Berkeley Earth Temperature Averaging Process

Robert Rohde[1], Judith Curry[2], Donald Groom[3],

Robert Jacobsen[3,4], Richard A. Muller[1,3,4], Saul Perlmutter[3,4],

Arthur Rosenfeld[3,4], Charlotte Wickham[5], Jonathan Wurtele[3,4]

Corresponding address for all authors:

Berkeley Earth Project

2831 Garber St.

Berkeley CA 94705

email: RAMuller@LBL.gov

[1]Novim Group, Berkeley Earth Surface Temperature Project; [2]Georgia Institute of Technology; [3]Lawrence Berkeley National Laboratory; [4] University of California, Berkeley; [5]Now at Oregon State University

Abstract

A new mathematical framework is presented for producing maps and large-scale averages of temperature changes from weather station data for the purposes of climate analysis. This allows one to include short and discontinuous temperature records, so that nearly all temperature data can be used. The framework contains a weighting process that assesses the quality and consistency of a spatial network of temperature stations as an integral part of the averaging process. This permits data with varying levels of quality to be used without compromising the accuracy of the resulting reconstructions. Lastly, the process presented here is extensible to spatial networks of arbitrary density (or locally varying density) while maintaining the expected spatial relationships. In this paper, this framework is applied to the Global Historical Climatology Network land temperature dataset to present a new global land temperature reconstruction from 1800 to present with error uncertainties that include many key effects. In so doing, we find that the global land mean temperature has increased by 0.911 ± 0.042 C since the 1950s (95% confidence for statistical and spatial uncertainties). This change is consistent with global land-surface warming results previously reported, but with reduced uncertainty.

1. Introduction

While there are many indicators of climate change, the long-term evolution of global surface temperatures is perhaps the metric that is both the easiest to understand and most closely linked to the quantitative predictions of climate models. It is also backed by the largest collection of raw data. According to the summary provided by the Intergovernmental Panel on Climate Change (IPCC), the mean global surface temperature (both land and oceans) has increased 0.64 ± 0.13 C from 1956 to 2005 at 95% confidence (Trenberth et al. 2007).

During the latter half of the twentieth century weather monitoring instruments of good quality were widely deployed, yet the quoted uncertainty on global temperature change during this time period is still ± 20%. Reducing this uncertainty is a major goal of this paper. Longer records may provide more precise indicators of change; however, according to the IPCC, temperature increases prior to 1950 were caused by a combination of anthropogenic factors and natural factors (e.g. changes in solar activity), and it is only since about 1950 that man-made emissions have come to dominate over natural factors. Hence constraining the post-1950 period is of particular importance in understanding the impact of greenhouse gases.

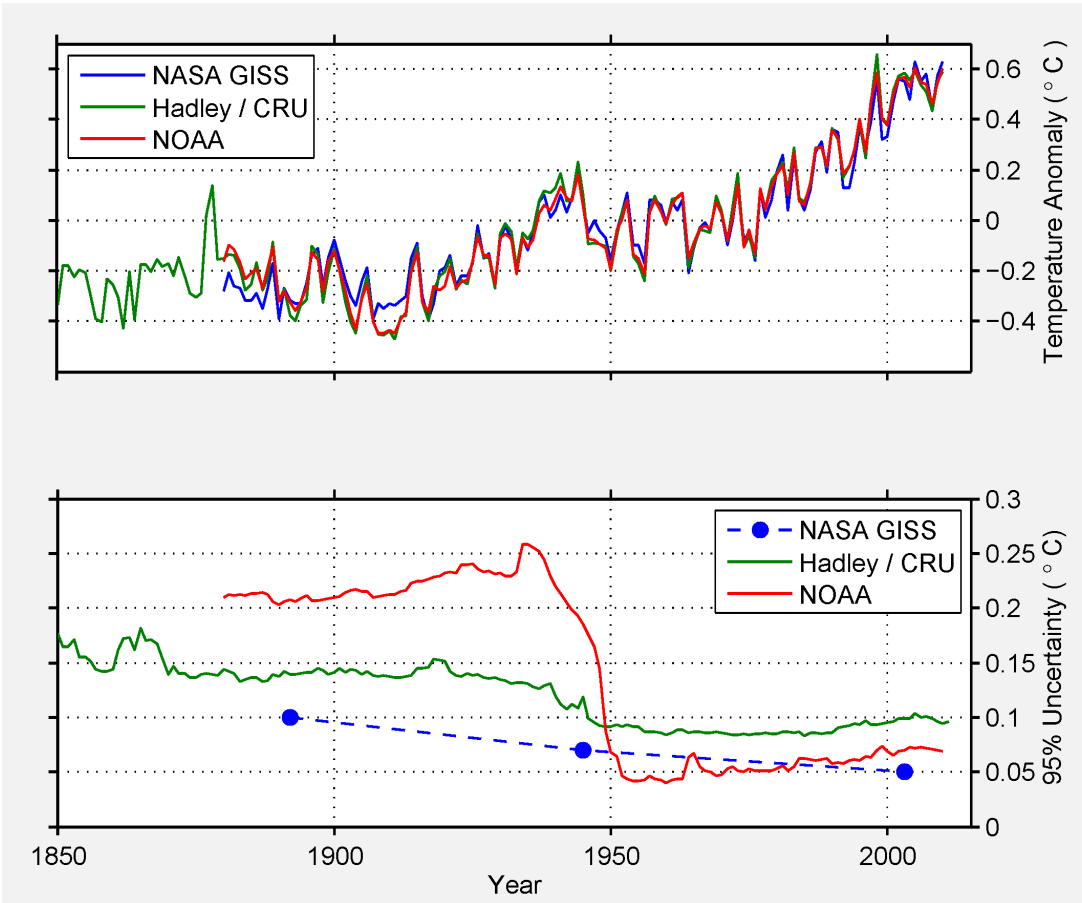
The Berkeley Earth Surface Temperature project was created to help refine our estimates of the rate of recent global warming. This is being approached through several parallel efforts to A) increase the size of the data set used to study global climate change, B) bring additional statistical techniques to bear on the problem that will help reduce the uncertainty in the resulting averages, and C) produce new analysis of systematic effects, including data selection bias, urban heat island effects, and the limitations of poor station siting. The current paper focuses on refinements in the averaging process itself and does not introduce any new data. The analysis framework described here includes a number of features to identify and handle unreliable data;

however, discussion of specific biases such as those associated with station siting and/or urban heat islands will also be published separately.

2. Averaging Methods of Prior Studies

Presently there are three major research groups that routinely produce a global average time series of instrumental temperatures for the purposes of studying climate change. These groups are located at the National Aeronautics and Space Administration Goddard Institute for Space Studies (NASA GISS), the National Oceanic and Atmospheric Administration (NOAA), and a collaboration of the Hadley Centre of the UK Meteorological Office with the Climate Research Unit of East Anglia (HadCRU). They have developed their analysis frameworks over a period of about 25 years and share many common features (Hansen and Lebedeff 1987; Hansen et al. 1999; Hansen et al. 2010; Jones et al. 1986; Jones and Moberg 2003; Brohan et al. 2006; Smith and Reynolds 2005; Smith et al. 2008). The global average time series for the three groups are presented in Figure 1 and their relative similarities are immediately apparent. Each group combines measurements from fixed-position weather stations on land with transient ships / buoys in water to reconstruct changes in the global average temperature during the instrumental era, roughly 1850 to present. Two of the three groups (GISS and HadCRU) treat the land-based and ocean problems as essentially independent reconstructions with global results only formed after constructing separate land and ocean time series. The present paper will present improvements and innovations for the processing of the land-based measurements. Though much of the work presented can be modified for use in an ocean context, we will not discuss that application at this time due to the added complexities and systematics involved in monitoring from mobile ships / buoys.

Figure 1. Comparison of the global annual averages and annual average uncertainty



In broad terms each land-based temperature analysis can be broken down into several overlapping pieces: A) the compilation of a basic dataset, B) the application of a quality control and "correction" framework to deal with erroneous, biased, and questionable data, and C) a process by which the resulting data is mapped and averaged to produce useful climate indices. The existing research groups use different but heavily overlapping data sets consisting of between 4400 and 7500 weather monitoring stations (Brohan et al. 2006, Hansen et al. 2010; Peterson and Vose 1997). Our ongoing work to build a climate database suggests that over 40000 weather station records have been digitized. All three temperature analysis groups derive a global average time series starting from monthly average temperatures, though daily data and records of maximum and minimum temperatures (as well as other variables such as precipitation) are of increasing used in other forms of climate analysis (Easterling et al. 1997, Klein and Können 2003, Alexander et al. 2006, Zhang et al. 2007). The selection of stations to include in climate analyses has been heavily influenced by algorithms that require the use of long, nearly-continuous records. Secondarily, the algorithms often require that all or most of a reference "baseline" period be represented from which a station's "normal" temperature is defined. Each group differs in how it approaches these problems and the degree of flexibility they have in their execution, but these requirements have served to exclude many temperature records shorter than 15 years from existing analyses (only 5% of NOAA records are shorter than 15 years).

The focus on methods that require long records may arise in part from the way previous authors have thought about the climate. The World Meteorological Organization (WMO) gives an operational definition of climate as the average weather over a period of 30 years (Arguez and Vose 2011). From this perspective, it is trivially true that individual weather stations must have very long records in order to perceive multi-decadal climate changes from a single site. However, as we will show, the focus on long record lengths is unnecessary when one can compare many station records with overlapping spatial and temporal coverage.

