provide one measure of the value of the GISTEMP averaging method.

Sampling Bias: Recent studies have shown the likely presence of bias in surface temperature products compared to the true global mean [*Simmons et al.*, 2010; *Cowtan and Way*, 2014; *Karl et al.*, 2015; *Jones*, 2016; *Simmons et al.*, 2016]. In addition, recent evidence from remote sensed temperature analyses suggest production GISTEMP may be underestimating Arctic warming [*Susskind et al.*, 2019]. To quantify the potential sampling biases due to limited station coverage, we introduce a potential systematic additive bias α_k and multiplicative bias β_k. Then, determining the variance of ϵ_k can be formulated as the univariate regression

$$\mu(t) = \alpha_k + \beta_k A_k(t) + \epsilon_k(t) \tag{7}$$

Since we are working with anomalies that are standardized over the entire time period of ERA5 (1979–2018), the additive bias $\alpha_k = 0$ for all decades as all of our grid-box time series are mean zero. However, we fit the full linear regression as a sanity check as it will have practically no effect on our estimation of β_k or the uncertainty. Since the ERA5 reanalysis currently spans 1979-2018, only the estimates for the 1980s though 2010s are representative of potential bias in operational GISTEMP. The estimates for decades pre-1980 do not reflect the actual bias in GISTEMP during their periods as the underlying climate variability is not properly accounted for. However, the estimates of bias due to limited coverage in early decade are useful for understanding the importance of station coverage for capturing the current pattern of global temperature change.

Limiting Uncertainty: Running the sampling uncertainty analysis in Section 5.1.2 with the assumption that we have station coverage for every land grid box provides a lower bound of the sampling uncertainty. We expect the limiting uncertainty to be greater than zero as the smoothing arising from interpolation increases the uncertainty in the global mean. Calculation of the limiting uncertainty is important to determine the relative values of increased data availability and methodological improvements for lowering the uncertainty of the global mean estimate. In addition to quantifying the lower uncertainty bound, we run the sampling bias analysis with the simulated full coverage to determine if the GISTEMP method has any systematic bias in an idealized case over the 1979–2018 period.

Comparison of Averaging Methods: In addition to using the GISTEMP band-average method, we run the sampling uncertainty analysis in Section 5.1.2 using a simple latitude-weighted mean. Comparison of the resulting LSAT sampling uncertainties shows the difference between the two averaging methods in accounting for missing data.

5.1.4 GHCN Homogenization Uncertainties

Station uncertainty due to homogenization of station series is quantified in the GHCNv4 analysis and incorporated in the GISTEMP uncertainty analysis with no modification[*Menne et al.*, 2018]. The GHCNv4 method divides the total homogenization uncertainty for land stations into two independent components: the parametric uncertainty associated with the Pairwise Homogenization Algorithm (PHA; *Menne and Williams* [2009]) used to homogenize the GHCNv4 monthly data and incomplete homogenization caused by artificial shifts in the data that remain undetected by the PHA.

The PHA detects artificial time series mean shifts due to changes in observing practice by comparing a station series with neighboring stations [*Menne and Williams*, 2009]. Various parameters, such as the minimum number of neighboring stations, are set in implementing the PHA and affect the sensitivity and accuracy of the method. Parametric uncertainty is quantified by running the PHA as an ensemble whose members have randomly varying parameter settings from a set of configurations that produced the best results when run on realistic benchmark datasets [*Williams et al.*, 2012]. For GHCNv4 monthly, 100 different versions of the PHA were used to homogenize the GHCNv4 data, yielding 100 different homogenized versions of each GHCN station record [*Menne et al.*, 2018]. The parameter uncertainty is determined by the sample standard deviation of the 100 feasible records.

While the PHA detects large (>0.2°C) breaks in time series, it (and other breakpoint detection methods) is unable to detect small shifts. This uncertainty associated with incomplete homogenization is estimated by adding small adjustments to the homogenization ensemble members at random dates and with random magnitudes. The frequency and magnitude of the added adjustments were determined by estimating the distribution of the missed (mostly small) breaks from the distribution of actual breaks detected by the PHA. Detected breaks in GHCNv4 have a bimodal distribution with peaks around ±0.5°C. In between these peaks is the so-called "missing middle" of the distribution, which *Menne et al.* [2018] estimated as having a mean of about −0.01°C and a standard deviation of 0.2 with an average frequency of occurrence of about 1 in 50 years. The number of missed adjustments for each station record in the ensemble was determined by sampling from a Poisson distribution with an average frequency of 1 in 59 years, and their magnitude was selected by a random draw from a normal distribution $\mathcal{N}(-0.01, 0.2)$.

5.1.5 Total Land Surface Temperature Uncertainty Methodology

As introduced in our discussion of sources of land uncertainty in Section 4.2, land surface temperature uncertainty arises due to station, bias, and sampling uncertainties.

Since the three sources are independent and we can ignore the bias uncertainty for means of large spatial scale, the total uncertainty is simply the sum of the station homogenization and sampling uncertainties. As these uncertainties are expressed as variances, it is critical that the variance for the homogenization and sampling uncertainties are added rather than the standard deviations or confidence intervals.

5.2 Sea Surface Temperature Uncertainty Methodology

We use the uncertainty analysis from ERSSTv4 to quantify the uncertainty in the ocean temperature in the GISTEMP analysis as ERSSTv5 did not make any changes to the underlying reconstruction or uncertainty methods [*Liu et al.*, 2015]. ERSSTv4 quantified their uncertainty through an ensemble of feasible SST fields rather than a single uncertainty field. The largest ensemble simulation contains 1,000 members and was constructed to quantify the parametric uncertainty in their prediction [*Huang et al.*, 2016]. Our analysis utilizes this 1,000 member large ensemble to understand how the uncertainty in the ERSST product impacts the GISTEMP uncertainty.

The parametric SST global and hemispheric uncertainty calculation closely follows the analysis performed by the ERSST team [*Huang et al.*, 2016]. We perform the GISTEMP averaging step with no land data for each of the 1,000 ensemble members resulting in 1,000 possible global and hemispheric time series. That is, we calculate the global mean with an ocean-only mask for each of the ERSST ensemble members. The 95% confidence interval for the parametric uncertainty of the SST model are calculated for each time point using the empirical 95% confidence interval of possible global mean sea surface temperature.

Our assumption in this calculation is that the ERSST large ensemble is symmetric about the median for global and hemispheric means and that ERSSTv5 is the median value of the ensemble. Both of these assumptions are not perfect, but reasonable for these large scale means. We find that the mean and median of the global sea surface temperature mean ensemble are nearly identical. Furthermore, the strong agreement between the operational and ensemble global mean (and thus global median from our result) in Figure 12 of *Huang et al.* [2016] supports the assumption that the global uncertainty is symmetric.

5.3 Total Global Uncertainty Methodology

The final step in the global uncertainty analysis is the combination of the separate land and ocean uncertainties into a total global uncertainty. If $\overline{A}(t)$ is the annual global mean anomaly for a year t and given an estimate of the global mean anomaly $\tilde{A}(t)$, we define the uncertainty of the global annual mean temperature as

$$\mathcal{E}(t) = \mathrm{Var}\left(\tilde{A}(t)\right) \tag{8}$$

The land-only uncertainty is comprised of the sampling uncertainty calculated using the method described in Section 5.1 with missing values for all of the ocean grid cells and the homogenization uncertainty according to the GHCNv4 analysis. Likewise, ocean-only uncertainty is calculated using the method described in Section 5.2 with missing values for all of the land grid cells. The resulting uncertainties then describe the uncertainty over a subset of the area of the Earth.

To calculate the global uncertainty, we account for the different regions and therefore area-proportions of Earth that the land and ocean cover. We define $\mathcal{E}_L$ as our LSAT uncertainty estimate from the reanalysis sub-sampling and $\mathcal{E}_S$ as the SST uncertainty estimate from the ERSST ensemble analysis. Using a_L and a_S (the area of the land and ocean on the Earth respectively) and assuming that the land and ocean uncertainty com-

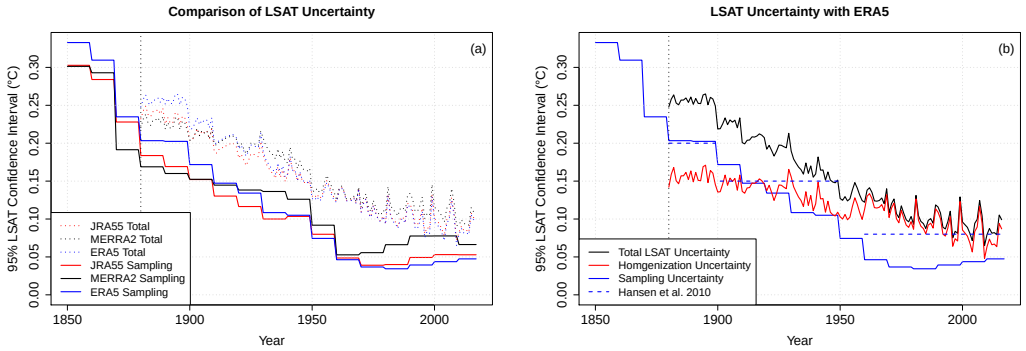


Figure 3. The total uncertainty (2σ) in the global annual mean land surface temperature decomposed into the sampling and homogenization uncertainty components where the homogenization uncertainty is found in an independent analysis and is currently limited to 1880 [*Menne et al.*, 2018]. (a) The sampling and resulting total LSAT uncertainty calculations using the three reanalyses. (b) The LSAT uncertainty as calculated with ERA5, the reanalysis selected for the analysis. The LSAT sampling uncertainty estimate from *Hansen et al.* [2010] is shown for comparison.

ponents are independent, the total global uncertainty variance is

$$\mathcal{E}(t) = \left(\frac{a_L}{a_L + a_S}\right)^2 \mathcal{E}_L(t) + \left(\frac{a_S}{a_L + a_S}\right)^2 \mathcal{E}_S(t) \tag{9}$$

Hemispheric and other regional combined land and ocean uncertainties are calculated similarly.

Uncertainty values from products that are not operational are assumed constant for time periods after the end of their record. For the SST uncertainty, the ERSST ensemble was only issued through 2014. Thus, we use the 2014 value for years 2015–2018 and will update the analysis as more data become available. Likewise, the GHCNv4 homogenization was conducted through 2016 resulting in the 2017 and 2018 homogenization uncertainties being set to the 2016 value.

6 Results

6.1 LSAT Uncertainty Results

The sampling and total uncertainty in the global annual land surface mean temperature as calculated by each of the three reanalyses is shown in Figure 3a. As expected, increased number of stations and coverage of stations as time progresses results in decreasing sampling uncertainty over time. The three reanalyses are in general agreement with any differences in the sampling uncertainty shrinking in the total LSAT uncertainty. In the early decades of the study period, sampling uncertainty and homogenization uncertainty are of similar magnitude.

Figure 3b shows the LSAT as found with the ERA5 sampling uncertainty analysis. We will use the ERA5 analysis for the LSAT estimates in the remainder of the study. The homogenization component includes both the parametric uncertainty as well as uncertainties due to missed breaks. Approaching the present, the global sampling uncertainty decreases as the majority of the land has some station coverage, but the global homogenization uncertainty remains high. In particular, the major drop in sampling uncertainty in 1950–1970 occurs due to the inclusion of Antarctica. The relative lack of decrease in

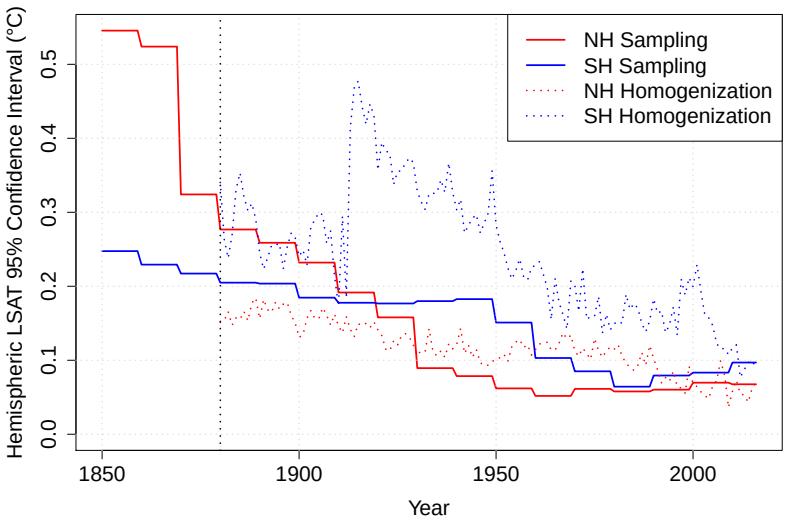


Figure 4. Annual land surface temperature anomaly sampling (solid) and homogenization (dotted) uncertainty (2σ) per hemisphere. As expected, the uncertainty in the southern hemisphere is greater in all decades, but reduces greatly to near the northern hemisphere uncertainty post-1960.

uncertainty in the global mean due to homogenization results from correction uncertainties in station records propagating forward in time [*Menne et al.*, 2018]. The minor contribution of sampling uncertainty to the total modern LSAT uncertainty illustrates how increasing coverage of temperature monitoring will not fix the uncertainty issue in the land surface temperature record.

The ERA5 analysis shows that the uncertainties in *Hansen et al.* [2010] were quite good for the early record, but overestimate the sampling uncertainty post-1950. In particular, we find nearly exact agreement over 1880–1900. The sampling uncertainty analysis also suggests that the GISTEMP annual mean time series may be extended to dates earlier than 1880 as is done in HadCRUT4 and Berkeley Earth, but not without suffering a large increase in sampling uncertainty, particularly if including data prior to 1870.

Separating the land uncertainty by hemisphere, we find that the southern hemisphere has greater sampling uncertainty post-1920 coinciding with improved northern hemispheric coverage of the Arctic land and sea ice region (Figure 4). We again see the effect of Antarctica on the southern hemisphere through the reduction in sampling uncertainty from 1950–1970. The hemispheric homogenization uncertainties are slowly decreasing as is the global mean with the exception of the large jump in southern hemisphere uncertainty in the mid 1920s which can be explained by limited number of stations in the southern hemisphere available for comparison.

We further break down the sampling uncertainty analysis to the GISTEMP band level to determine the latitudinal regions where the station record may be unreliable for the broader region. Figure 5 shows the time series for each of the 8 latitudinal bands used in the GISTEMP analysis. The polar series confirm that these regions are driving the decrease in sampling uncertainty for both hemisphere.