Additionally, though the focus of existing work has been on long records, it is unclear that such records are ultimately more accurate for any given time interval than are shorter records covering the same interval. The consistency of long records is affected by changes in instrumentation, station location, measurement procedures, local vegetation and many other factors that can introduce artificial biases in a temperature record (Folland et al. 2001, Peterson and Vose 1997, Brohan et al. 2006, Menne et al. 2009, Hansen et al. 2001). A previous analysis of the 1218 stations in US Historical Climatology Network found that on average each record has one spurious shift in mean level greater than about 0.5 C for every 15-20 years of record (Menne et al. 2009). Existing detection algorithms are inefficient for biases less than 0.5 C, suggesting that the typical length of record reliability is likely to be even shorter. All three groups have developed procedures to detect and "correct" for such biases by introducing adjustments to individual time series. Though procedures vary, the goal is generally to detect spurious changes in a record and use neighboring series to derive an appropriate adjustment. This process is generally known as "homogenization", and has the effect of making the temperature network more spatially homogeneous but at the expense that neighboring series are no longer independent. For all of the existing groups, this process of bias adjustment is a separate step conducted prior to constructing a global average.

After homogenization (and other quality control steps), the existing groups place each "corrected" time series in its spatial context and construct a global average. The simplest process, conducted by HadCRU, divides the Earth into 5° x 5° latitude-longitude grid cells and associates the data from each station time series with a single cell. Because the size of the cells varies with latitude, the number of records per cell and weight per record is affected by this gridding process in a way that has nothing to do with the nature of the underlying climate. In contrast, GISS uses an 8000-element equal-area grid, and associates each station time series with multiple grid cells by defining the grid cell average as a distance-weighted function of temperatures at many nearby station locations. This captures some of the spatial structure and is resistant to many of the gridding artifacts that can affect HadCRU. Lastly, NOAA has the most sophisticated treatment of spatial structure. NOAA's process, in part, decomposes an estimated spatial covariance matrix into a collection of empirical modes of spatial variability on a 5° x 5°

grid. These modes are then used to map station data onto the grid according to the degree of covariance expected between the weather at a station location and the weather at a grid cell center. (For additional details, and explanation of how low-frequency and high-frequency modes are handled differently, see Smith and Reynolds 2005). In principle, NOAA's method should be the best at capturing and exploiting spatial patterns of weather variability. However, their process relies on defining spatial modes during a relatively short modern reference period (1982-1991 for land records, Smith and Reynolds 2005), and they must assume that the patterns of spatial variation observed during that interval are adequately representative of the entire history. Further, if the goal is to understand climate change then the assumption that spatial patterns of weather variability are time-invariant is potentially confounding.

In all three of these prior approaches, every record used in gridded averaging is assumed to be equally reliable. More precisely, they make the assumption that their quality control and homogenization processes address erroneous and biased data prior to the gridding and averaging step in such a way that each resulting time series is deserving of equal weight. (GISS makes a partial exception in that a corrective model for urban heat island biases is applied after gridding.) This has the effect that records subject to many bias "corrections" can be given the same weight in an average as a record where no bias adjustments were found to be necessary. In such cases, the differences in data quality may play a role in how the uncertainty is assessed, but not in the construction of the global average.

All three of the averaging processes currently being used rely on the concept of a "baseline" parameter to define the "normal" weather. The baseline can either be introduced for each record before gridding (e.g. HadCRU) or it can be introduced after gridding and defined at the level of the grid cell average (e.g. NASA). The intent of the baseline temperature parameter is to capture the "normal" climate at that location by reference to the average weather over some specific reference period (e.g. 1960-1980). Each time series is then replaced by an "anomaly" time series consisting of the differences from the baseline. This approach is motivated by the observation that temperatures change rapidly with latitude (about 1 C per 150 km poleward) and altitude (about 1 C for every 220 m of surface elevation), and that these changes are quite large compared to the approximately 1 C / century of global warming that one wants to investigate. In effect, the baseline parameters are meant to capture most of the spatial variability between sites. In particular, the average of anomaly series should be much less sensitive to biases due to the start and stop of individual records. Without some adjustment for such spatial variability, an excess of high (or low) latitude stations could erroneously pull the corresponding global average to lower (or higher) values.

The use of an individual baseline parameter per station (or grid cell) makes no assumptions about the underlying spatial structure. This means the maximum spatial information can in principle be removed from each record; however, several trade-offs are incurred in doing so. First, the use of predefined reference intervals will limit the usability of stations that were not active during the corresponding period (though other compensating approaches are often used). Secondly, by defining all stations to have zero anomaly during the reference period, one may suppress true structure in the temperature field at that time. Specifically, reconstructions using this method will have lower spatial variability during the reference interval than at other times due to the artificial constraint that all regions have the same mean value during the reference period.

Lastly, after gridding the data and creating anomaly series, each existing group creates a large-scale average using an area-weighted average of non-empty grid cells. HadCRU and GISS add an additional nuance, as they apply a post-stratification procedure prior to their final average. Specifically, they create averages of specific latitude bands (or hemispheres in HadCRU's case), and then combine those average to create the final global average. This has the effect that each missing cell in a latitude band is essentially replaced by the average of the valid cells in the band before constructing the ultimate global average. To a degree this approach also compensates for the fact that certain areas (e.g. the Northern Hemisphere) tend to have much greater historical coverage than others. Monte Carlo tests we conducted generally confirm that latitudinal banding improves the accuracy of the overall average given the techniques employed by HadCRU and GISS; however, we observe that such approaches are largely an indirect means of incorporating information about the spatial structure of the temperature field that could be modeled more directly.

3. New Averaging Model

The global average temperature is a simple descriptive statistic that aims to characterize the Earth. Operationally, the global average may be defined as the integral average of the temperatures over the surface of the Earth as would be measured by an ideal weather station sampling the air at every location. As the true Earth has neither ideal temperature stations nor infinitely dense spatial coverage, we can never capture the ideal global average temperature completely; however, we can use the data we do have to constrain its value.

As described in the preceding section, the existing global temperature analysis groups use a variety of well-motivated algorithms to generate a history of global temperature change. However, none of their approaches would generally correspond to a statistical model in the more formal sense.

Let $T(\vec{x},t)$ be the global temperature field in space and time. We define the decomposition:

$$T(\vec{x},t) = \theta(t) + C(\vec{x}) + W(\vec{x},t) \quad [1]$$

Uniqueness can be guaranteed by applying the constraints:

$$\int_{\text{Earth's surface}} C(\vec{x})d\vec{x} = 0,$$
$$\int_{\text{Earth's surface}} W(\vec{x},t)d\vec{x} = 0, \text{ for all } t, \quad [2]$$
$$\int_{\text{Earth's surface}} W(\vec{x},t)dt = 0, \text{ for all locations } \vec{x}$$

Given this decomposition, we see that $\theta(t)$ corresponds to the global mean temperature as a function of time. $C(\vec{x})$ captures the time-invariant spatial structure of the temperature field, and hence can be seen as a form of spatial "climatology", though it differs from the normal definition of a climatology by a simple additive factor corresponding to the long-term average of $\theta(t)$. The last term, $W(\vec{x},t)$, is meant to capture the "weather", i.e. those fluctuations in temperature over space and time that are neither part of the long-term evolution of the average nor part of the stable spatial structure. In this paper, we show how it is possible to estimate the global temperature field by simultaneously constraining all three pieces of $T(\vec{x},t)$ using the available data. (Because we are introducing a large number of symbols, we summarize all the key symbols in the Appendix.)

As our study is based solely on the use of land-based temperature data, we choose to restrict the spatial integrals in equation [2] to only the Earth's land surface. As a result, our study will identify $\theta(t)$ with the land temperature average only. Rather than defining a specific base interval (e.g. 1950-1980) as has been common in prior work, we will show below how it is possible to reconcile all time periods simultaneously. As a result, the time integral in equation [2] should be understood as occurring over the full multi-century period from which data is available. As a side-effect of this approach, $W(\vec{x},t)$ will also incorporate some multi-decadal changes that might more typically be described as changes in climate rather than "weather".

We further break $C(\vec{x})$ into a number of additional components:

$$C(\vec{x}) = \lambda(\text{ latitude}(\vec{x})\text{ }) + h(\text{ elevation}(\vec{x})\text{ }) + G(\vec{x}) \qquad [3]$$

Here λ depends only on the latitude of $\vec{x}$, h depends only on the elevation of $\vec{x}$, and $G(\vec{x})$ is the "geographic anomaly", i.e. the spatial variations in mean climatology that can't be explained solely by latitude and elevation. With appropriate models for λ and h it is possible to explain about 95% of the variance in annual mean temperatures over the surface of the Earth in terms of just latitude and elevation. The functional forms of λ, h, and $G(\vec{x})$ will be discussed below.

Consider a temperature monitoring station at location $\vec{x}_i$, we expect the temperature datum $d_i(t_j)$ to ideally correspond to $T(\vec{x}_i, t_j) = \theta(t_j) + C(\vec{x}_i) + W(\vec{x}_i, t_j)$. More generally, we assert that:

$$d_i(t_j) = \theta(t_j) + b_i + W(\vec{x}_i, t_j) + \epsilon_{i,j} \qquad [4]$$

Where $\epsilon_{i,j}$ is defined to be error in the *i*-th station and the *j*-th time step, and b_i is the "baseline" temperature for the *i*-th station necessary to minimize the error. With this definition

$$a_i = b_i - C(\vec{x}_i) \qquad [5]$$

is a measure of the bias at the *i*-th station relative to the true climatology.

For each of the parameters and fields we have discussed we shall use the "hat" notation, e.g. $\hat{\theta}(t_j)$, $\hat{b}_i$, to denote values that are estimated from data and to distinguish them from the true fields specified by definition. Given equation [4], it is natural to consider finding fields that minimize expressions of the form

$$SSD = \sum_{i,j} \left(d_i(t_j) - \hat{\theta}(t_j) - \hat{b}_i - \widehat{W}(\vec{x}_i, t_j) \right)^2 \approx \sum_{i,j} \epsilon_{i,j}{}^2 \quad [6]$$

Where SSD denotes the sum of square deviations and such a minimization would attempt to minimize the error terms. Though appealing, [6] is ultimately misguided as $d_i(t_j)$ is distributed highly non-uniformly in both space and time, and the temperature histories at neighboring stations are highly correlated. A naïve application of [6] would result in $\hat{\theta}(t_j)$ biased towards the most densely sampled regions of the globe.