Combining our improved total LSAT uncertainty with the GISTEMP land surface temperature time series gives a intuitive description of the certainty of the land warm-

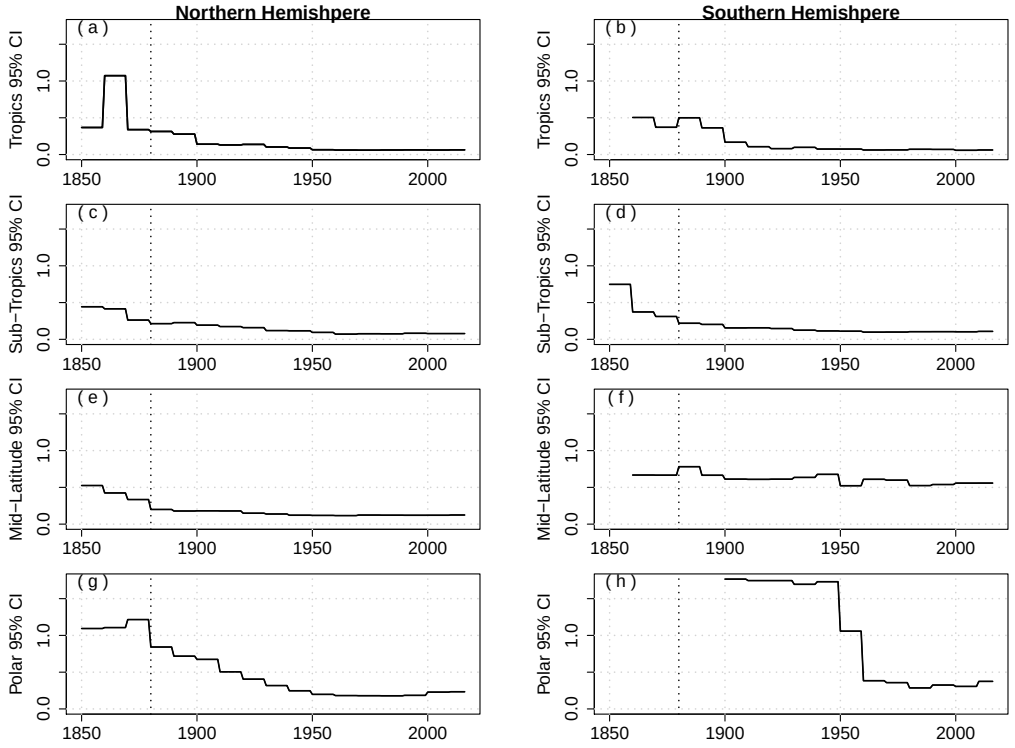


Figure 5. Annual land surface temperature anomaly (°C) uncertainty (2σ) per latitudinal band on the GISTEMP grid. The tropics (a)/(b) are 0–23.6°, the sub-tropics (c)/(d) are 23.6–44.4°, the mid-latitudes (e)/(f) are 44.4–64.2°, and the polar regions (g)/(h) are 64.2–90°. The dotted line marks 1880, the current start date of production GISTEMP.

ing trend over the modern record period. Figure 6 shows the LSAT time series from the operational GISTEMP analysis with confidence intervals according to the sampling and homogenization uncertainties. The magnitude in the trend is many times greater than the uncertainty at any period. Additionally, the uncertainty is much lower in the 1960–present period in which much of the warming has occurred.

6.2 LSAT Extensions Results

6.2.1 Sampling Bias Results

Since the results of the sampling bias assessment were not robust among reanalyses, we present the results for all three reanalyses in Figure 7. In general, the JRA55 and ERA5 products agree, with MERRA being an outlier. We find no evidence for sampling bias for the in-sample 1980–Present time period when using the ERA analysis. We also have the smallest confidence intervals of the three analyses for ERA5 demonstrating that the non-significance of the bias is a robust result.

As mentioned, the major caveat in the bias calculation is that the climate has been highly non-stationary over the past 150 years and we are calculating the bias due to a particular incomplete sampling using the climate changes over the ERA period of 1979–2018. That is, we are determining how good of a job a particular station arrangement could do at observing the climate change that has occurred from 1979–2018; a period in which we believe that the Arctic is warming faster than the rest of the land. In addition, we are making the assumption that the arctic temperature is changing at a fixed multiple of the global

average. This assumption is reasonable as model studies have shown that modeling the amplification trend linearly is a reasonable choice over recent decades *Serreze and Barry* [2011]; *Cohen et al.* [2014].

The large and significant cool biases in the ERA and JRA reanalyses in the early record describe how undersampling the observed 1979–Present temperature change would lead to a biased calculation in the global mean. The approach of the estimates to unbiased mirrors the global coverage shown in Figure 2. The relationship between coverage and bias in estimating the 1979–present warming makes sense, particularly because we know that station coverage in polar regions was limited or nonexistent in the early record and arctic temperature changed more rapidly over the past few decades.

6.2.2 Limiting Uncertainty Results

Running the sampling uncertainty analysis assuming perfect coverage suggests that 0.034 is the limiting sampling 95% confidence interval for the GISTEMP method. In other words, adding additional station observations will not reduce the sampling uncertainty below this level. The current coverage is already quite close to this value as shown in Figure 8 implying that we are close to the limiting coverage for the GISTEMP model. Roughly speaking, the limiting uncertainty decreases with the amount of smoothing in the interpolation. As station coverage continues to improve, the choice of interpolation in GISTEMP should be revisited.

Our limiting sampling bias is found to be significant, albeit small. We find that the GISTEMP procedure overestimates the true global mean LSAT over the ERA5 record by 1.5% with a 95% confidence interval of (1.0%, 2.0%). A small limiting bias again suggests a reduction in the smoothing radius as full coverage is approached. In the context of the results in the previous section, we interpret production GISTEMP as being nearly unbiased, even in the pathological limiting case.

6.2.3 Averaging Method Comparison Results

Figure 8 compares the LSAT sampling uncertainty from the simple latitude-weighted mean and GISTEMP band mean methods. We find that the GISTEMP method almost always outperforms the simple method with the 1890s and 1900s being the only exception. Furthermore, we see that the GISTEMP method outperforms the simple method by up to 50% in the 1930s and 1940s, primarily due to the added arctic coverage providing better NH polar band estimates. The results demonstrate the value added by the GISTEMP averaging scheme by leveraging the zonal correlation of temperature anomalies.

6.3 Ocean

The global uncertainty from the ERSST large ensemble using the GISTEMP averaging scheme resembles the global uncertainty calculated by the ERSST team. Similar uncertainty is expected as the GISTEMP averaging scheme converges to a latitudinal-weighted grid cell average as missing data approaches zero and the ERSST large ensemble has complete coverage of the oceans. The GISTEMP operational global annual average SST time series is shown in Figure 9. As in the LSAT global time series, the magnitude of the warming trend dominates the uncertainty of the calculation.

Looking at the hemispheric uncertainty in the annual SST anomaly, we see that there are minor differences between the two hemispheres (Figure 10). The larger uncertainty in the southern hemisphere post 1945 drives the global uncertainty as the southern hemisphere has double the area occupied by ocean compared to the northern hemisphere.

6.4 Total Global Uncertainty

We are now able to combine our total global uncertainty with the production GISTEMP global annual mean surface temperature anomaly time series. Figure 11 shows the production GISTEMP global time series with the 95% confidence interval calculated in this study. The confidence interval has been added to the distributed GISTEMP time series facilitating uncertainty quantification in studies that utilize the GISTEMP product. As in both the SST and LSAT time series, the warming signal is greater than the underlying uncertainty. We investigate the possible uncertainty of the signal in the following section.

As in the land and ocean analyses, we decompose the global uncertainty into Northern and Southern hemisphere uncertainties (Figure 12). Following the larger land uncertainty and comparable ocean uncertainty, we see that the total uncertainty on the annual hemispheric mean is almost always larger in the Southern hemisphere.

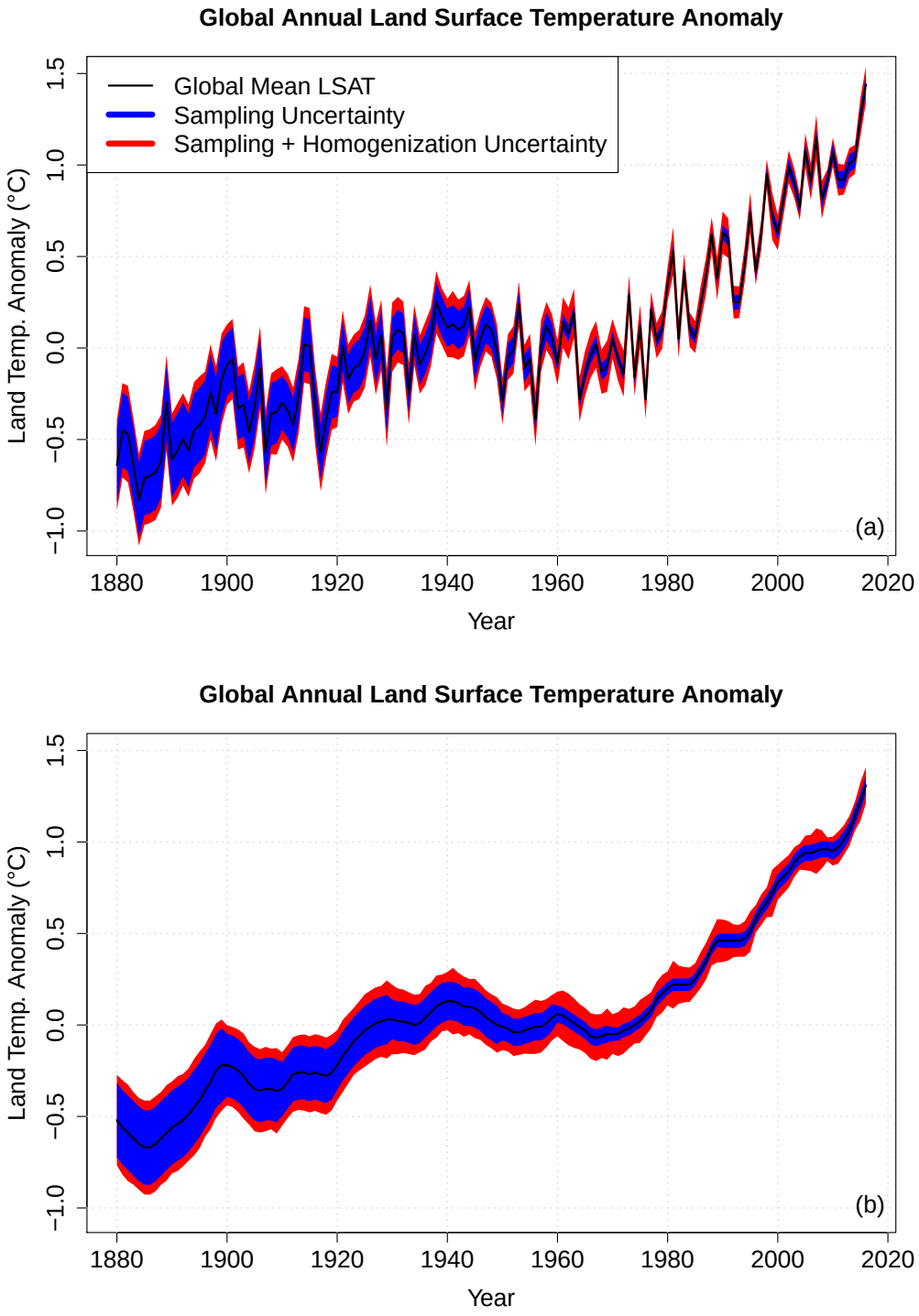


Figure 6. The GISTEMP land-only mean with 95% confidence intervals for (a) annual mean and (b) annual mean smoothed by LOWESS with 5-year bandwidth. For both plots, the envelopes show the annual uncertainty of the sampling uncertainty alone as well as the total uncertainty when including the homogenization. Anomalies are calculated with respect to the 1951-1980 climatology. We include the annual uncertainty on the 5-year smoothed series to illustrate that the trend has much larger magnitude than the uncertainty.

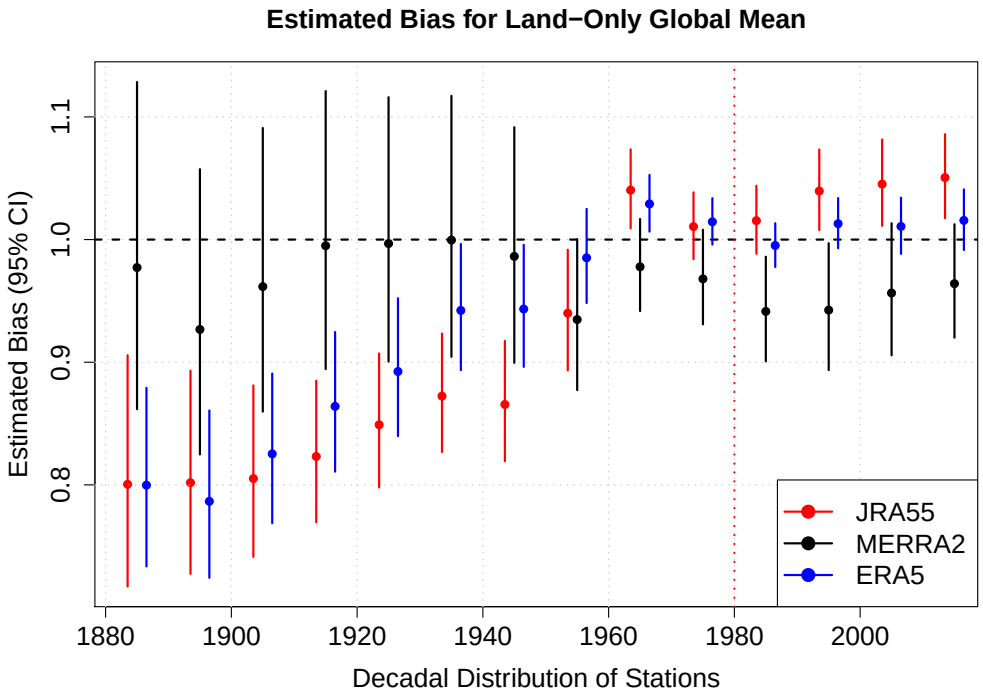


Figure 7. Estimates of the scaling bias on the global mean anomaly due to the decadal incomplete sampling in the LSAT for each of the three reanalyses. The line at 1.0 signifies an unbiased estimate and confidence intervals larger or smaller than this value signify statistically significant bias. The red line signifies the start date of the products; decades after this point can be interpreted as a measure of the bias in the global mean of GISTEMP.

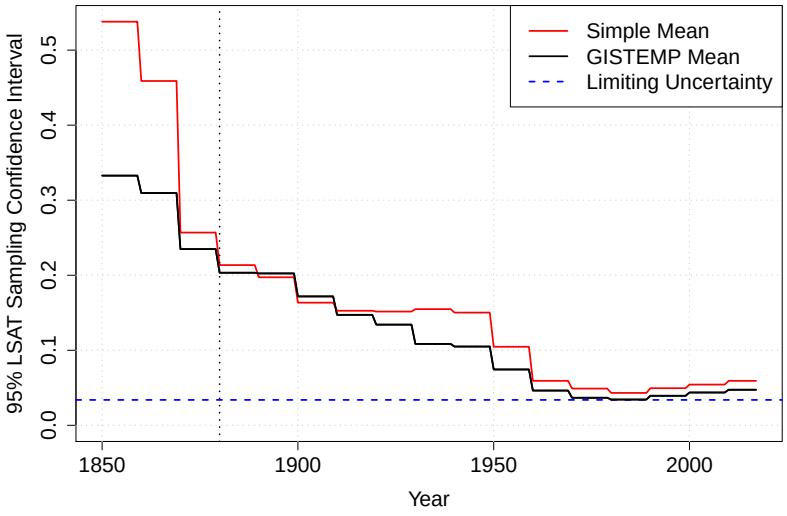


Figure 8. A comparison of the sampling uncertainty in the global land-only annual mean temperature anomaly when using the GISTEMP averaging scheme and a simple cosine-weighted mean. The limiting mean is sampling uncertainty found in the ERA5 sampling analysis assuming that there is a station at every grid point and represented the uncertainty introduced into the estimate by the interpolation.

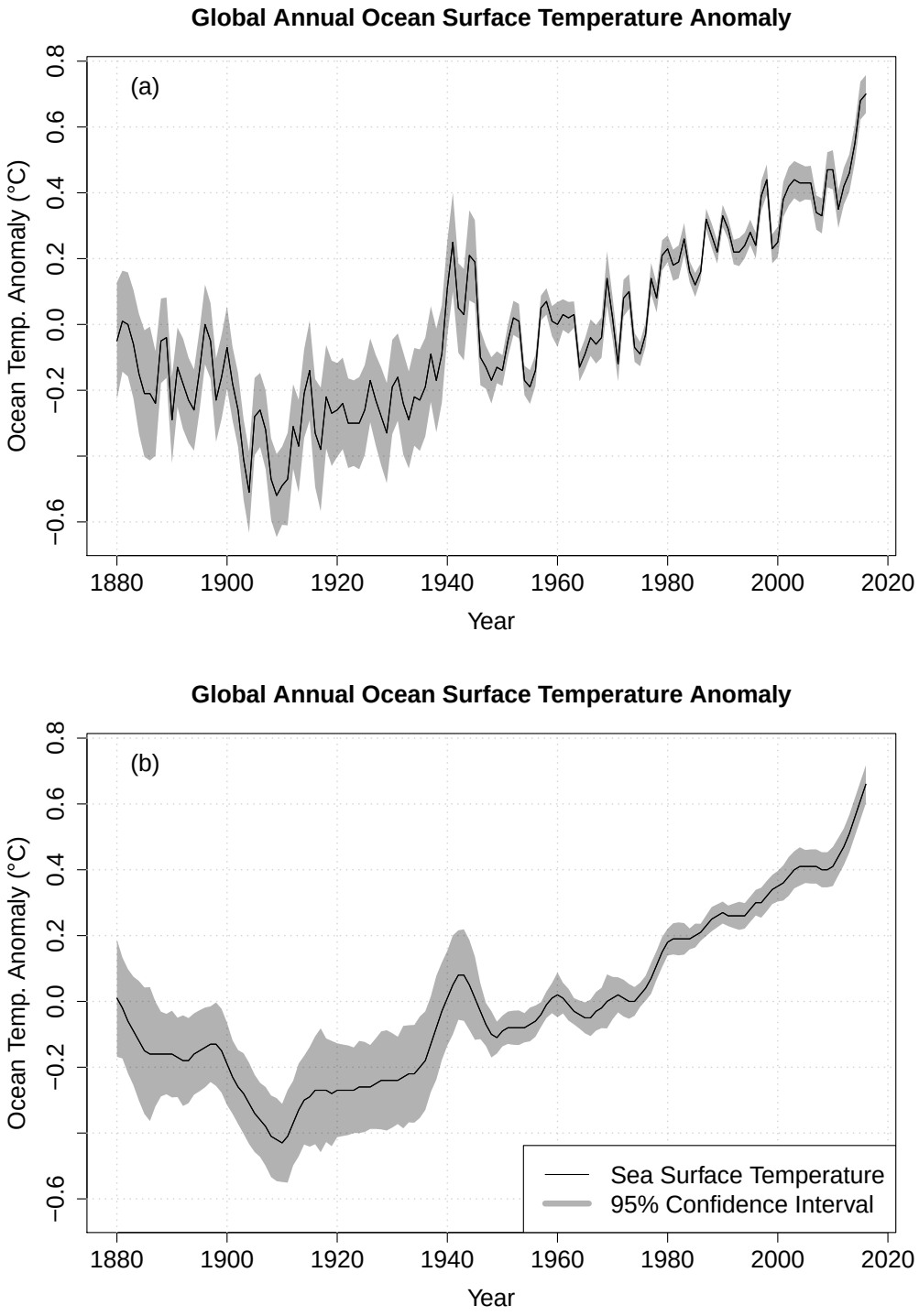
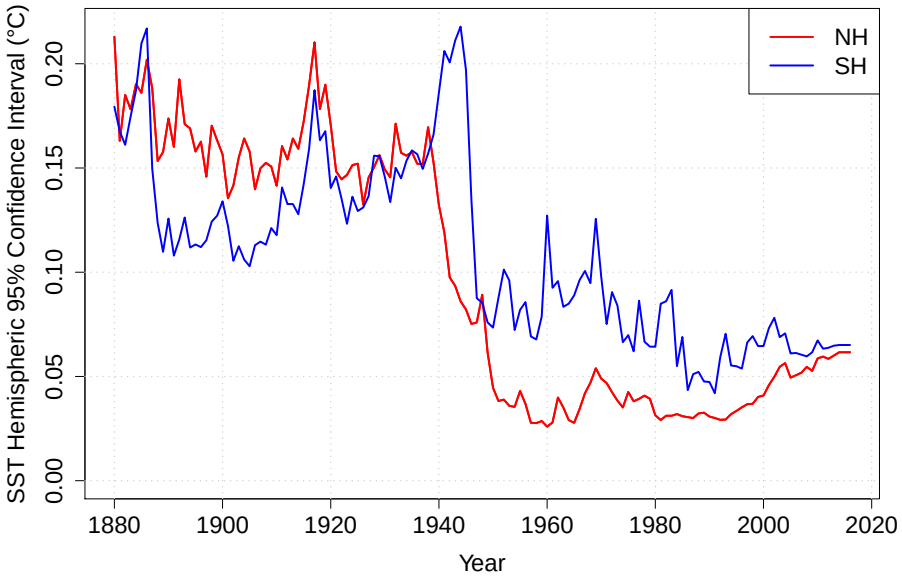


Figure 9. The GISTEMP ocean-only mean with 95% confidence intervals for (a) annual mean and (b) annual mean smoothed by LOWESS with 5-year bandwidth. The envelopes show the annual SST parametric uncertainty as calculated from the ERSSTv4 large ensemble. Anomalies are calculated with respect to the 1951-1980 climatology. We include the annual uncertainty on the 5-year smoothed series to illustrate that the trend has much larger magnitude than the uncertainty.



683 **Figure 10.** Annual sea surface temperature anomaly parametric uncertainty (2σ) per hemisphere calculated
684 using the ERSSTv4 large ensemble with the GISTEMP averaging scheme.

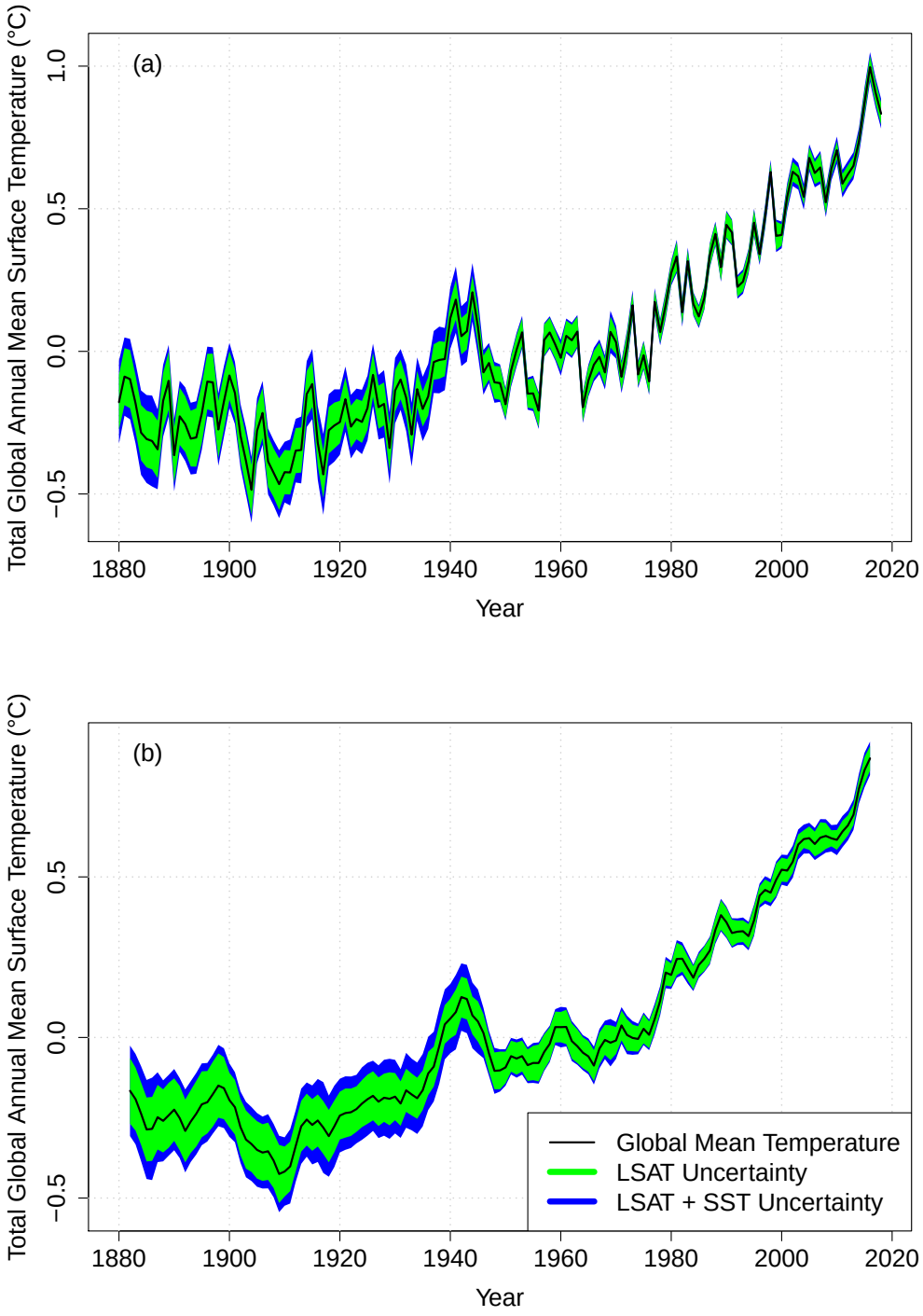
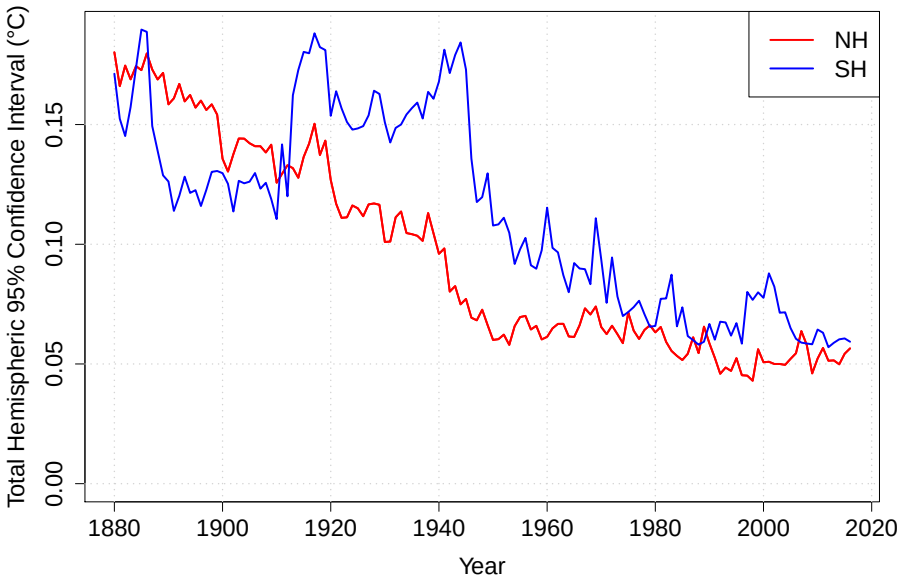
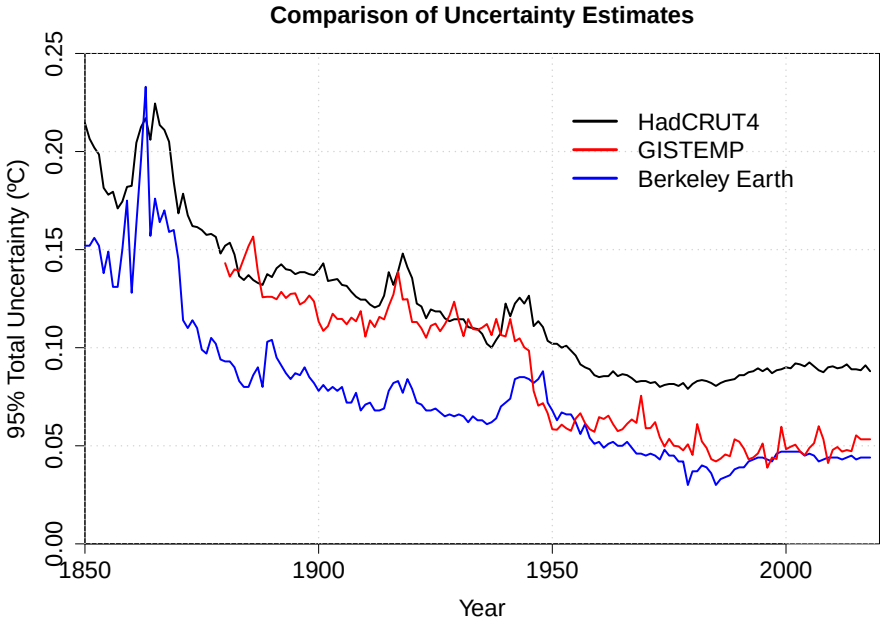


Figure 11. The production GISTEMP global mean temperature time series with the total (LSAT and SST) 95% confidence interval calculated in our study for (a) annual mean temperature and (b) annual mean temperature smoothed with LOWESS with 5-year bandwidth. Anomalies are calculated with respect to the 1951-1980 climatology. We include the annual uncertainty on the 5-year smoothed series to illustrate that the trend has much larger magnitude than the uncertainty.



713 **Figure 12.** Annual mean temperature anomaly total uncertainty (2σ) per hemisphere.