However, [6] does inspire our first natural set of constraint equations, namely

$$\hat{b}_i = \frac{\sum_j \left(d_i(t_j) - \hat{\theta}(t_j) - \widehat{W}(\vec{x}_i, t_j) \right)}{\sum_j 1} \quad [7]$$

Since $\hat{b}_i$ is specific to a single station, there is no disadvantage to simply stating that it be chosen to minimize the error at that specific station.

To determine the other fields, it is instructive to consider the properties that we expect $\widehat{W}(\vec{x}_i, t_j)$ to have. To begin, it should have (at least approximately) zero mean over space and time in accordance with equation [2]. Secondly, we expect that weather fluctuations should be highly correlated over short distances in space. These considerations are very similar to the fundamental assumptions of the spatial statistical analysis technique known as Kriging (Krige 1951, Cressie 1990, Journel 1989). Provided the assumptions of Kriging are met, this technique provides best linear unbiased estimator of the underlying spatial field.

The simple Kriging estimate of a field, $M(\vec{x})$, from a collection of measurements M_i having positions $\vec{x}_i$ is:

$$\widehat{M}(\vec{x}) = \sum_{i=1}^{N} K_i(\vec{x})\, M_i \qquad [8]$$

266

$$\begin{pmatrix} K_1(\vec{x}) \\ \vdots \\ K_N(\vec{x}) \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \mathrm{Cov}(\vec{x}_1, \vec{x}_2) & \cdots & \mathrm{Cov}(\vec{x}_1, \vec{x}_N) \\ \mathrm{Cov}(\vec{x}_2, \vec{x}_1) & \sigma_2^2 & & \mathrm{Cov}(\vec{x}_2, \vec{x}_N) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(\vec{x}_N, \vec{x}_1) & \mathrm{Cov}(\vec{x}_N, \vec{x}_2) & \cdots & \sigma_N^2 \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{Cov}(\vec{x}, \vec{x}_1) \\ \vdots \\ \mathrm{Cov}(\vec{x}, \vec{x}_N) \end{pmatrix} \qquad [9]$$

267 Where σ_i^2 is the variance at the *i*-th site and $\mathrm{Cov}(\vec{a}, \vec{b})$ is the covariance between sites $\vec{a}$
268 and $\vec{b}$. If the covariance is known and M_i are sampled from an underlying population having
269 zero mean, then equation [8] provides the best linear unbiased estimate of the field $M(\vec{x})$. In
270 particular, Kriging describes a natural way to adjust the weight that each record receives in order
271 to avoid overweighting densely sampled regions. This adjustment for station density is an
272 intrinsic part of the inverse covariance matrix.

273 In order to take advantage of the statistical properties of simple Kriging, it is necessary
274 that the data field on which the interpolation is based have zero mean. However, this limitation
275 is removed by "ordinary" Kriging where the addition of extra parameter(s) is used to transform
276 the data set by removing known spatial structure (Journel 1989, Cressie 1990). In our case, it is
277 natural to identify the sampled data as:

$$M_i = d_i(t_j) - \hat{\theta}(t_j) - \hat{b}_i \qquad [10]$$

278 which would be expected to have zero mean per equation [4]. For the "ordinary" Kriging
279 approach the ideal parameterization is found by choosing parameters $\hat{\theta}$ and $\hat{b}_i$ that minimize the
280 average variance of the field, e.g.

$$\text{Minimize: } \int_{\text{Earth's surface}} \left(M(\vec{x},t)\right)^2 d\vec{x} \qquad [11]$$

In most practical uses of Kriging it is necessary to estimate or approximate the covariance matrix in equation [9] based on the available data (Krige 1951, Cressie 1990, Journel 1989). NOAA also requires the covariance matrix for their optimal interpolation method; they estimate it by first constructing a variogram during a time interval with dense temperature sampling and then decomposing it into empirical spatial modes that are used to model the spatial structure of the data (Smith and Reynolds 2005). Their approach is nearly ideal for capturing the spatial structure of the data during the modern era, but has several weaknesses. Specifically this method assumes that the spatial structures are adequately constrained during a brief calibration period and that such relationships remain stable even over an extended period of climate change.

We present an alternative that preserves many of the natural spatial considerations provided by Kriging, but also shares characteristics with the local averaging approach adopted by GISS (Hansen et al 1999, Hansen and Lebedeff 1987). If the variance of the underlying field changes slowly as a function of location, then the covariance function can be replaced with the correlation function, $R(\vec{a},\vec{b})$, which leads to the formulation that:

$$\begin{pmatrix} S_{a_1}(\vec{x},t_j) \\ \vdots \\ S_{a_N}(\vec{x},t_j) \end{pmatrix} = \begin{pmatrix} 1 & R(\vec{x}_{a_1},\vec{x}_{a_2}) & \cdots & R(\vec{x}_{a_1},\vec{x}_{a_N}) \\ R(\vec{x}_{a_2},\vec{x}_{a_1}) & 1 & & R(\vec{x}_{a_2},\vec{x}_{a_N}) \\ \vdots & & \ddots & \vdots \\ R(\vec{x}_{a_N},\vec{x}_{a_1}) & R(\vec{x}_{a_N},\vec{x}_{a_2}) & \cdots & 1 \end{pmatrix}^{-1} \begin{pmatrix} R(\vec{x},\vec{x}_{a_1}) \\ \vdots \\ R(\vec{x},\vec{x}_{a_N}) \end{pmatrix} \qquad [12]$$

Where $a_1 \dots a_N$ denotes the collection of stations active at time t_j, and thus

$$\widehat{W}(\vec{x},t_j) = \sum_{i=1}^{N} S_{a_i}(\vec{x},t_j)\left(d_{a_i}(t_j) - \hat{\theta}(t_j) - \hat{b}_{a_i}\right) \qquad [13]$$

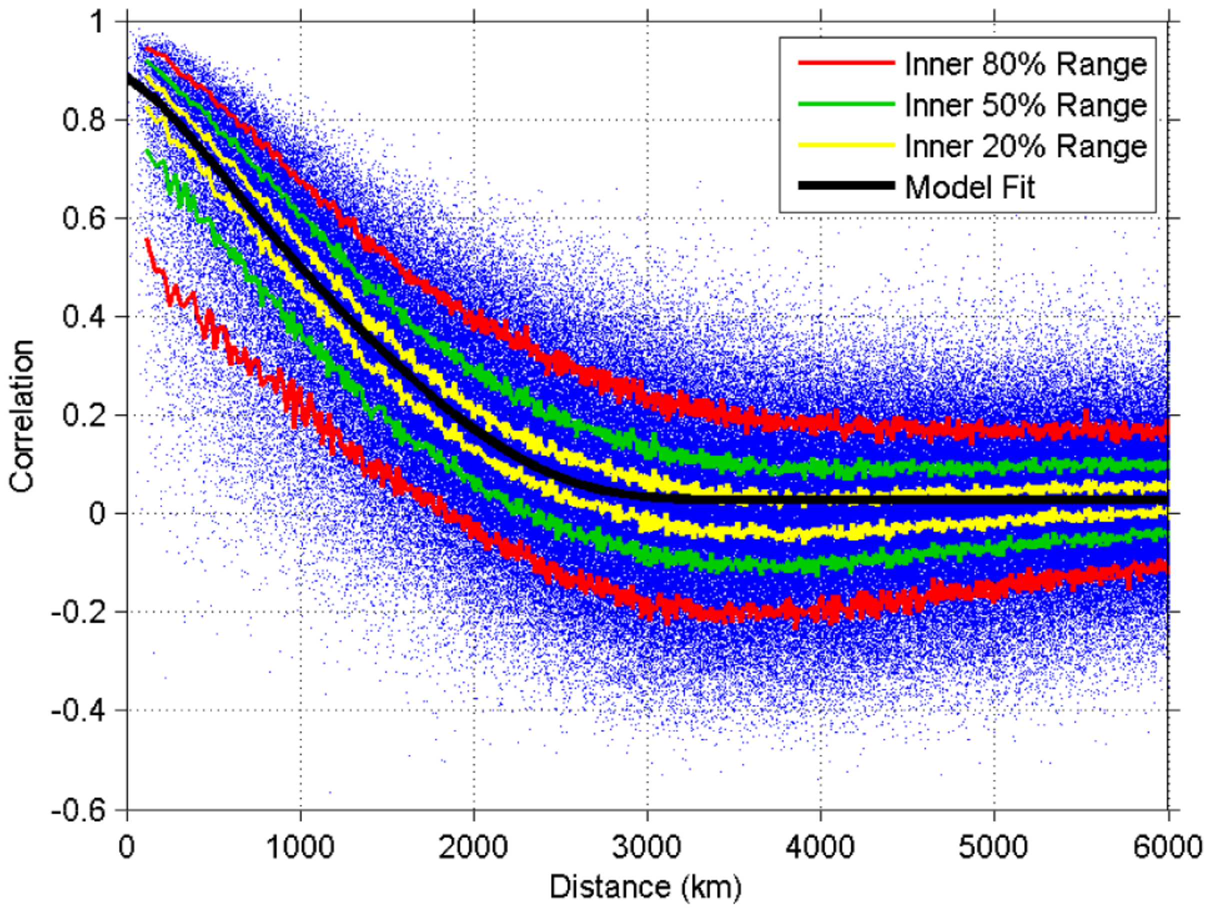
The Kriging formulation is most efficient at capturing fluctuations which have a scale length comparable to the correlation length; however, it also permits the user to find finer structure if more densely positioned data is provided. In particular, in the limit of infinitely dense data, the Kriging estimate of the field will necessarily match the field exactly. This is in direct contrast to the GISS and HadCRU averaging approaches which will always smooth over fine structure.

A further modification is made by assuming that $R(\vec{a},\vec{b}) \approx R(d)$, where $d = |\vec{a} - \vec{b}|$ denotes the distance between $\vec{a}$ and $\vec{b}$. This allows us to parameterize the correlation field as a simple function of one variable, though it admittedly neglects differences in correlation that might be related to factors such as latitude, altitude, and local vegetation, etc. The correlation function is parameterized using:

$$R(d) = e^{-(\alpha+\beta d+\gamma d^2+\varepsilon d^3+\eta d^4)} + \mu \qquad [14]$$

This is compared to a reference data set based on randomly selecting 500,000 pairs of stations, and measuring the correlation of their non-seasonal temperature fluctuations provided they have at least ten years of overlapping data. The resulting data set and fit are presented in Figure 2. Pair selection was accomplished by choosing random locations on the globe and locating the nearest temperature records, subject to a requirement that it be no more than 100 km from the chosen random location. The small constant term μ measures the correlation over the very largest distance scale; however, for the purposes of equation [12] it is computationally advantageous to set $\mu = 0$ which we did while scaling the rest of equation [14] by $1/(1-\mu)$ to compensate near $d = 0$. This allows us to treat stations at distances greater than ~4000 km as completely uncorrelated, which greatly simplifying the matrix inversion in equation [12] since a majority of the matrix elements are now zeros. Figure 2 shows that the correlation structure is substantial out to a distance of ~1000 km, and non-trivial to ~2000 km from each site.

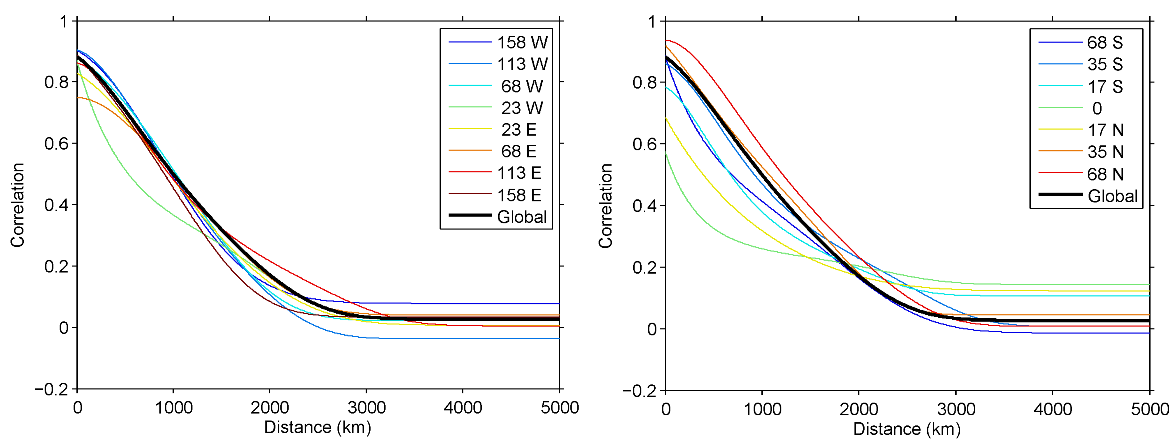
Figure 2. Mean correlation versus distance curve



Based on the data, the best fit values in equation [14] were $\alpha = 0.1276$, $\beta = 2.4541 \times 10^{-4}$ / km, $\gamma = 5.3881 \times 10^{-7}$ / km^2, $\varepsilon = -2.7452 \times 10^{-11}$ / km^3, $\theta = 8.3007 \times 10^{-14}$ / km^4 and $\mu = 0.0272$. These were the values we used in the Berkeley Earth temperature reconstruction method.

In Figure 3 we show similar fits using station pairs restricted by either latitude or longitude. In the case of longitude, we divide the Earth into 8 longitude bands and find that the correlation structure is very similar across each. The largest deviation occurs in the band centered at 23 W with reduced correlation at short distances. This band is one of several that include relatively few temperature stations as it spans much of the Atlantic Ocean, and so this deviation might be primarily a statistical fluctuation. The deviations observed in Figure 3 for latitude bands are more meaningful however. We note that latitude bands show decreasing short-range correlation as one approaches the equator and a corresponding increase in long-range correlation. Both of these effects are consistent with decreased weather variability in most tropical areas. These variations, though non-trivial, are relatively modest for most regions. For the current presentation we shall restrict ourselves to the simple correlation function given by equation [14], though further refinements of the correlation function are likely to be a topic of future research.

Figure 3. Correlation versus distance fits using only stations selected from portions of the Earth



We note that the correlation in the limit of zero distance, $R(0) = 0.8802$, has a natural and important physical interpretation. It is an estimate of the correlation that one expects to see between two typical weather monitors placed at the same location. By extension, if we assume such stations would report the same temperature except that each is subject to random and uncorrelated error, then it follows that $1 - R(0) = 12.0\%$ of the non-seasonal variation in the typical station record is caused by measurement noise that is unrelated to the variation in the underlying temperature field. Since the average root-mean-square non-seasonal variability is ~2.0 C, it follows that an estimate of the short-term instrumental noise for the typical month at a typical station is ~0.47 C at 95% confidence. This estimate is much larger than the approximately 0.06 C typically used for the random monthly measurement error (Folland et al. 2001). Our correlation analysis suggests that such estimates may understate the amount of random noise introduced by local and instrumental effects. However, we note that the same authors assign an uncertainty of 0.8 C to the homogenization process they use to remove longer-term biases. We suspect that the difficulty they associate with homogenization is partially caused by the same short-term noise that we observe. However, our correlation estimate would not generally include long-term biases that cause a station to be persistently too hot or too cold, and so the estimates are not entirely comparable. The impact of short-term local noise on the ultimate temperature reconstruction can be reduced in regions where stations are densely located and thus provide overlapping coverage. The simple correlation function described above would imply that each temperature station captures $\frac{\iint R(\vec{x})^2 d\vec{x}}{\iint 1 d\vec{x}} = 0.58\%$ of the Earth's temperature field; equivalently, 180 ideally distributed weather stations would be sufficient to capture nearly all of the expected structure in the Earth's monthly mean anomaly field. This is similar to the estimate of 110 to 180 stations provided by Jones 1994. We note that the estimate of 180 stations includes the effect of measurement noise. Removing this consideration, we would find that the underlying monthly mean temperature field has approximately 115 independent degrees of freedom. In practice though, quality control and bias correction procedures will substantially increase the number of records required.

The new Kriging coefficients $S_i(\vec{x}, t_j)$ defined by equation [12] also have several natural interpretations. Firstly the average of $S_i(\vec{x}, t_j)$ over land:

$$0 \leq \frac{\int S_i(\vec{x}, t_j)\, d\vec{x}}{\int 1\, d\vec{x}} < 1 \qquad \textbf{[15]}$$

can be interpreted as the total weight in the global land-surface average attributed to the *i*-th station at time t_j. Secondly, the use of correlation rather than covariance in our construction, gives rise to a natural interpretation of the sum of $S_i(\vec{x}, t_j)$ over all stations. Because Kriging is linear and our construction of R is positive definite, it follows that:

$$0 \leq F(\vec{x}, t_j) \equiv \sum_i S_i(\vec{x}, t_j) \leq 1 \qquad \textbf{[16]}$$

Where $F(\vec{x}, t_j)$ has the qualitative interpretation as the fraction of the $W(\vec{x}, t_j)$ field that has been effectively constrained by the data. The above is true even though individual terms $S_i(\vec{x}, t_j)$ may in general be negative. Since the true temperature anomaly is

$$\begin{aligned} \hat{\theta}(t_j) + \widehat{W}(\vec{x}, t_j) &= \hat{\theta}(t_j) + \sum_i S_i(\vec{x}, t_j)\left(d_i(t_j) - \hat{b}_i - \hat{\theta}(t_j)\right) \\ &= \left(1 - F(\vec{x}, t_j)\right)\hat{\theta}(t_j) + \sum_i S_i(\vec{x}, t_j)\left(d_i(t_j) - \hat{b}_i\right) \end{aligned} \qquad \textbf{[17]}$$

we see that in the limit $F(\vec{x}, t_j) \rightarrow 1$, the temperature estimate at $\vec{x}$ depends only on the local data $d_i(t_j)$, while in the limit $F(\vec{x}, t_j) \ll 1$ the temperature field at $\vec{x}$ is estimated to have the same value as the global average of the data. For diagnostic purposes it is also useful to define:

$$\bar{F}(t_j) = \frac{\int F(\vec{x}, t_j)\, d\vec{x}}{\int 1\, d\vec{x}} \qquad \textbf{[18]}$$

which provides a measure of total field completeness as a function of time.

Under the ordinary Kriging formulation, we would expect to find the parameters $\hat{\theta}(t_j)$ and $\hat{b}_i$ by minimizing a quality of fit metric:

$$\int \widehat{W}\left(\vec{x}, t_j\right)^2 \, d\vec{x} \tag{19}$$

Minimizing this quantity can be shown to be equivalent to satisfying at all times the set of equations given by

$$\int \widehat{W}\left(\vec{x}, t_j\right) F\left(\vec{x}, t_j\right) d\vec{x} = 0 \tag{20}$$

This is nearly identical to the constraint in equation [2] that:

$$\int W\left(\vec{x}, t_j\right) d\vec{x} = 0 \tag{21}$$

This latter criterion is identical to equation [20] in both the limit $F\left(\vec{x}, t_j\right) \rightarrow 1$, indicating dense sampling, and the limit $F\left(\vec{x}, t_j\right) \rightarrow 0$, indicating an absence of sampling since $W\left(\vec{x}, t_j\right)$ also becomes 0 in this limit. We choose to accept equation [2] as our fundamental constraint equation rather than equation [20]. This implies that our solution is only an approximation to the ordinary Kriging solution in the spatial mode; however, making this approximation confers several advantages. First, it ensures that $\hat{\theta}\left(t_j\right)$ and $\hat{b}_i$ retain their natural physical interpretation. Secondly, computational advantages are provided by isolating the $S_i\left(\vec{x}, t_j\right)$ so that the integrals might be performed independently for each station.