714 **Figure 13.** Comparison of total uncertainty (95% confidence interval) in three independent global analyses,
715 HadCRUT4, GISTEMP (this paper), and Berkeley Earth.

7 Discussion

Since the first GISTEMP estimates in the 1980s, there have been large increases in the amount of data ingested, improvements in the homogenization of station data to remove non-climatic effects, and the incorporation of ocean data, but not much change to the global mean calculation methodology. These data changes have produced variations over time of the global annual mean record that, while not a controlled exploration, are indicative of the structural uncertainties in the product that arise indirectly through changes in data availability and processing. The new analysis presented here is far more complete, but it is appropriate that recent versions of GISTEMP fall within the uncertainties shown in Figure 11.

The improved assessment of uncertainty in the GISTEMP product is a function of three new developments: the Monte Carlo ensembles that have been done for the input data (ERSST and GHCN), the upgrading of the GISTEMP code base, and the evolving standards in uncertainty quantification in climate science. These threads have made the current study far more tractable than it would have been a decade ago.

The existence of the new uncertainty product now allows us to be more rigorous in assessing the strength of claims of records and trends in the data itself, but also to improve the propagation of that uncertainty into, for instance, detection and attribution exercises for constraining anthropogenic climate change.

One persistent question is whether it makes sense to extend the GISTEMP product prior to 1880, to perhaps as early as 1850 (for instance, to help estimate a 19th Century baseline climatology [*Hawkins et al.*, 2017]). Figure 4.2 demonstrates that the sampling in the 1870s is not that much worse than the 1880s, but unfortunately, the homogenization analysis does not extend before 1880, and nor does the ERSST data. This is an issue we will continue to explore.

7.1 Probability of a new warmest year record

The addition of the global annual mean uncertainty values calculated in this study to the widely distributed GISTEMP surface temperature product will enable users to include more informed probabilistic statements of uncertainty in their research. One such example is the probability of warmest year calculation which is often cited in scientific and popular literature.

Given the strong trend in global mean temperatures since the 1970s, NASA/GISS has frequently reported on new records for annual means over the instrumental period (11 times since 1988). This naturally leads to the question of how confident we can be in declaring that any particular record year in the GISTEMP index, was in fact the warmest year in the real world since 1880. Discussion of this uncertainty has been a focus of the NOAA and NASA annual briefings since 2014, which at the time was the warmest year in the record [*NASA Public Affairs*, 2015]. With the major El Niño event in 2015/2016, both subsequent years were notably warmer [*NASA Public Affairs*, 2016, 2017], but how certain can we be of that?

We make a Monte Carlo estimate of the warmest year by determining which year has the highest temperature anomaly after either independent or autoregressive simulations of the uncertainties. The probability that a given year was the warmest year on record to date is then the number of simulations in which it is the warmest year divided by the total number of simulations. We use this method to reassess how well NASA's recent statements on the probability of warmest years match up to our updated uncertainty calculations.

In January 2015, NASA reported that 2014 was likely the warmest year with 38% likelihood [*NASA Public Affairs*, 2015] based on a simple assumption of linearly increasing

uncertainty based on the *Hansen et al.* [2010] estimates. We now find that this was conservative, and that 2014 actually had a 79% chance of truly being the warmest year in the instrumental period. Assuming autocorrelated uncertainties, this reduces slightly to 75% since the next most probable warmest years were non-consecutive (2010 and 2005). The following year, NASA reported a likelihood that 2015 was the new record warmest year was 96%, which compares to a 99.99% probability calculated now (regardless of whether we use independent or autocorrelated uncertainties).

Assuming that uncertainties in the annual mean are independent from year to year, we find that 2016 is likely the warmest year in the last 139 (1880–2018) years with 86.2% certainty. The other years that could plausibly have been the warmest were 2017 (12.5 % probability), 2018 (1.2% probability), and 2015 (<0.1% probability). While the GISTEMP-estimated mean global temperature is larger in 2015 than in 2018, the uncertainty in the 2018 mean is larger, primarily due to an increase in the LSAT homogenization uncertainty. Therefore, 2015 will rank higher on the warmest years than 2018 on average, but the additional uncertainly in the 2018 mean gives it a greater chance of being the warmest year.

We can also calculate this probability using autoregressive uncertainties. Unlike the uncertainty in temperature change, autoregressive uncertainties give more certainty to 2016 being the warmest year with a simulated 87.2% certainty. Since all of the candidate years have occurred in consecutive years, positive autocorrelation reduces expected difference in uncertainty.

While the AR(1) calculation is a reasonable choice for comparing anomalies over a short time period, such a calculation is not statistically sound for longer-term analyses using the uncertainties calculated in our study. Components of the uncertainty, particularly the homogenization uncertainty, persist over many decades reflecting large shifts in the record that propagate in time. These types of uncertainties are best represented in an uncertainty ensemble which has not yet been created for GISTEMP.

7.2 Comparison to other uncertainty estimates

Two of the other products shown in Fig. 1 have independently derived total uncertainties, specifically HadCRUT4 and BEST. Figure 13 shows the comparison of the three 95% confidence intervals. The overall magnitudes are similar, with close agreement with the HadCRUT4 uncertainty pre-1945 and with BEST post-1945. The character of the change around 1945 is driven primarily by the reduction in SST uncertainty in ERSST and reduction in the greater reduction in GISTEMP LSAT sampling uncertainty relative to HadCRUT4.

8 Conclusion

Our new uncertainty quantification of the global annual mean surface temperature anomaly in the GISTEMP product brings this analysis up to the enhanced standards of its peers and we hope that this will aid the interpretation and utility of this widely used product. This paper has focused on the global and hemispheric annual means, but the procedure can equally be used to improve the uncertainty analysis of regional and monthly data products and these will be pursued in further work.

Acknowledgements

We thank both the reviewers and editor for the insightful comments. Their suggested extensions led to interesting discoveries and a much stronger analysis and paper. The GISTEMP analysis is funded from grants from the NASA Modeling, Analysis and Prediction program in the Science Mission Directorate. N.L. was also supported by the

National Science Foundation Graduate Research Fellowship under Grant No. NSF DGE 16-44869. Datasets from GHCN and ERSST are supported by NOAA's National Centers for Environmental Information. The Antarctic READER data are supported by SCAR. Thanks to Boyin Huang (NOAA) for access to the ERSST large parameter ensemble. Special thanks to the ClearClimateCode project, Nick Barnes and David R. Jones, for converting the original GISTEMP codebase to Python. The analysis was performed in the open source language R [*R Core Team*, 2016] and the data, code, and intermediate steps needed to generate all figures in this report are available on the GISTEMP website (`https://data.giss.nasa.gov/gistemp/uncertainty`).

References

Barnes, N., and D. Jones (2011), Clear Climate Code: Rewriting legacy science software for clarity, *IEEE Software*, *28*(6), 36–42, doi:10.1109/ms.2011.113.

Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones (2006), Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850, *Journal of Geophysical Research*, *111*(D12), D12,106, doi:10.1029/2005JD006548.

Callendar, G. S. (1938), The artificial production of carbon dioxide and its influence on temperature, *Quarterly Journal of the Royal Meteorological Society*, *64*(275), 223–240, doi:10.1002/qj.49706427503.

Callendar, G. S. (1961), Temperature fluctuations and trends over the earth, *Quarterly Journal of the Royal Meteorological Society*, *87*(371), 1–12, doi:10.1002/qj.49708737102.

Cohen, J., J. A. Screen, J. C. Furtado, M. Barlow, D. Whittleston, D. Coumou, J. Francis, K. Dethloff, D. Entekhabi, J. Overland, and J. Jones (2014), Recent arctic amplification and extreme mid-latitude weather, *Nature Geoscience*, *7*, 627 EP –.

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), `https://cds.climate.copernicus.eu/cdsapp#!/home`.

Cowtan, K., and R. G. Way (2014), Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends, *Quarterly Journal of the Royal Meteorological Society*, *140*(683), 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, *Geophysical Research Letters*, *42*(15), 6526–6534, doi:10.1002/2015gl064888.

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 (2016), ICOADS release 3.0: a major update to the historical marine climate record, *International Journal of Climatology*, *37*(5), 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.-K. 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), *Journal of Climate*, *30*(14), 5419–5454, doi:10.1175/JCLI-D-16-0758.1.

Hansen, J., and S. Lebedeff (1987), Global trends of measured surface air temperature, *Journal of Geophysical Research: Atmospheres*, *92*(D11), 13,345–13,372, doi:10.1029/JD092iD11p13345.

Hansen, J., D. Johnson, A. Lacis, S. Lebedeff, P. Lee, D. Rind, and G. Russell (1981), Climate impact of increasing atmospheric carbon dioxide, *Science*, *213*(4511), 957–966, doi:10.1126/science.213.4511.957.

Hansen, J., R. Ruedy, J. Glascoe, and M. Sato (1999), GISS analysis of surface temperature change, *Journal of Geophysical Research: Atmospheres*, *104*(D24), 30,997–31,022, doi:10.1029/1999JD900835.

Hansen, J., M. Sato, R. Ruedy, P. Kharecha, A. Lacis, R. Miller, L. Nazarenko, K. Lo, G. A. Schmidt, G. Russell, I. Aleinov, S. Bauer, E. Baum, B. Cairns, V. Canuto, M. Chandler, Y. Cheng, A. Cohen, A. D. Genio, G. Faluvegi, E. Fleming, A. Friend, T. Hall, C. Jackman, J. Jonas, M. Kelley, N. Y. Kiang, D. Koch, G. Labow, J. Lerner, S. Menon, T. Novakov, V. Oinas, J. Perlwitz, J. Perlwitz, D. Rind, A. Romanou, R. Schmunk, D. Shindell, P. Stone, S. Sun, D. Streets, N. Tausnev, D. Thresher, N. Unger, M. Yao, and S. Zhang (2007), Climate simulations for 1880–2003 with GISS modelE, *Climate Dynamics*, *29*(7-8), 661–696, doi:10.1007/s00382-007-0255-8.

Hansen, J., R. Ruedy, M. Sato, and K. Lo (2010), Global surface temperature change, *Reviews of Geophysics*, *48*(4), doi:10.1029/2010RG000345, rG4004.

Hawkins, E., and P. D. Jones (2013), On increasing global temperatures: 75 years after Callendar, *Quarterly Journal of the Royal Meteorological Society*, *139*(677), 1961–1963, doi:10.1002/qj.2178.

Hawkins, E., P. Ortega, E. Suckling, A. Schurer, G. Hegerl, P. Jones, M. Joshi, T. J. Osborn, V. Masson-Delmotte, J. Mignot, P. Thorne, and G. J. van Oldenborgh (2017), Estimating changes in global temperature since the preindustrial period, *Bulletin of the American Meteorological Society*, *98*(9), 1841–1856, doi:10.1175/BAMS-D-16-0007.1.

Huang, B., V. F. Banzon, E. Freeman, J. Lawrimore, W. Liu, T. C. Peterson, T. M. Smith, P. W. Thorne, S. D. Woodruff, and H.-M. Zhang (2015), Extended reconstructed sea surface temperature version 4 (ERSSTv4). Part I: Upgrades and intercomparisons, *Journal of Climate*, *28*(3), 911–930, doi:10.1175/JCLI-D-14-00006.1.

Huang, B., P. W. Thorne, T. M. Smith, W. Liu, J. Lawrimore, V. F. Banzon, H.-M. Zhang, T. C. Peterson, and M. Menne (2016), Further exploring and quantifying uncertainties for extended reconstructed sea surface temperature (ERSST) version 4 (v4), *Journal of Climate*, *29*(9), 3119–3142, 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.-M. Zhang (2017), Extended reconstructed sea surface temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons, *Journal of Climate*, *30*(20), 8179–8205, doi:10.1175/JCLI-D-16-0836.1.

Ishihara, K. (2006), Calculation of global surface temperature anomalies with COBE-SST, *Weather Service Bulletin*, *73*, S19–S25.

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

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.

Kennedy, J. J., N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby (2011a), Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties, *Journal of Geophysical Research*, *116*(D14), D14,103, doi:10.1029/2010JD015218.

Kennedy, J. J., N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby (2011b), Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization, *Journal of Geophysical Research*, *116*(D14), D14,104, doi:10.1029/2010JD015220.

Kobayashi, S., Y. Ota, Y. Harada, A. Ebita, M. Moriya, H. Onoda, K. Onogi, H. Kamahori, C. Kobayashi, H. Endo, K. Miyaoka, and K. Takahashi (2015), The JRA-55 reanalysis: General specifications and basic characteristics, *J. Meteorol. Soc. Japan. Ser. II*, *93*(1), 5–48, doi:10.2151/jmsj.2015-001.

Liu, W., B. Huang, P. W. Thorne, V. F. Banzon, H.-M. Zhang, E. Freeman, J. Lawrimore, T. C. Peterson, T. M. Smith, S. D. Woodruff, W. Liu, B. Huang, P. W. Thorne, V. F. Banzon, H.-M. Zhang, E. Freeman, J. Lawrimore, T. C. Peterson, T. M. Smith, and S. D. Woodruff (2015), Extended Reconstructed Sea Surface Temperature Version 4 (ERSST.v4): Part II. Parametric and Structural Uncertainty Estimations, *Journal of Climate*, *28*(3), 931–951, doi:10.1175/JCLI-D-14-00007.1.

McCarty, W., L. Coy, R. Gelaro, A. Huang, D. Merkova, E. B. Smith, M. Sienkiewicz, and K. Wargan (2016), Merra-2 input observations: Summary and assessment, *NASA Technical Report Series on Global Modeling and Data Assimilation*, *46*.

Menne, M. J., and C. N. Williams (2009), Homogenization of temperature series via pairwise comparisons, *Journal of Climate*, *22*(7), 1700–1717, doi:10.1175/2008jcli2263.1.

Menne, M. J., C. N. Williams, and R. S. Vose (2009), The U.S. Historical Climatology Network monthly temperature data, version 2, *Bulletin of the American Meteorological*

Society, *90*(7), 993–1008, doi:10.1175/2008bams2613.1.

Menne, M. J., C. N. Williams, and M. A. Palecki (2010), On the reliability of the U.S. surface temperature record, *Journal of Geophysical Research*, *115*(D11), doi:10.1029/2009jd013094.

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, *Journal of Climate*, doi:10.1175/jcli-d-18-0094.1.

Mitchell, J. M. (1961), Recent secular changes of global temperature, *Annals of the New York Academy of Sciences*, *95*(1), 235–250, doi:10.1111/j.1749-6632.1961.tb50036.x.

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 data set, *Journal of Geophysical Research: Atmospheres*, *117*(D8), doi:10.1029/2011JD017187.

NASA Public Affairs (2015), NASA, NOAA find 2014 warmest year in modern record, `https://www.giss.nasa.gov/research/news/20150116/`, last-accessed July 9, 2018.

NASA Public Affairs (2016), NASA, NOAA analyses reveal record-shattering global warm temperatures in 2015, `https://www.giss.nasa.gov/research/news/20160120/`, last-accessed July 9, 2018.