Given equations [7] and [13] imposing criterion [2] actually constrains the global average temperature $\hat{\theta}(t_j)$ nearly completely. Though not immediately obvious, constraints [7], [13] and [21] leave a single unaccounted for degree of freedom. Specifically one can adjust all $\hat{\theta}(t_j)$ by any arbitrary additive factor provided one makes a compensating adjustment to all $\hat{b}_i$. This last degree of freedom can be removed by specifying the climatology $C(\vec{x})$, applying the zero mean criterion from equation [2] and assuming that the local anomaly distribution (equation [5]) will also have mean 0. This implies:

$$C(\vec{x}_i) = \lambda(\vec{x}_i) + h(\vec{x}_i) + G(\vec{x}_i) \ \approx \hat{b}_i \tag{22}$$

We parameterize $h(\vec{x})$ as a simple quadratic function of elevation and parameterize $\lambda(\vec{x})$ as a piece-wise linear function of the absolute value of latitude with 11 knots equally spaced in the cosine of latitude. For $G(\vec{x})$ we reuse the Kriging formulation developed above, with a modification

$$\begin{pmatrix} B_1(\vec{x}) \\ \vdots \\ B_N(\vec{x}) \end{pmatrix} = \begin{pmatrix} \frac{1+(n_1-1)R(0)}{n_1} & R(\vec{x}_1, \vec{x}_2) & \cdots & R(\vec{x}_1, \vec{x}_N) \\ R(\vec{x}_2, \vec{x}_1) & \frac{1+(n_2-1)R(0)}{n_2} & & R(\vec{x}_2, \vec{x}_N) \\ \vdots & & \ddots & \vdots \\ R(\vec{x}_N, \vec{x}_1) & R(\vec{x}_N, \vec{x}_2) & \cdots & \frac{1+(n_N-1)R(0)}{n_N} \end{pmatrix}^{-1} \begin{pmatrix} R(\vec{x}, \vec{x}_1) \\ \vdots \\ R(\vec{x}, \vec{x}_N) \end{pmatrix} \tag{23}$$

$$\hat{G}(\vec{x}) = \sum_{i=1}^{N} B_i(\vec{x}) * \left(\hat{b}_i - \hat{\lambda}(\vec{x}) - \hat{h}(\vec{x})\right) \tag{24}$$

where n_i is the number of months of data for the i-th station. The modified diagonal terms on the correlation matrix are the natural effect of treating the value $\hat{b}_i$ as if it were entered into the Kriging process n_i times, which appropriately gives greater weight to values of $\hat{b}_i$ that are more precisely constrained. As noted previously the factors associated with latitude and altitude collectively capture ~95% of the variance in the stationary climatology field. Most of the remaining structure is driven by dynamical processes (e.g. ocean and atmospheric circulation) or by boundary conditions such as the nearness to an ocean.

This final normalization described here has the effect of placing the $\hat{\theta}(t_j)$ on an absolute scale such that these values are a true measure of mean temperature and not merely a measure of a temperature anomaly. In practice, we find that the normalization to an absolute scale is considerably more uncertain than the determination of relative changes in temperature. This occurs due to the large range of variations in $\hat{b}_i$ from nearly 30 C at the tropics to about -50 C in Antarctica. This large variability makes it relatively difficult to measure the spatial average temperature, and as a result there is more measurement uncertainty in the estimate of the absolute temperature normalization than there is in the measurement of changes over time.

The preceding outline explains the core of our analysis process. However, we make other modifications to address issues of bias correction and station reliability. Whereas other groups use a procedure they refer to as *homogenization*, our approach is different; we call it the *scalpel*.

4. Homogenization and the Scalpel

Temperature time series may be subject to many measurement artifacts and microclimate effects (Folland et al. 2001, Peterson and Vose 1997, Brohan et al. 2006, Menne et al. 2009, Hansen et al. 2001). Measurement biases often manifest as abrupt discontinuities arising from changes in instrumentation, site location, nearby environmental changes (e.g. construction), and similar artifacts. They can also derive from gradual changes in instrument quality or calibration, for example, fouling of a station due to accumulated dirt or leaves can change the station's thermal or air flow characteristics. In addition to measurement problems, even an accurately recorded temperature history may not provide a useful depiction of regional scale temperature changes due to microclimate effects at the station site that are not representative of large-scale climate patterns. The most widely discussed microclimate effect is the potential for "urban heat islands" to cause spuriously large temperature trends at sites in regions that have undergone urban development (Hansen et al. 2010, Oke 1982, Jones et al. 1990). At noted in the prior section, we estimate that on average 12% of the non-seasonal variance in a typical monthly temperature time series is caused by short-term local noise of one kind or another. All of the existing temperature analysis groups use processes designed to detect various discontinuities in a temperature time series and "correct" them by introducing adjustments that make the presumptively biased time series look more like neighboring time series and/or regional averages

(Menne and Williams 2009, Jones and Moberg 2003, Hansen et al. 1999). This data correction process is called "homogenization."

Rather than correcting data, we rely on a philosophically different approach. Our method has two components: 1) Break time series into independent fragments at times when there is evidence of abrupt discontinuities, and 2) Adjust the weights within the fitting equations to account for differences in reliability. The first step, cutting records at times of apparent discontinuities, is a natural extension of our fitting procedure that determines the relative offsets between stations, encapsulated by $\hat{b}_i$, as an intrinsic part of our analysis. We call this cutting procedure the *scalpel*. Provided that we can identify appropriate breakpoints, the necessary adjustment will be made automatically as part of the fitting process. We are able to use the scalpel approach because our analysis method can use very short records, whereas the methods employed by other groups generally require their time series be long enough to contain a reference interval.

The addition of breakpoints will generally improve the quality of fit provided they occur at times of actual discontinuities in the record. The addition of unnecessary breakpoints (i.e. adding breaks at time points which lack any real discontinuity), should be trend neutral in the fit as both halves of the record would then be expected to tend towards the same $\hat{b}_i$ value; however, unnecessary breakpoints can amplify noise and increase the resulting uncertainty in the record (discussed below).

There are in general two kinds of evidence that can lead to an expectation of a discontinuity in the data. The first is "metadata", such as documented station moves or instrumentation changes. For the current paper, the only "metadata" cut we use is based on gaps in the record; if a station failed to report temperature data for a year or more, then we consider that gap as evidence of a change in station conditions and break the time series into separate records at either side of the gap. In the future, we will extend the use of the scalpel to processes such as station moves and instrumentation changes; however, the analysis presented below is based on the GHCN dataset which does not provide the necessary metadata to make those cuts. The second kind of evidence requiring a breakpoint is an apparent shift in the statistical properties of the data itself (e.g. mean, variance) when compared to neighboring time series that are expected to be highly correlated. When such a shift is detected, we can divide the data at that time, making what we call an "empirical breakpoint". The detection of empirical breakpoints is a well-developed field in statistics (Page 1955, Tsay 1991, Hinkley 1971, Davis 2006), though relatively little work has been done to develop the case where spatially correlated data are widely available. As a result, the existing groups have each developed their own approach to empirical change point detection (Menne and Williams 2009; Jones and Moberg 2003, Hansen et al. 1999). In the present paper, we use a simple empirical criterion that is not intended to be a complete study of the issue. Like prior work, the present criterion must be applied prior to any averaging. In principle, change point detection could be incorporated into an iterative averaging process that uses the immediately preceding average to help determine a set of breakpoints for the next iteration; however, no such work has been done at present. For the present paper, we follow NOAA in considering the neighborhood of each station and identifying the most highly correlated adjacent stations. A local reference series is then constructed by a weighted average of the neighboring stations. This is compared to the station's records, and a breakpoint is introduced at places where there is an abrupt shift in mean larger than 4 standard deviations. This empirical technique results in approximately 1 cut for every 12.2 years of record, which is somewhat more than the changepoint occurrence rate of one every 15-20 years reported by

Menne et al. 2009. Future work will explore alternative cut criteria, but the present effort is meant merely to incorporate the most obvious change points and show how our averaging technique can incorporate the discontinuity adjustment process in a natural way.

5. Outlier Weighting

The next potential problem to consider is point outliers, i.e. single data points that vary greatly from the expected value as determined by the local average. Removal of outliers is done by defining the difference between a temperature stations report and the expected value at that same site:

$$\Delta_i(t_j) = d_i(t_j) - \hat{b}_i - \hat{\theta}(t_j) - \widehat{W}^{\dagger}(\vec{x}_i, t_j) \qquad [25]$$

where $\widehat{W}^{\dagger}(\vec{x}_i, t_j)$ approximates the effect of constructing the $\widehat{W}(\vec{x}_i, t_j)$ field without the influence of the i-th station:

$$\widehat{W}^{\dagger}(\vec{x}_i, t_j) = \widehat{W}(\vec{x}_i, t_j) - S_i(\vec{x}_i, t_j)(d_i(t_j) - \hat{b}_i - \hat{\theta}(t_j)) \qquad [26]$$

The scale of the typical measurement error ($e \approx 0.55$ C) is estimated from:

$$e^2 = \frac{\sum_{i,j}\left(\Delta_i(t_j)\right)^2}{\sum_{i,j} 1} \qquad [27]$$

The outlier weight adjustment is defined as

$$O_{i,j} = \begin{cases} 1 & \text{if } \left(\Delta_i(t_j)\right)^2 \leq (2.5e)^2 \\ 2.5e/\left|\Delta_i(t_j)\right| & \text{otherwise} \end{cases} \qquad [28]$$

Equation [28] specifies a downweighting term to be applied for point outliers that are more than $2.5e$ from the modeled expectation. This outlier weighting is used to define a modified expression for $\hat{b}_i$:

$$\hat{b}_i^{\,*} = \frac{\sum_j O_{i,j}\left(d_i(t_j) - \hat{\theta}(t_j) - \widehat{W}(\vec{x}_i, t_j)\right)}{\sum_j O_{i,j}} \qquad [29]$$

and also incorporated into the site weighting discussed below.