NASA Public Affairs (2017), NASA, NOAA data show 2016 warmest year on record globally, `https://www.giss.nasa.gov/research/news/20170118/`, last-accessed July 9, 2018.

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

R Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rao, Y., S. Liang, and Y. Yu (2018), Land surface air temperature data are considerably different among BEST-LAND, CRU-TEM4v, NASA-GISS, and NOAA-NCEI, *J. Geophys. Res. Atmos.*, *123*(11), 5881–5900, doi:10.1029/2018jd028355.

Rennie, J. J., J. H. Lawrimore, B. E. Gleason, P. W. Thorne, C. P. Morice, M. J. Menne, C. N. Williams, 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. L. 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, *Geoscience Data Journal*, *1*(2), 75–102, doi:10.1002/gdj3.8.

Rohde, R., R. 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, *Geoinformatics & Geostatistics: An Overview*, doi:10.4172/2327-4581.1000101.

Rohde, R., R. Muller, R. Jacobsen, S. Perlmutter, A. Rosenfeld, J. Wurtele, J. Curry, C. Wickham, and S. Mosher (2013b), Berkeley Earth temperature averaging process, *Geoinformatics & Geostatistics: An Overview*, *1*, 20–100, doi:10.4172/2327-4581.1000103.

Serreze, M. C., and R. G. Barry (2011), Processes and impacts of arctic amplification: A research synthesis, *Global and Planetary Change*, *77*(1), 85 – 96, doi:https://doi.org/10.1016/j.gloplacha.2011.03.004.

Simmons, A. J., K. M. Willett, P. D. Jones, P. W. Thorne, and D. P. Dee (2010), Low-frequency variations in surface atmospheric humidity, temperature, and precipitation: Inferences from reanalyses and monthly gridded observational data sets, *Journal of Geophysical Research*, *115*(D1), doi:10.1029/2009jd012442.

Simmons, A. J., P. Berrisford, D. P. Dee, H. Hersbach, S. Hirahara, and J.-N. Thépaut (2016), A reassessment of temperature variations and trends from global reanalyses and

monthly surface climatological datasets, *Quarterly Journal of the Royal Meteorological Society*, *143*(702), 101–119, doi:10.1002/qj.2949.

Smith, T. M., and R. W. Reynolds (2005), A global merged land–air–sea surface temperature reconstruction based on historical observations (1880–1997), *Journal of Climate*, *18*(12), 2021–2036, doi:10.1175/JCLI3362.1.

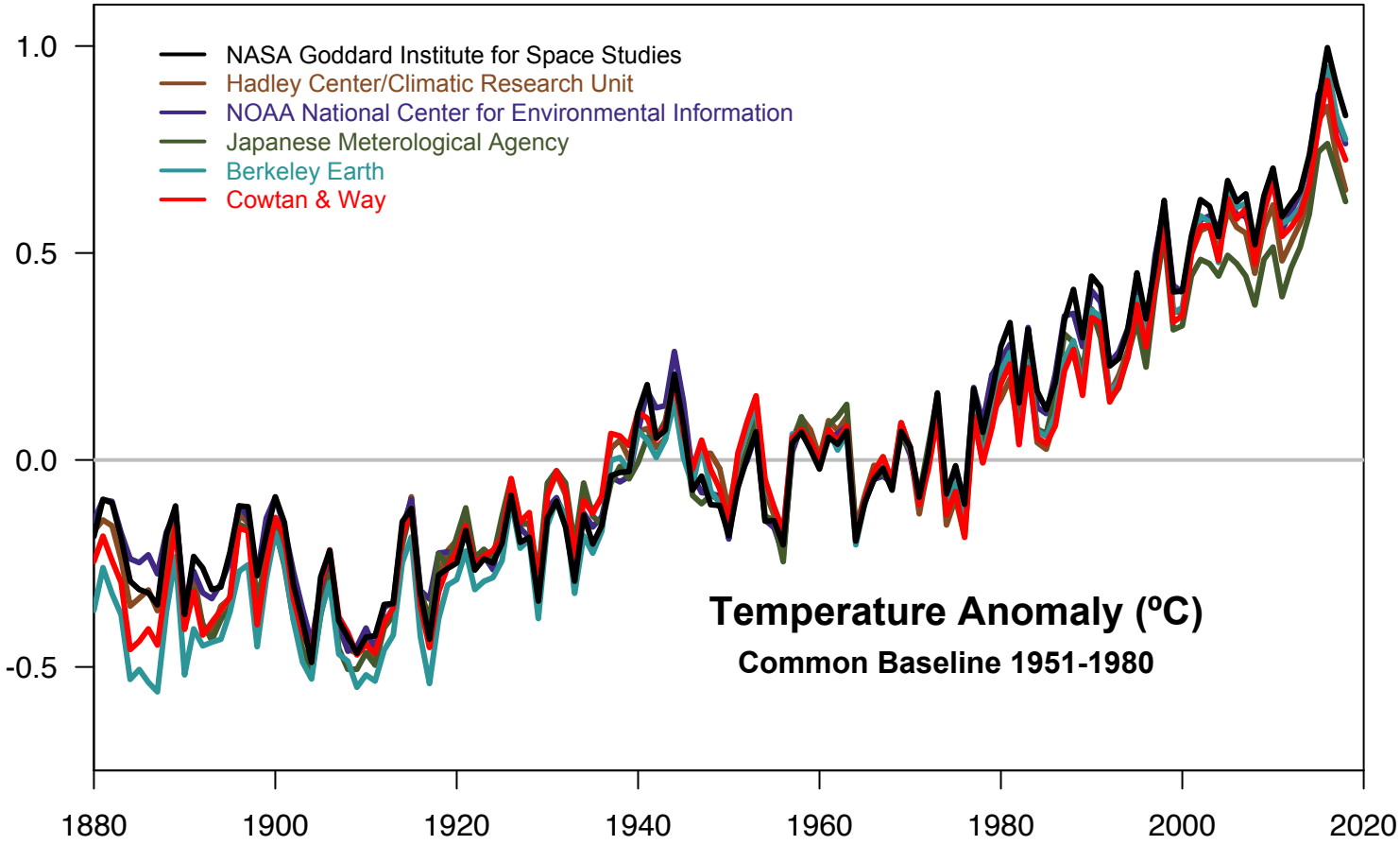
Susskind, J., G. A. Schmidt, J. N. Lee, and L. Iredell (2019), Recent global warming as confirmed by AIRS, *Environmental Research Letters*, *14*(4), 044030, doi:10.1088/1748-9326/aafd4e.

Thorne, P. W., H. J. Diamond, B. Goodison, S. Harrigan, Z. Hausfather, N. B. Ingleby, P. D. Jones, J. H. Lawrimore, D. H. Lister, A. Merlone, T. Oakley, M. Palecki, T. C. Peterson, M. de Podesta, C. Tassone, V. Venema, and K. M. Willett (2018), Towards a global land surface climate fiducial reference measurements network, *International Journal of Climatology*, *38*(6), 2760–2774, doi:10.1002/joc.5458.

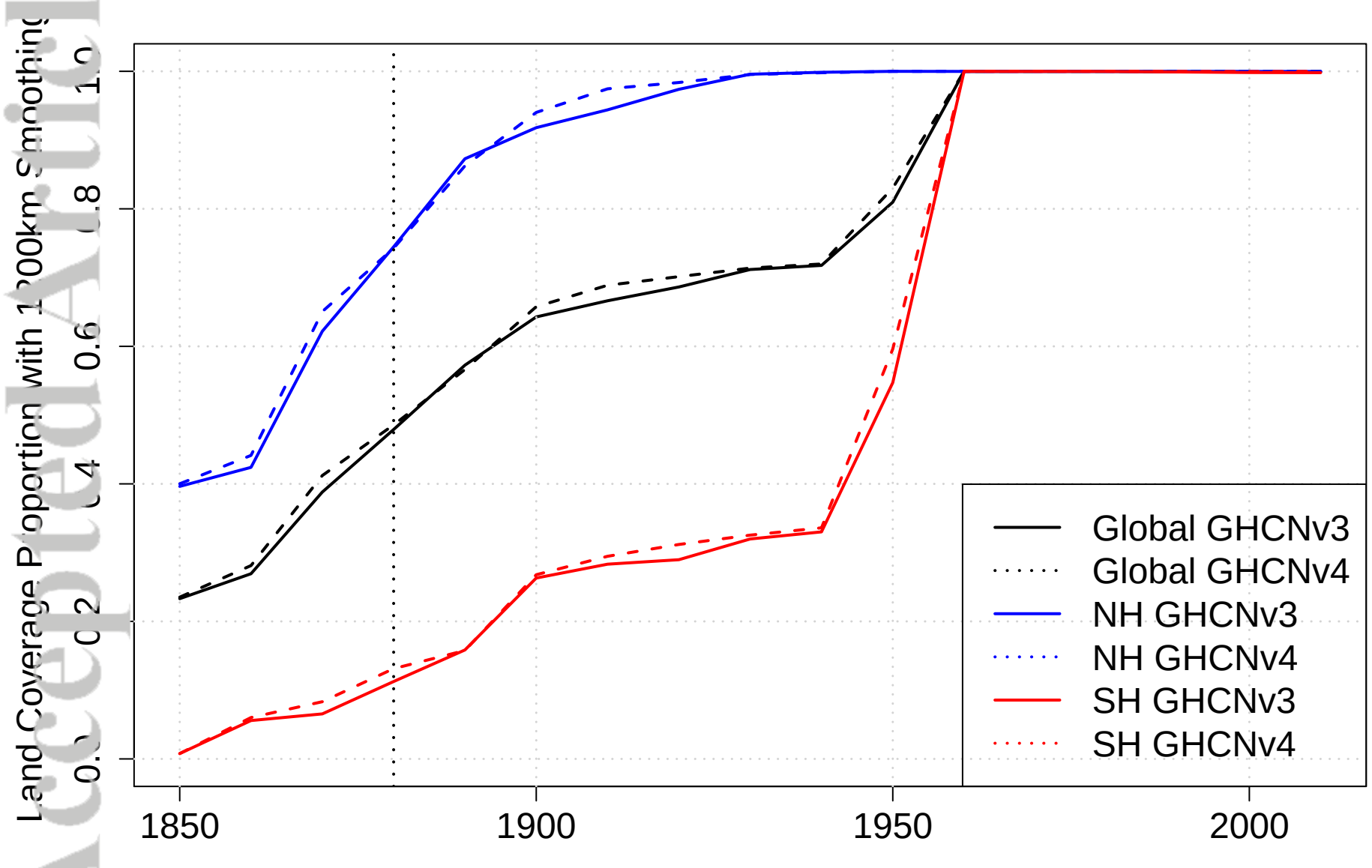
Vose, R. S., D. Arndt, V. F. Banzon, D. R. Easterling, B. Gleason, B. Huang, E. Kearns, J. H. Lawrimore, M. J. Menne, T. C. Peterson, R. W. Reynolds, T. M. Smith, C. N. W. Jr., and D. B. Wuertz (2012), Noaa's merged land–ocean surface temperature analysis, *Bulletin of the American Meteorological Society*, *93*(11), 1677–1685, doi:10.1175/BAMS-D-11-00241.1.

Williams, C. N., M. J. Menne, and P. W. Thorne (2012), Benchmarking the performance of pairwise homogenization of surface temperatures in the United States, *Journal of Geophysical Research: Atmospheres*, *117*(D5), D05116, doi:10.1029/2011jd016761.

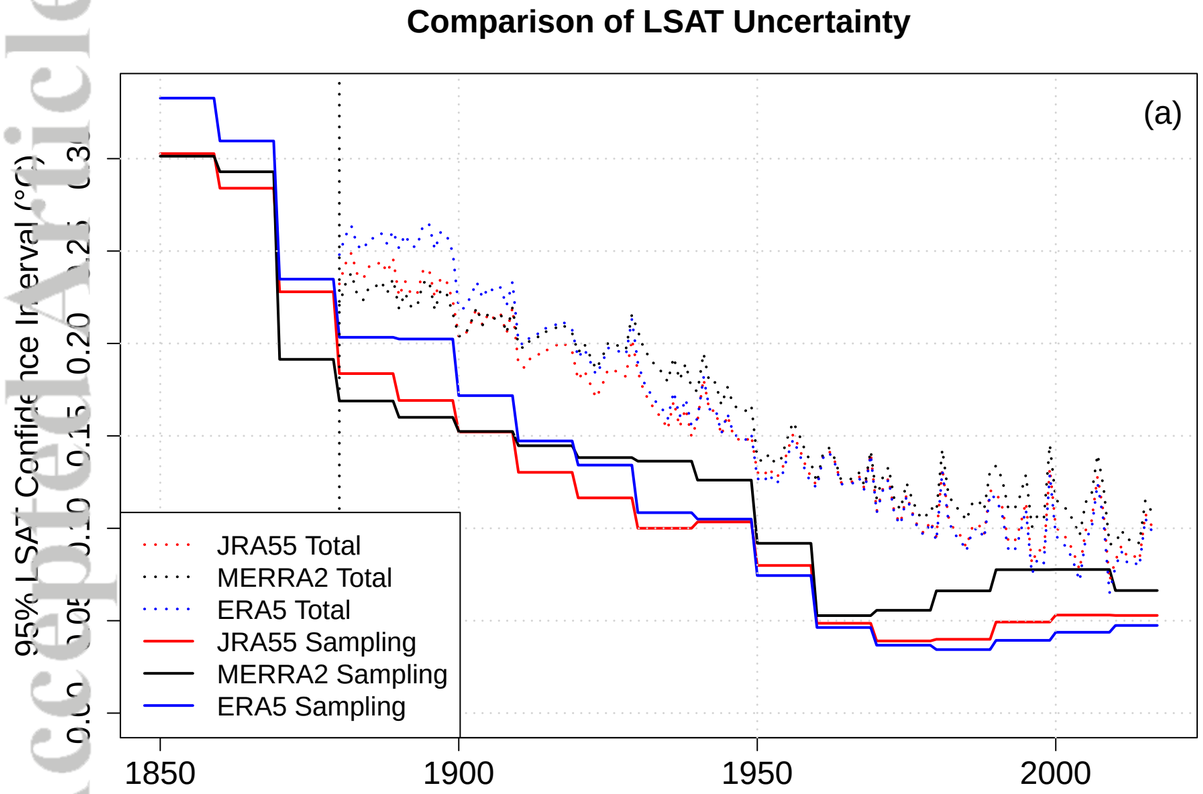
Revision Figure 1.

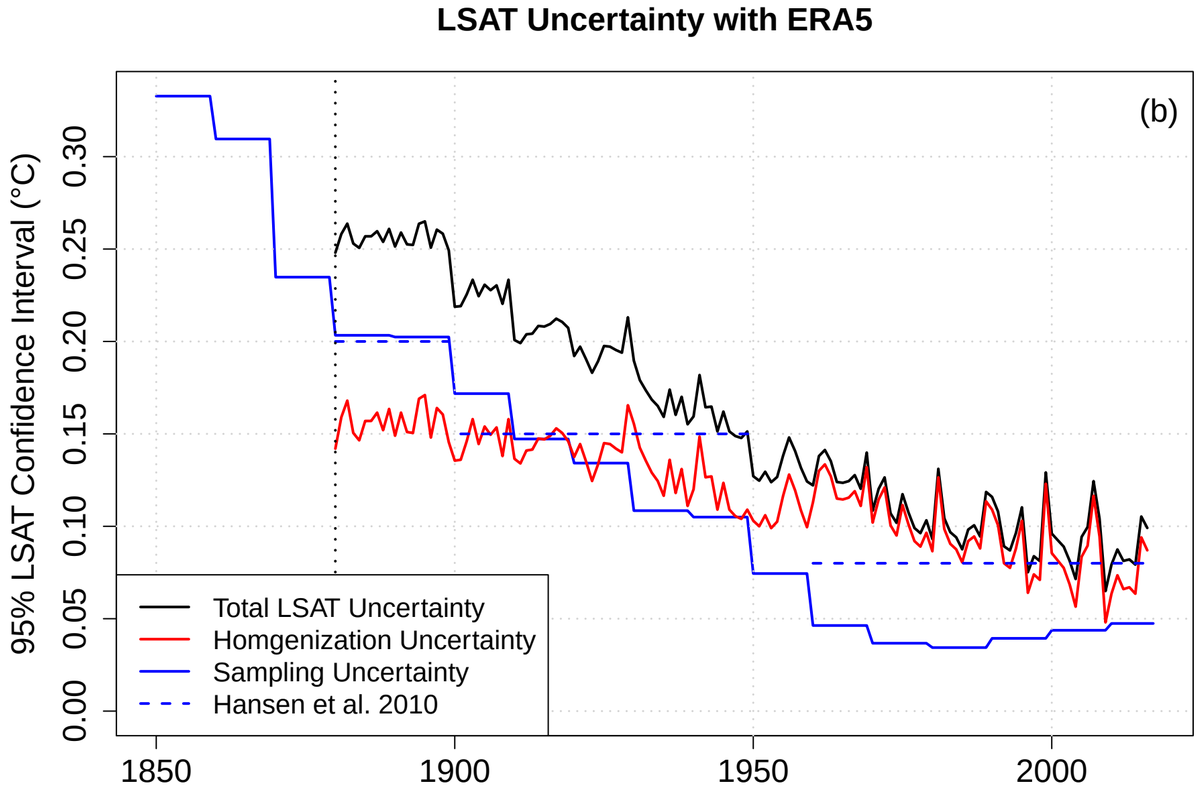


Revision Figure 2.

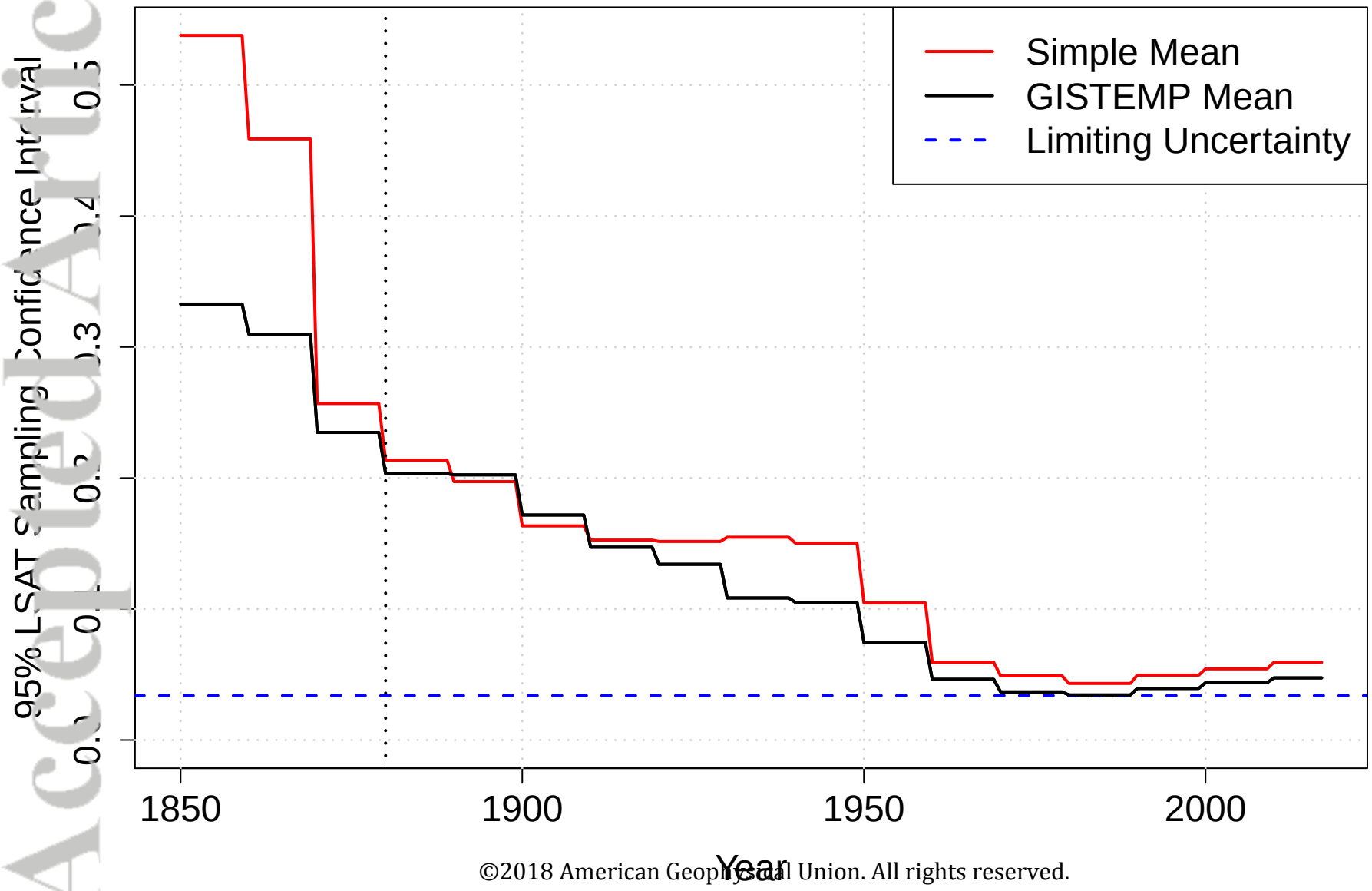


Revision Figure 3.

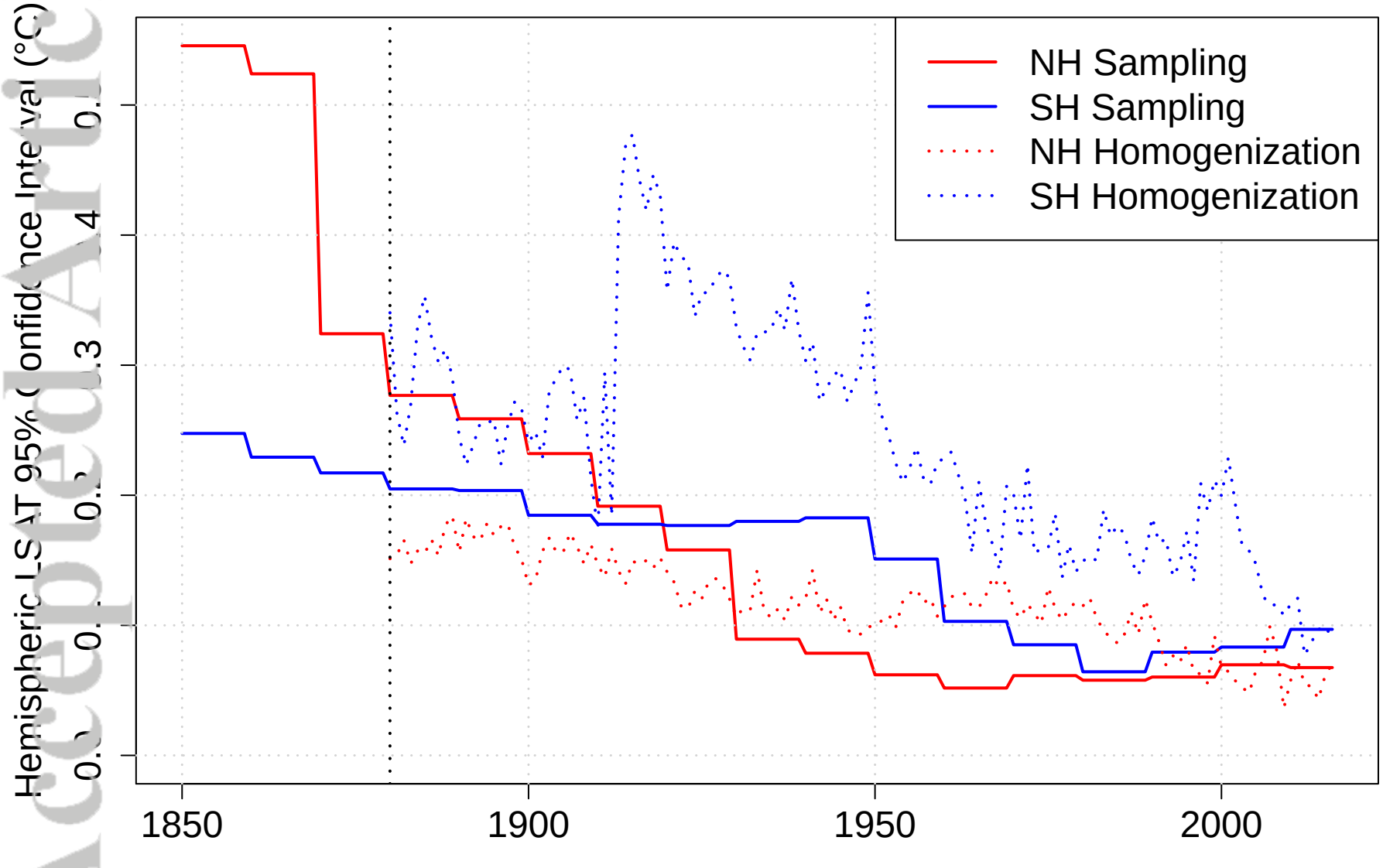




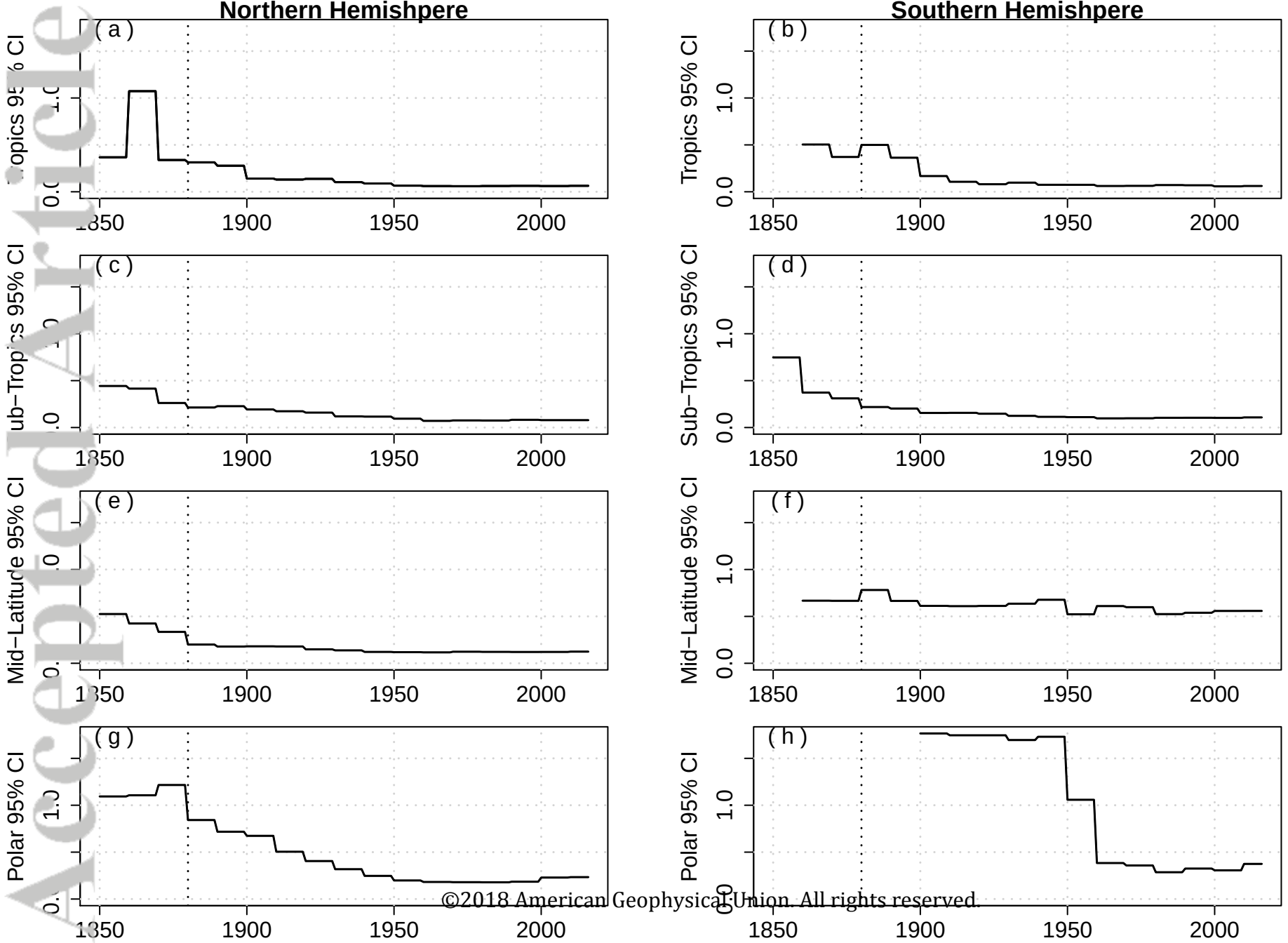
Revision Figure 4.