This choice of target threshold, $2.5e$, is partly arbitrary but was selected with the expectation that most of the measured data should be unaffected. If the underlying data fluctuations were normally distributed, we would expect this process to crop 1.25% of the data. In practice, we observe that the data fluctuation distribution tends to be intermediate between a normal distribution and a Laplace distribution. In the Laplace limit, we would expect to crop 2.9% of the data, so the actual exclusion rate can be expected to be intermediate between 1.25% and 2.9% for the typical station record.

Of course, the goal is not to remove legitimate data, but rather to limit the impact of erroneous outliers. In defining equation [28], we adjusted the weight of outliers to a fixed target, $2.5e$, rather than to simply downweight them to zero. This helps to ensure numerical stability.

6. Reliability Weighting

In addition to point outliers, climate records often vary for other reasons that can affect an individual record's reliability at the level of long-term trends. For example, we also need to consider the possibility of gradual biases that lead to spurious trends. In this case we assess the overall "reliability" of the record by measuring each record's average level of agreement with the expected field $\hat{T}(\vec{x}, t)$ at the same location.

For each station we compute a measure of the quality of fit:

$$e_i^2 = \frac{\sum_j \min\left\{\left(\Delta_i(t_j)\right)^2, 25e^2\right\}}{\sum_j 1} \qquad [30]$$

The "min" is used to avoid giving too great a weight to the most extreme outliers when judging the reliability of the series. The station weight is then defined as:

$$\omega_i = \frac{2e^2}{e^2 + e_i^2} \qquad [31]$$

Due to the limits on outliers from the previous section, the station weight has a range between 1/13 and 2, effectively allowing a "perfect" station record to receive up to 26 times the weight of a "terrible" record. This functional form was chosen for the station weight due to several desirable qualities. The typical record is expected to have a weight near 1, with poor records being more severely downweighted than good records are enhanced. Using a relationship that limits the potential upweighting of good records was found to be necessary in order to ensure efficient convergence and numerical stability. A number of alternative weighting and functional forms with similar properties were also considered, but we found that the construction of global temperature time series were not very sensitive to the details of how the downweighting of inconsistent records was handled.

After defining the station weight, we need to incorporate this information into the spatial averaging process, e.g. equation [13], by adjusting the associated Kriging coefficients. Ideally, one might use the station weights to modify the correlation matrix (equation [12]) and recompute the Kriging coefficients. However, it is unclear what form of modification would be appropriate, and frequent recomputation of the required matrix inverses would be computationally impractical. So, we opted for a more direct approach to the reweighting of the Kriging solution. We define updated spatial averaging coefficients:

$$S_i^*(\vec{x}, t_j) = \frac{\omega_i S_i(\vec{x}, t_j)}{\left(\sum_m \omega_m S_m(\vec{x}, t_j)\right) + \left(1 - F(\vec{x}, t_j)\right)} \qquad [32]$$

This expression is motivated by the representation of the true anomaly in equation [17] as:

$$\hat{\theta}(t_j) + \widehat{W}(\vec{x}, t_j) = \left(1 - F(\vec{x}, t_j)\right)\hat{\theta}(t_j) + \sum_i S_i(\vec{x}, t_j)\left(d_i(t_j) - \hat{b}_i\right) \qquad [33]$$

and the desire to leave the expected variance of the right hand side unchanged after reweighting. Because $F(\vec{x}, t_j) = \sum_m S_m(\vec{x}, t_j)$ it follows that $S_i^*(\vec{x}, t_j)$ is equal to $S_i(\vec{x}, t_j)$ if all the station weights are set to 1. The $\left(1 - F(\vec{x}, t_j)\right)$ term in the denominator can be understood as measuring the influence of the global mean field, rather than the local data, in the construction of the local average temperature estimate. The omission of this term in equation [32] would lead to a weighting scheme that is numerically unstable.

It is important to note that equation [32] merely says that the local weather average $\widehat{W}(\vec{x}, t_j)$ should give proportionally greater weight to more reliable records. However, if all of the records in a given region have a similar value of ω_i, then they will all receive about the same weight regardless of the actual numerical value of ω_i. Specifically, we note ω_i does not directly influence $\hat{\theta}(t_j)$. This behavior is important as some regions of the Earth, such as Siberia, tend to have broadly lower values of ω_i due to the high variability of local weather conditions. However, as long as all of the records in a region have similar values for ω_i, then the individual stations will still receive equal and appropriate weight in the global average. This avoids a potential problem that high variability regions could be underrepresented in the construction the global time series $\hat{\theta}(t_j)$.

As noted above, the formulation of equation [32] is not necessarily ideal compared to processes that could adjust the correlation matrix directly, and hence this approach should be considered as an approximate approach for incorporating station reliability differences. In particular, the range bounds shown for $S_i(\vec{x}, t_j)$, such as that given for equation [16], will not necessarily hold for $S_i^*(\vec{x}, t_j)$.

Equation [32] leads to a natural expression for the outlier and reliability adjusted weather field

$$\widehat{W}^*(\vec{x}, t_j) = \sum_{i=1}^{N} O_{i,j} S_{a_i}{}^*(\vec{x}, t_j)\left(d_{a_i}(t_j) - \hat{\theta}(t_j) - \hat{b}_{a_i}{}^*\right) \qquad [34]$$

$\hat{b}_i{}^*$ and $\widehat{W}^*(\vec{x}, t_j)$ are now used to replace the original values in the execution of the model. In order to ensure robustness, this process of determining site and outlier weights is repeated many times until the parameter values stabilize. We find that we typically require 10 to 30 iterations to satisfy our convergence criteria.

Implicit in the discussion of station reliability considerations are several assumptions. Firstly, we assume that the local weather function constructed from many station records, $\widehat{W}(\vec{x}, t_j)$, will be a better estimate of the local temperature than any individual record could be. This assumption is generally characteristic of all averaging techniques; however, we can't rule out the possibility of large scale systematic biases. Our reliability adjustment techniques can work well when one or a few records are noticeably inconsistent with their neighbors, but large scale biases affecting many stations could cause such comparative estimates to fail. Secondly, we assume that the reliability of a station is largely invariant over time. This will in general be false; however, the scalpel procedure discussed previously will help us here. By breaking records into multiple pieces on the basis of metadata changes and/or empirical discontinuities, we then also have the opportunity to assess the reliability of each fragment individually. A detailed comparison and contrast of our results with those obtained using other approaches to deal with inhomogeneous data will be presented elsewhere.

7. Uncertainty Analysis

We consider there to be two essential forms of quantifiable uncertainty in the Berkeley Earth averaging process:

1. Statistical / Data-Driven Uncertainty: This is the error made in estimating the parameters $\hat{b}_i$ and $\hat{\theta}(t_j)$ due to the fact that the data, $d_i(t_j)$, may not be an accurate reflection of the true temperature changes at location $\vec{x}_i$.
2. Spatial Incompleteness Uncertainty: This is the expected error made in estimating the true land-surface average temperature due to the network of stations having incomplete coverage of all land areas.

In addition, there is "structural" or "model-design" uncertainty, which describes the error a statistical model makes compared to the real-world due to the design of the model. Given that it is impossible to know absolute truth, model limitations are generally assessed by attempting to validate the underlying assumptions that a model makes and comparing those assumptions to other approaches used by different models. For example, we use a site reliability weighting procedure to reduce the impact of anomalous trends (such as those associated with urban heat islands), while other models (such as those developed by GISS) attempt to remove anomalous trends by applying various corrections. Such differences are an important aspect of model design. In general, it is impossible to directly quantify structural uncertainties, and so they are not a factor in our standard uncertainty model. However, one may be able to identify model limitations by drawing comparisons between the results of the Berkeley Average and the results of other groups. Discussion of our results and comparison to those produced by other groups will be provided below.

Another technique for identifying structural uncertainty is to run the same model on multiple data sets that differ primarily based on factors that one suspects may give rise to unaccounted for model errors. For example, one can perform an analysis of rural data and compare it to an analysis of urban data to look for urbanization biases. Such comparisons tend to be non-trivial since it is rare that one can construct data sets that isolate the experimental variables without introducing other confounding variations. We will not provide any such analysis of such experiments in this paper; however, additional papers submitted by our group (Wickham et al. submitted; Muller et al. submitted) find that objective measures of station quality and urbanization do not have with a statistically significant impact on our results over most of the available record. In other words, the averaging techniques combined with the bias adjustment procedures we have described appear adequate for dealing with those data quality issues to within the limits of the uncertainties that nonetheless exist from other sources. The one possible exception is that Wickham et al. observed that rural stations may slightly overestimate global land-surface warming during the most recent decade. The suggested effect is small and opposite in sign to what one would expect from an urban heat island bias. At the present time we are not incorporating any explicit uncertainty to account for such factors, though the data driven uncertainty will implicitly capture the effects of variations in data behavior across the field.

The other analysis groups generally discuss a concept of "bias error" associated with systematic biases in the underlying data (e.g. Brohan et al. 2006; Smith and Reynolds 2005). To a degree these concepts overlap with the discussion of "structural error" in that the prior authors tend to add extra uncertainty to account for factors such as urban heat islands and instrumental changes in cases when they do not directly model them. Based on graphs produced by HadCRU, such "bias error" was considered to be a negligible portion of total error during the critical 1950-

2010 period of modern warming, but leads to an increase in total error up to 100% circa 1900 (Brohan et al. 2006). In the current presentation we will generally ignore these additional uncertainties, which will be discussed once future papers have examined the various contributing factors individually.

8. Statistical Uncertainty – Overview

Statistical uncertainty is a reflection of the errors introduced into the determination of model parameters due to the fact that the basic data, $d_i(t_j)$, may not be an accurate reflection of the true temperature history. In order to place uncertainties on the global mean temperature time series $\hat{\theta}(t_j)$, we apply two approaches, a systematic "sampling" method, and a "jackknife" method (Miller 1974, Tukey 1958, Quenouille 1949).