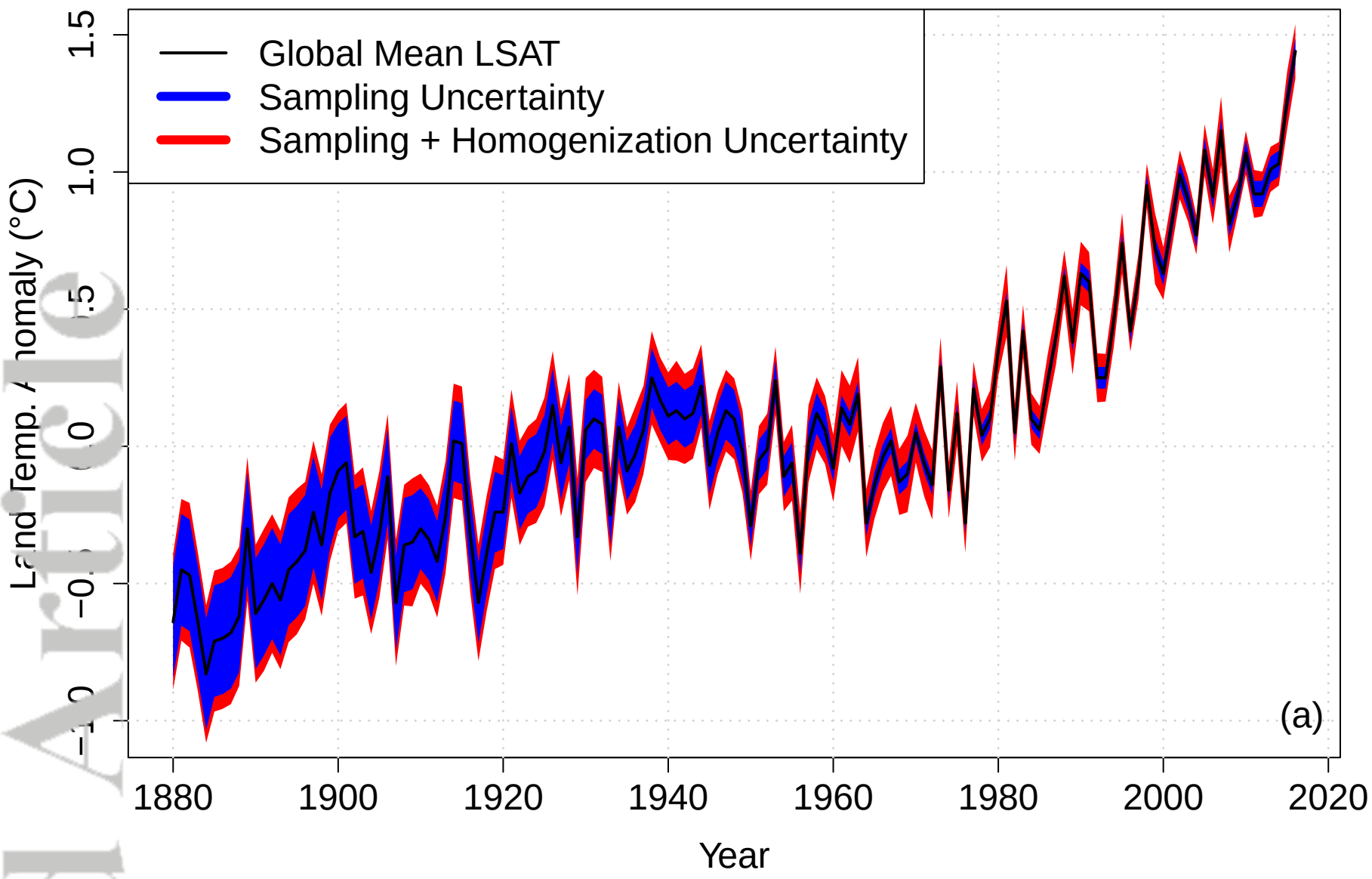
Revision Figure 5.

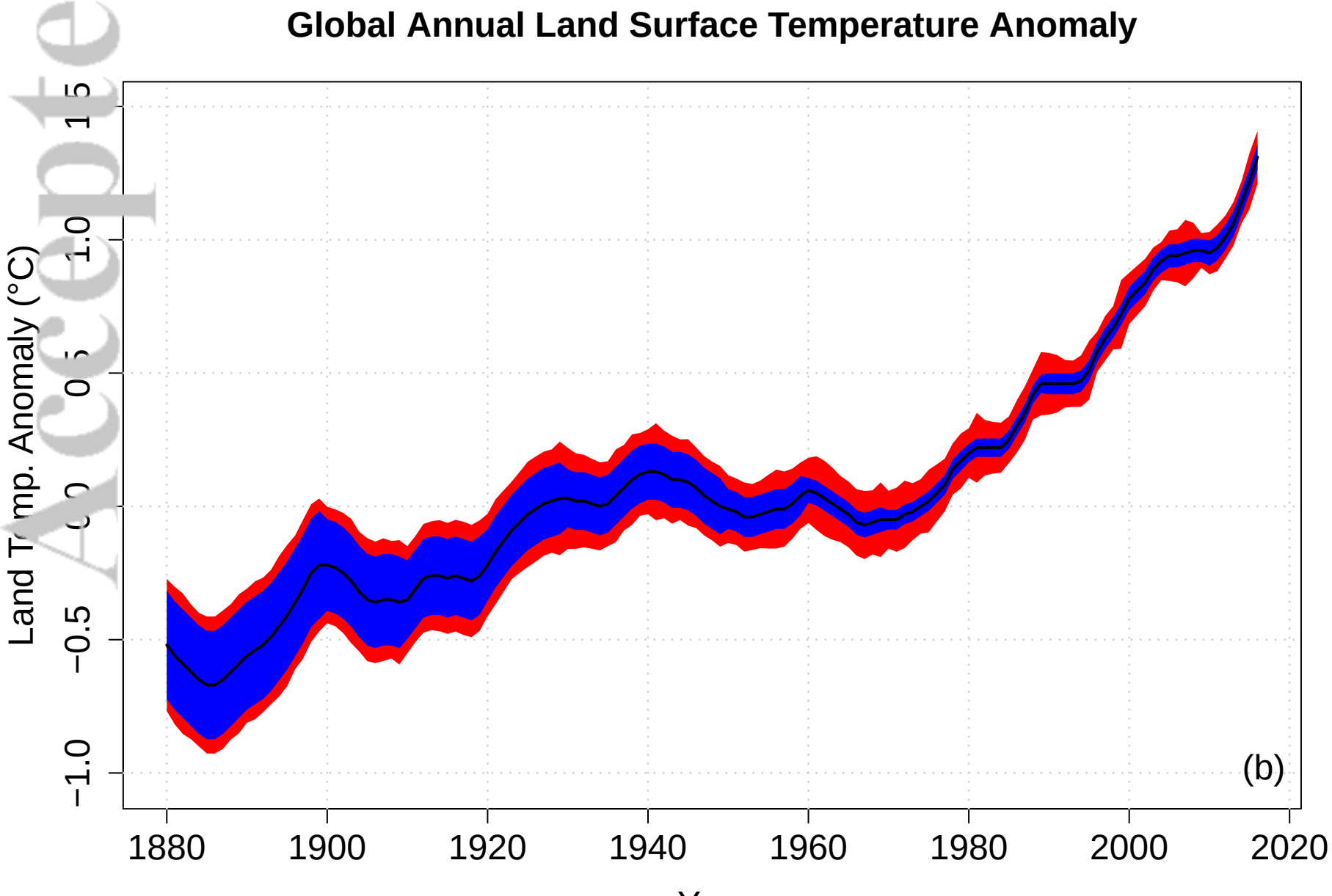


Revision Figure 6.

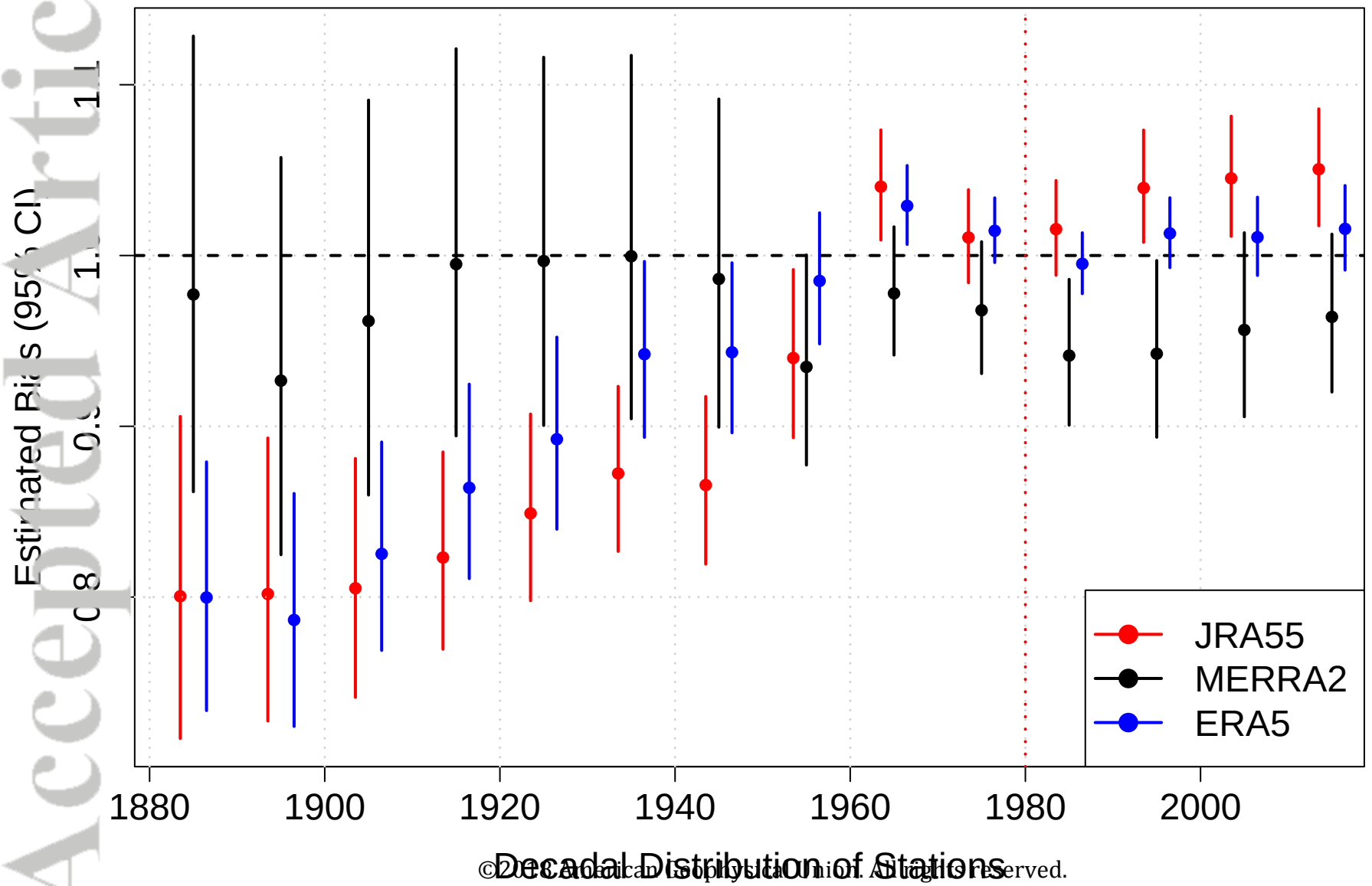


Revision Figure 7.