These approaches are both different from the approaches that have been commonly used in the past. Prior groups generally assess uncertainly from the bottom-up by assigning uncertainty to the initial data and all of the intermediate processing steps. This is a complicated process due to the possibility of correlated errors and the risk that those uncertainties may interact in unexpected ways. Further, one commonly applies the same amount of data uncertainty to all records, even though we would expect that some records are more accurate than others.

As an alternative, we approach the statistical uncertainty quantification from a top-down direction. At its core, this means measuring how much our result would change if there were variations in the amount of data available. By performing the entire analysis chain with small variations in the amount of data available we can assess the impact of data noise in a way that bypasses concerns over correlated error and varying record uncertainty. For a complex analysis system this will generally provide a more accurate measure of the statistical uncertainty, though there are some additional nuances.

9. Statistical Uncertainty – Sampling Method

The sampling method we apply relies on subsampling the station network, recomputing the temperature time series, and examining the variance in the results across the different samples. In the implementation we used for the current paper, each station is randomly assigned to one of five groups. Each of these groups can be expected to have similar, but somewhat diminished, spatial coverage compared to the complete sample. For each group of stations we reapply the averaging process. This leads to a set of new temperature time series $\hat{\theta}_n(t_j)$, where the *n* index denotes the subsample number. As each of these new time series is created from a completely independent station network, we are justified in treating their results as statistically independent.

For each subsampled network, we compute the mean temperature for an arbitrary period, e.g. Jan 1950 to Dec 2000, and subtract this from the data; this gives us five subsampled records that have the same temperature "anomaly." We do this to separate out the uncertainty associated with relative changes in the global land-surface time series from the larger uncertainty associated with the estimation of the Earth's absolute mean temperature. We then estimate the statistical uncertainty of $\hat{\theta}(t_j)$ as the standard error in the mean of the subsampled values, namely

$$\sigma_{\text{sampling}}(t_j) = \sqrt{\frac{\sum_n \left(\hat{\theta}_n(t_j) - \langle \hat{\theta}_n(t_j) \rangle\right)^2}{\sum_n 1}} \qquad [35]$$

Where $\langle \hat{\theta}_n(t_j) \rangle$ denotes the mean value. In general, the denominator will be 5 at times where all five subsamples report a value. However, since the different subsamples may have somewhat different time coverage, the number of records reported at early times may be different. We require at least three subsamples report a value in order for an uncertainty to be reported. Examples of subsampled temperature series and the resulting uncertainty will be provided with the discussion of GHCN results.

The sampling value could be further refined. One method would be to repeat this entire process of creating five subsamples through multiple iterations and average the results.

Unfortunately, though conceptually simple and computationally efficient the sampling method suffers from a flaw that leads to a systematic underestimation of the statistical uncertainty in our context. In forming each subsampled network, 80% of stations must be eliminated. This increases the effect of spatial uncertainty associated with each of these subsamples. Further, due to the highly heterogeneous history of temperature sampling the newly unsampled regions in each subnetwork will tend to overlap to a substantial degree leading to correlated errors in the uncertainty calculation. Based on a variety of Monte Carlo experiments, we concluded that the sampling estimates of uncertainty tend to understate the true error by between 10 and 100% depending on the distribution of the temperature monitoring network at the time.

10. Statistical Uncertainty – Jackknife Method

The "jackknife", a method developed by Quenoille and John Tukey, is our primary method for determining statistical uncertainty (Tukey 1958, Quenoille 1949, Miller 1974). It is a special modification of the sampling approach, finding its traditional use when the number of data points is too small to give a good result using ordinary sampling. Given the fact that we have many thousands of stations in our records, each with typically hundreds of data points, it was surprising to us that this method would prove so important. But despite our large set of data, there are time and places that are sparsely sampled. As noted above, the presence of this sparse sampling tends to cause the sampling technique to underestimate the statistical uncertainty.

We use the jackknife method in the following way. Given a set of stations (7280, when using the GHCN compilation) we construct 8 station groups, each consisting of 7/8 of the data, with a different 1/8 removed from each group. The data from each of these data samples is then run through the entire Berkeley Average machinery to create 8 records $\hat{\theta}_k(t_j)$ of average global land temperature vs. time. Following Quenouille and Tukey, we then create a new set of 8 "effectively independent" temperature records $\hat{\theta}_k^{\dagger}(t_i)$ by the jackknife formula

$$\hat{\theta}_k^{\dagger}(t_i) = 8\,\hat{\theta}_k(t_j) - 7\,\hat{\theta}(t_j) \qquad [36]$$

where $\hat{\theta}(t_j)$ is the reconstructed temperature record from the full (100%) sample. Hence we calculate the standard error among the effectively independent samples:

$$\sigma_{\text{jackknife}}(t_j) = \sqrt{\frac{\sum_k \left(\hat{\theta}_k^{\dagger}(t_j) - \langle \hat{\theta}_k^{\dagger}(t_j) \rangle\right)^2}{\sum_k 1}} \qquad [37]$$

We indeed found that the typical statistical uncertainties estimated from the jackknife were, in general, larger than those estimated from the sampling method. As the jackknife constructs its temperature average using a station network that is nearly complete, it is more robust against spatial distribution effects. In addition, we can more easily increase the number of samples without worrying that the network would become too sparse (as could happen if one increased the number of divisions in the sampling approach).

We studied the relative reliability of the sampling and jackknife methods using over 10,000 Monte Carlo simulations. For each of these simulations, we created a toy temperature model of the "Earth" consisting of 100 independent climate regions. We simulated data for each region, using a distribution function that was chosen to mimic the distribution of the real data; so, for example, some regions had many sites, but some had only 1 or 2. This model verified that sparse regions caused problems for the sampling method. In these tests we found that the jackknife method gave a consistently accurate measure of the true error (known since in the Monte Carlo we knew the "truth") while the sampling would consistently underestimate the true error.

When we discuss the results for our reanalysis of the GHCN data we will show the error uncertainties calculated both ways. The jackknife uncertainties are larger than those computed via sampling, but based on our Monte Carlo tests, we believe them to be more accurate.

11. Spatial Uncertainty

Spatial uncertainty measures the amount of error that is likely to occur in our averages due to incomplete sampling of land surface areas. Our primary technique in this case is empirical. We look at the sampled area available at past times, superimpose it on the modern day, and ask how much error would be incurred in measuring the modern temperature field given only the limited sample area available in the past. For example, if one only knew the temperature anomalies for Europe and North America, how much error would be incurred by using that measurement as an estimate of the global average temperature anomaly? The process for making this estimate involves applying the coverage field, $F(\vec{x}, t_j)$, that exists at each time and superimposing it on the nearly complete temperature anomaly fields $\widehat{W}(\vec{x}, t_j)$ that exist for late times, specifically $1960 \leq t_j \leq 2000$ when spatial land coverage approached 100%. We define the estimated average weather anomaly at time t_m based on the sample field available at time t_j to be:

$$\tau(t_j, t_m) = \frac{\int F(\vec{x}, t_j)\widehat{W}(\vec{x}, t_m)d\vec{x}}{\int F(\vec{x}, t_j)d\vec{x}} \tag{38}$$

And then define the spatial uncertainty in $\hat{\theta}(t_j)$ as:

$$\sigma_{\text{spatial}}(t_j) = \sqrt{\frac{\sum_{t_m=1960}^{2000}\left(\tau(t_j, t_m) - \tau(t_m, t_m)\right)^2}{\sum_{t_m=1960}^{2000} 1}} \tag{39}$$

Ideally $F(\vec{x}, t_j)$ would be identically 1 during the target interval $1960 \leq t_j \leq 2000$ used as a calibration standard, which would imply that $\tau(t_m, t_m) = 0$, via equation [21]. However, in practice these late time fields are only 90-98% complete. As a result, $\sigma_{\text{spatial}}(t_j)$ computed via this process will tend to slightly underestimate the uncertainty at late times.

An alternative is to use the correlated error propagation formula:

$$\sigma_{\text{spatial}}(t_j) \approx \sqrt{\int\int\left(1 - \frac{F(\vec{x}, t_j)}{F_{land}(t_j)}\right)\left(1 - \frac{F(\vec{y}, t_j)}{F_{land}(t_j)}\right)\hat{V}(\vec{y})\hat{V}(\vec{x})R(\vec{x}, \vec{y})d\vec{x}d\vec{y}} \tag{40}$$

Where $R(\vec{x}, \vec{y})$ is the correlation function estimated in equation [14], $F_{land}(t_j)$ is the spatial completeness factor defined in equation [18], and $\hat{V}(\vec{x})$ is square root of the variance at $\vec{x}$ estimated as:

$$H(\vec{x}, t_j) = \begin{cases} F(\vec{x}, t_j) & \text{if } F(\vec{x}, t_j) \geq 0.4 \\ 0 & \text{otherwise} \end{cases} \tag{41}$$

$$\hat{V}(\vec{x}) = \sqrt{\frac{\sum_j H(\vec{x}, t_j)\left(\frac{\widehat{W}(\vec{x}, t_j)}{F(\vec{x}, t_j)}\right)^2}{\sum_j H(\vec{x}, t_j)}} \tag{42}$$

The new symbol $H(\vec{x}, t_j)$ is introduced to focus the estimates of local variance on only those times when at least 40% of the variance has been determined by the local data. In addition, the term $\frac{\widehat{W}(\vec{x}, t_j)}{F(\vec{x}, t_j)}$ provides a correction to the magnitude of the fluctuations in $\widehat{W}(\vec{x}, t_j)$ in the presence of incomplete sampling. Recall that $\widehat{W}(\vec{x}, t_j) \to 0$ as $F(\vec{x}, t_j) \to 0$, which reflects the fact that there can be no knowledge of the local fluctuations in the field when no data is available in the local neighborhood.