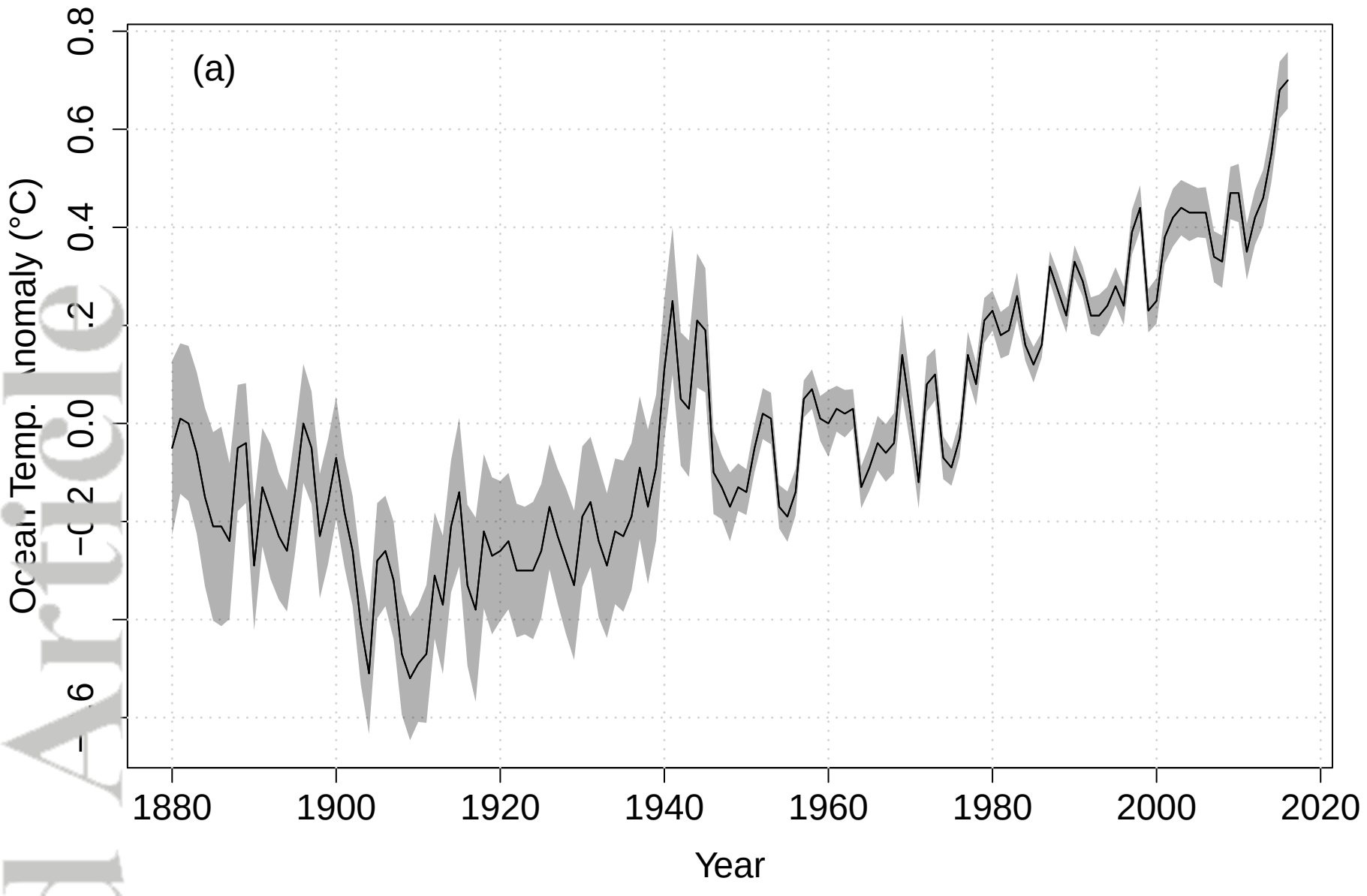


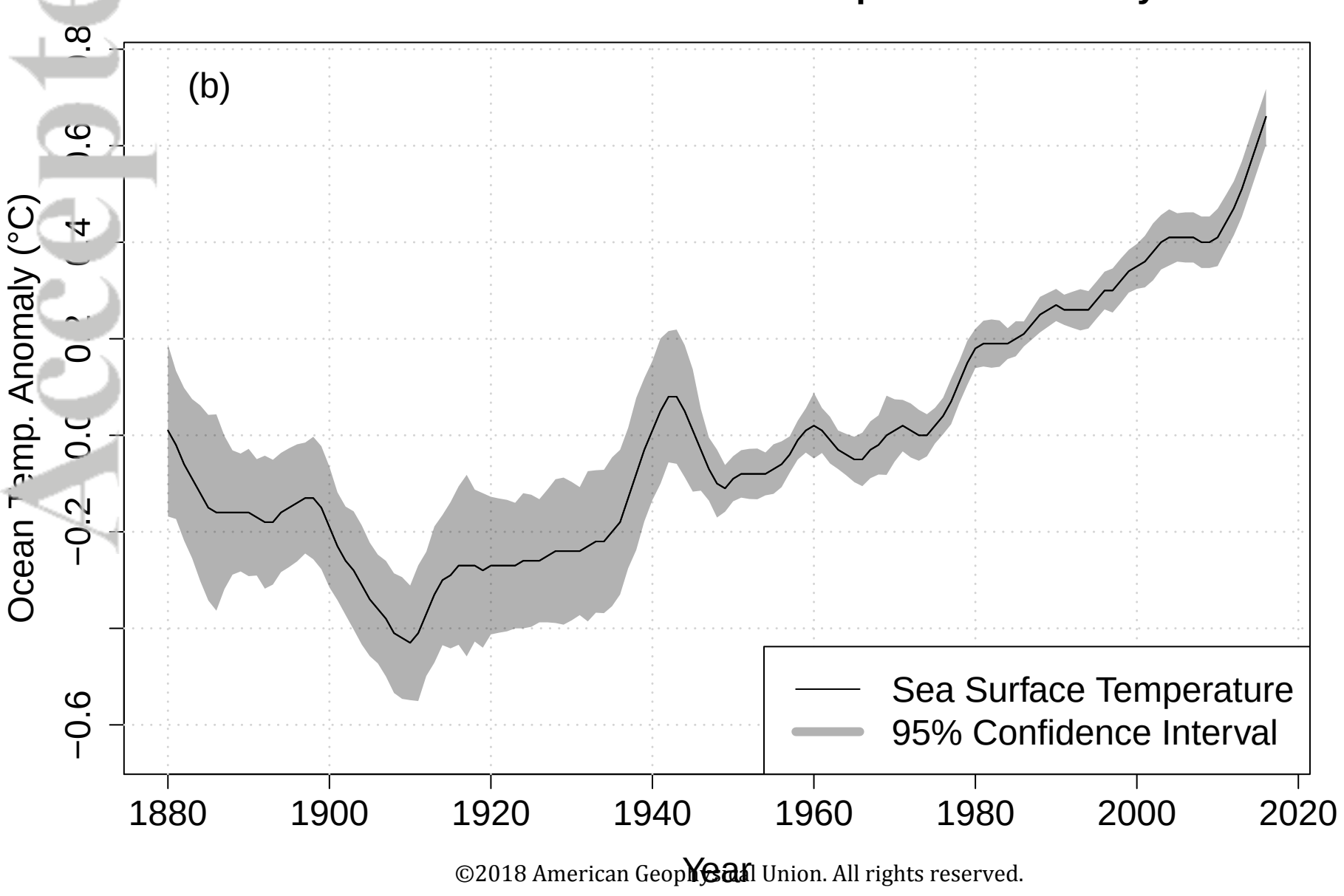


Revision Figure 8.

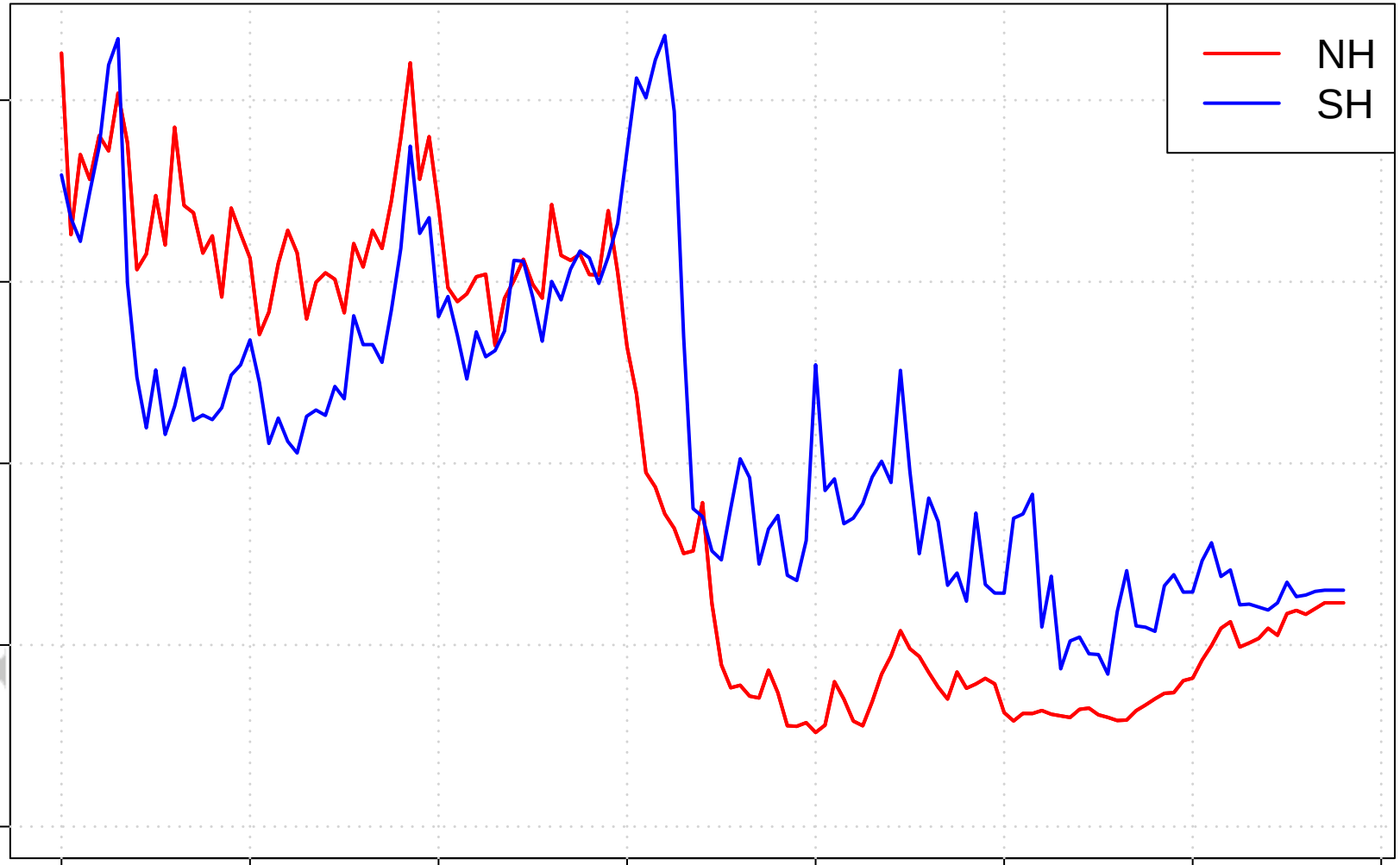


Revision Figure 9.

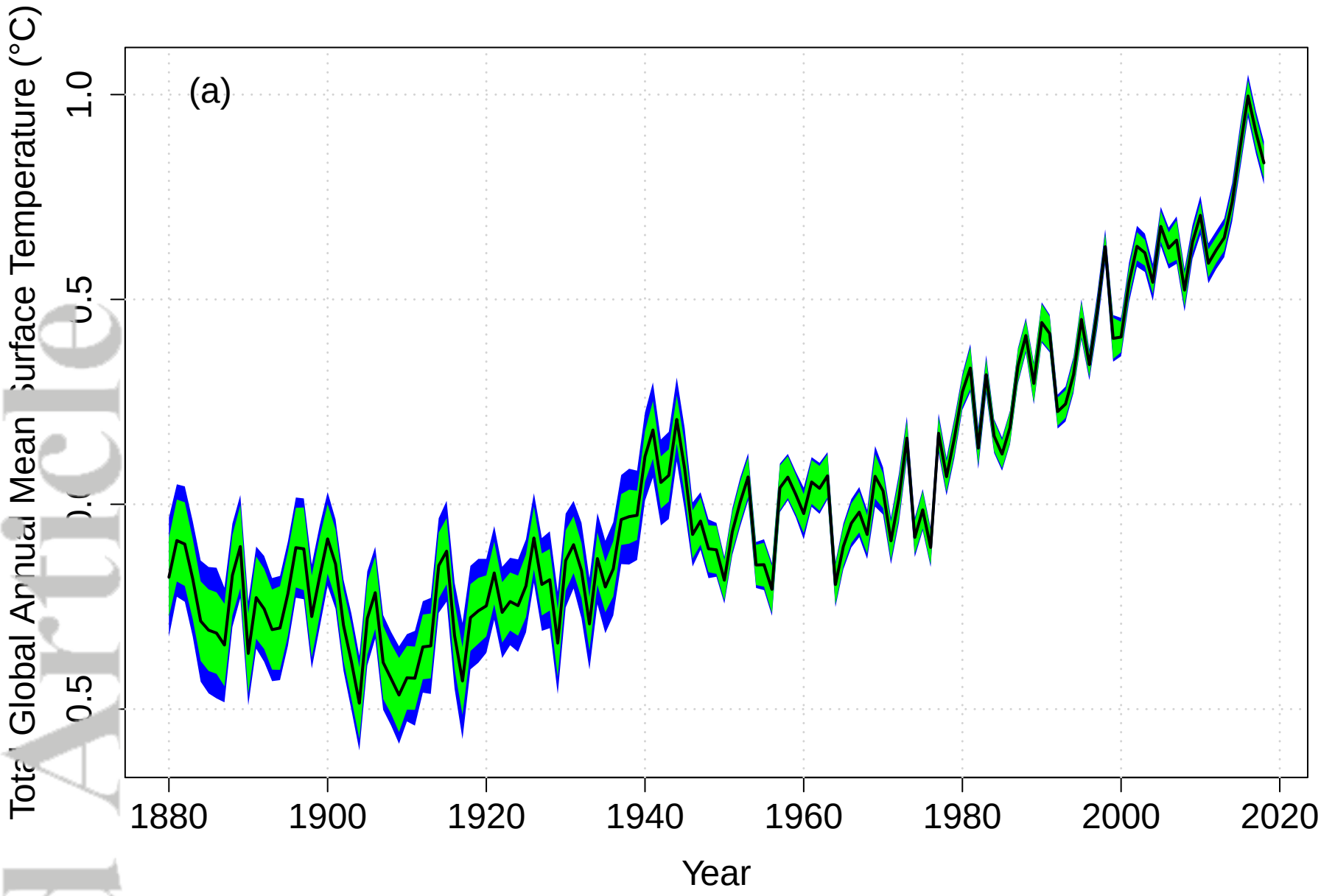




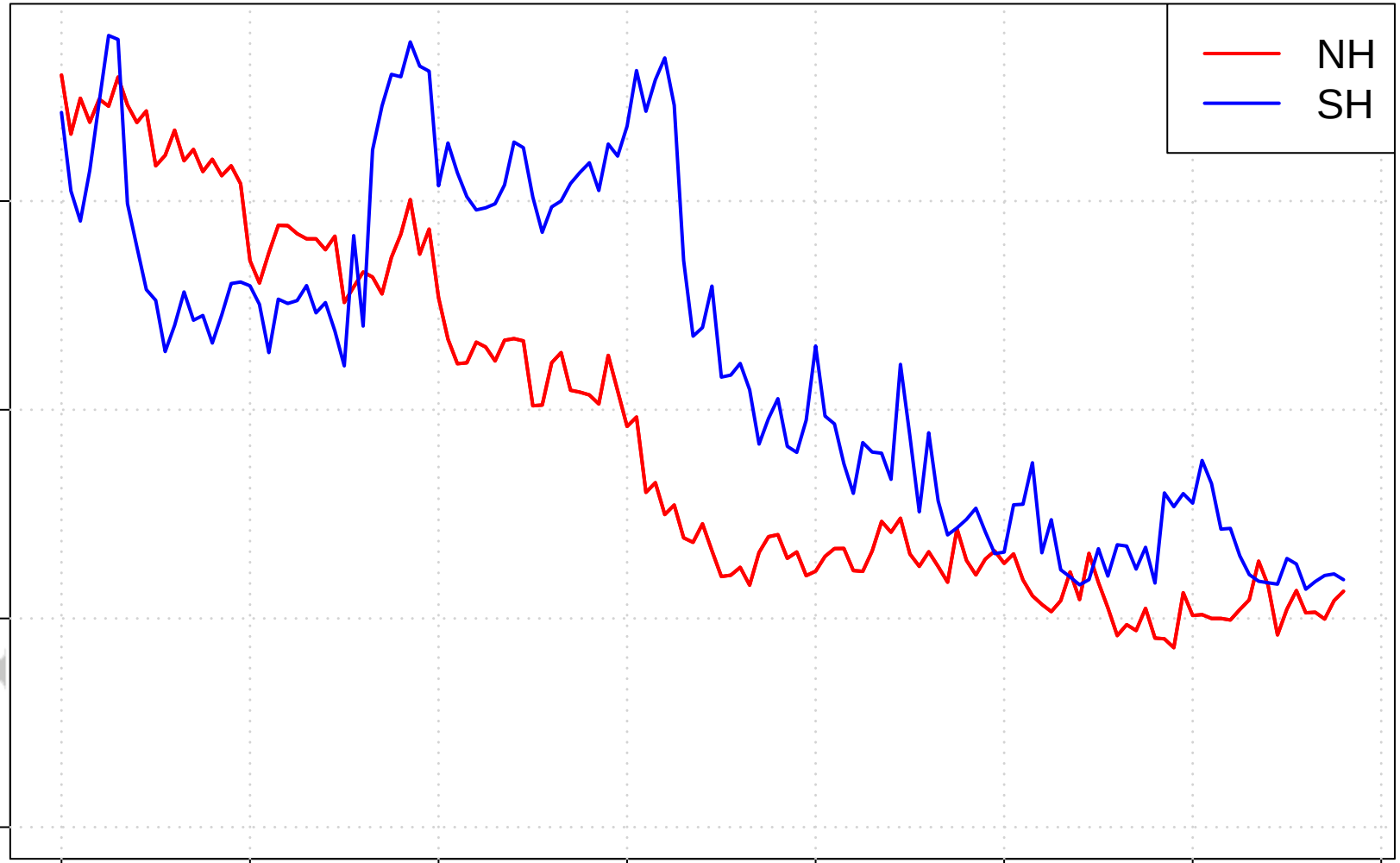
Revision Figure 10.



Revision Figure 11.



Revision Figure 12.



Revision Figure 13.