The estimate of $\sigma_{\text{spatial}}(t_j)$ from equation [39] tends to be 30-50% larger than the result of equation [40] at early times (e.g. pre-1940). We believe this is because the linearized error propagation formula in equation [40] and the approximate correlation function estimated in equation [14] don't capture enough of the structure of the field, and that the formulation in equation [39] is likely to be superior at early times. At late times the two results are nearly identical; however, both estimates of the uncertainty due to spatial incompleteness at late times tend be far lower than the statistical uncertainty at late times. In other words, at times where the spatial coverage of the Earth's land surface is nearly complete, the uncertainty is dominated by statistical factors rather the spatial ones.

As noted above, the empirical uncertainty estimate of equation [39] is partially limited due to incomplete sampling during the target interval. To compensate for this we add a small analytical correction, determined via equation [40] in the computation of our final spatial uncertainty estimates at regions with incomplete sampling. This correction is essentially negligible except at late times.

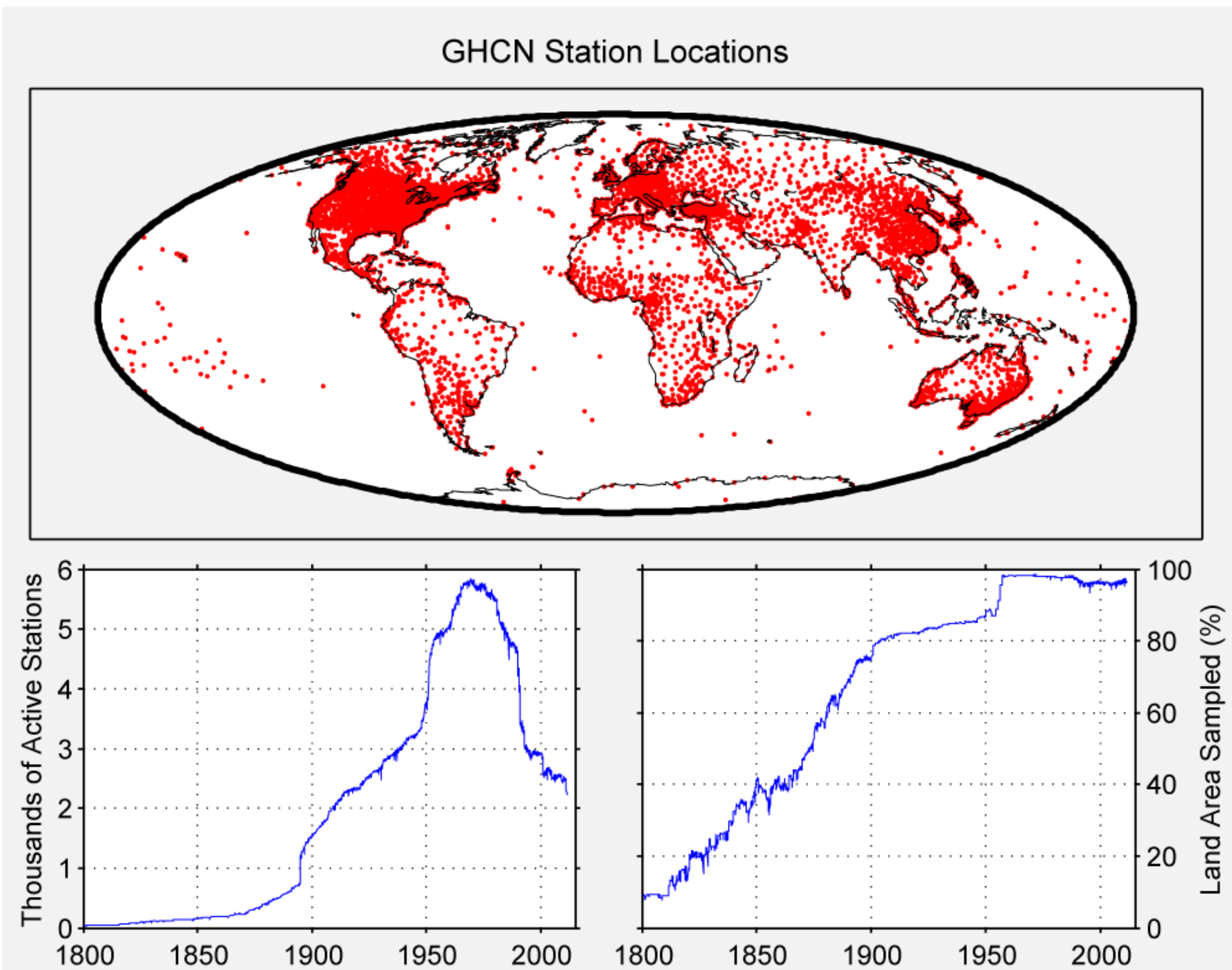
12. GHCN Results

The analysis method described in this paper has been applied to the 7280 weather stations in the Global Historical Climatology Network (GHCN) monthly average temperature data set developed by Peterson and Vose 1997; Menne and Williams 2009. We used the non-homogenized data set, with none of the NOAA corrections for inhomogeneities included; rather, we applied our scalpel method to break records at any documented discontinuity. We used the empirical scalpel method described earlier to detect undocumented changes; using this, the original 7,280 data records were broken into 47,282 record fragments. Of the 30,590 cuts, 5218 were based on gaps in record continuity longer than 1 year and the rest were found by our empirical method. We also found a small number of nonsense data points in the raw data, for example, values exceeding 70 C, records filled with zeros, or other repeated strings of data; these were eliminated by a pre-filtering process. In total, 0.8% of the data points were eliminated for such reasons. The NOAA analysis process uses their own pre-filtering in their homogenization and averaging processes, but we chose to handle them directly due to our preference for using the raw GHCN data with no prior corrections. A further 0.2% of data was eliminated because after cutting and filtering the resulting record was either too short to process (minimum length ≥6 months) or it occurred at a time with fewer than 5 total stations active.

It is worth making a special point of noting that after cutting and processing, the median length of a temperature time series processed by the Berkeley Average was only 7.1 years. Further, the inner 50% range for station record lengths was 2.7 to 12.8 years. As already stated, our climate change analysis system is designed to be very tolerant of short and discontinuous records which will allow us to work with a wider variety of data than is conventionally employed.

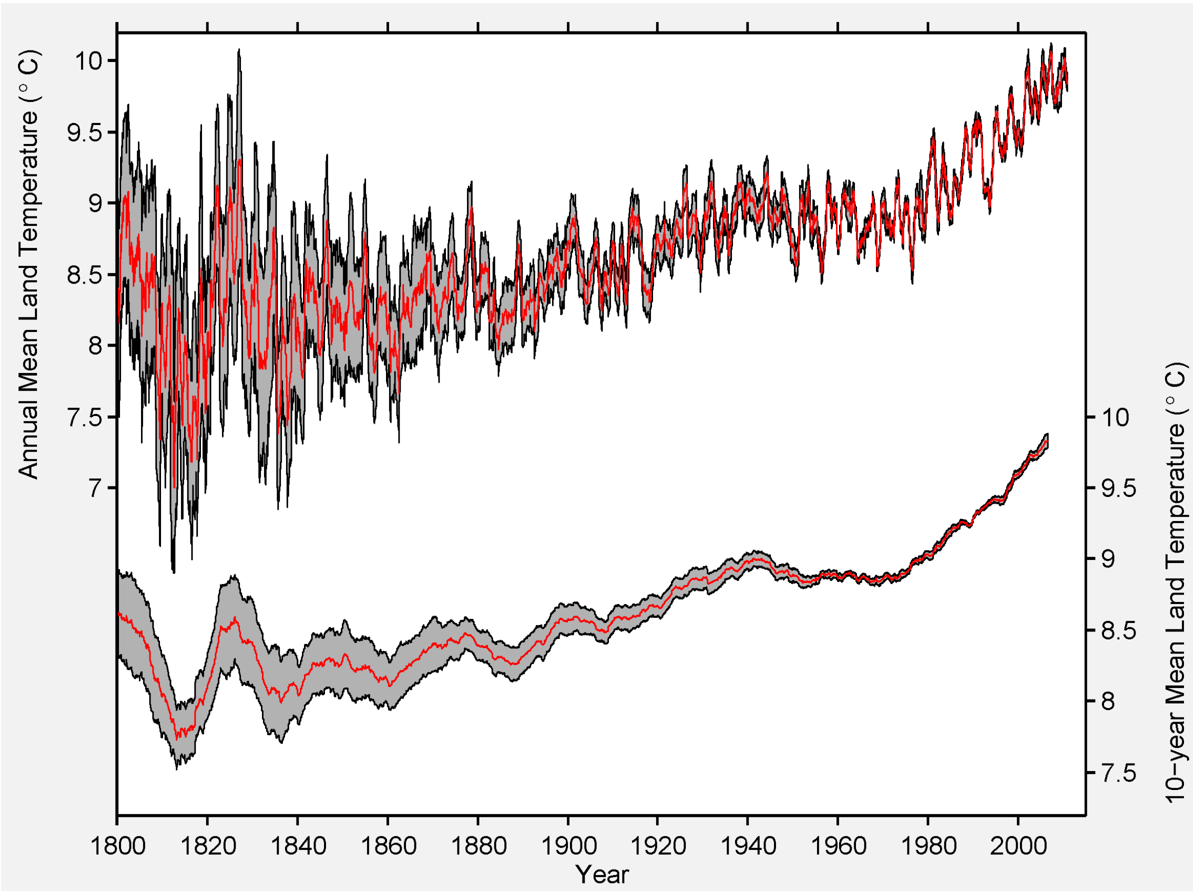
Figure 4 shows the station locations used by GHCN, the number of active stations vs. time, and the land area sampled vs. time (calculated using the method described in equation [18]). The sudden drop in the number of stations ca. 1990 is largely a result of the methodology used in compiling the GHCN dataset; GHCN generally only accepts records for stations that explicitly issue a monthly summary report however many stations have stopped reporting monthly results and only reported daily ones. Despite this drop, Figure 4(c) shows that the coverage of the Earth's land surface remained above 95%, reflecting the broad distribution of the stations that did remain.

Figure 4. Station locations for GHCN dataset, number of active stations over time, and percentage of the Earth's land area sampled



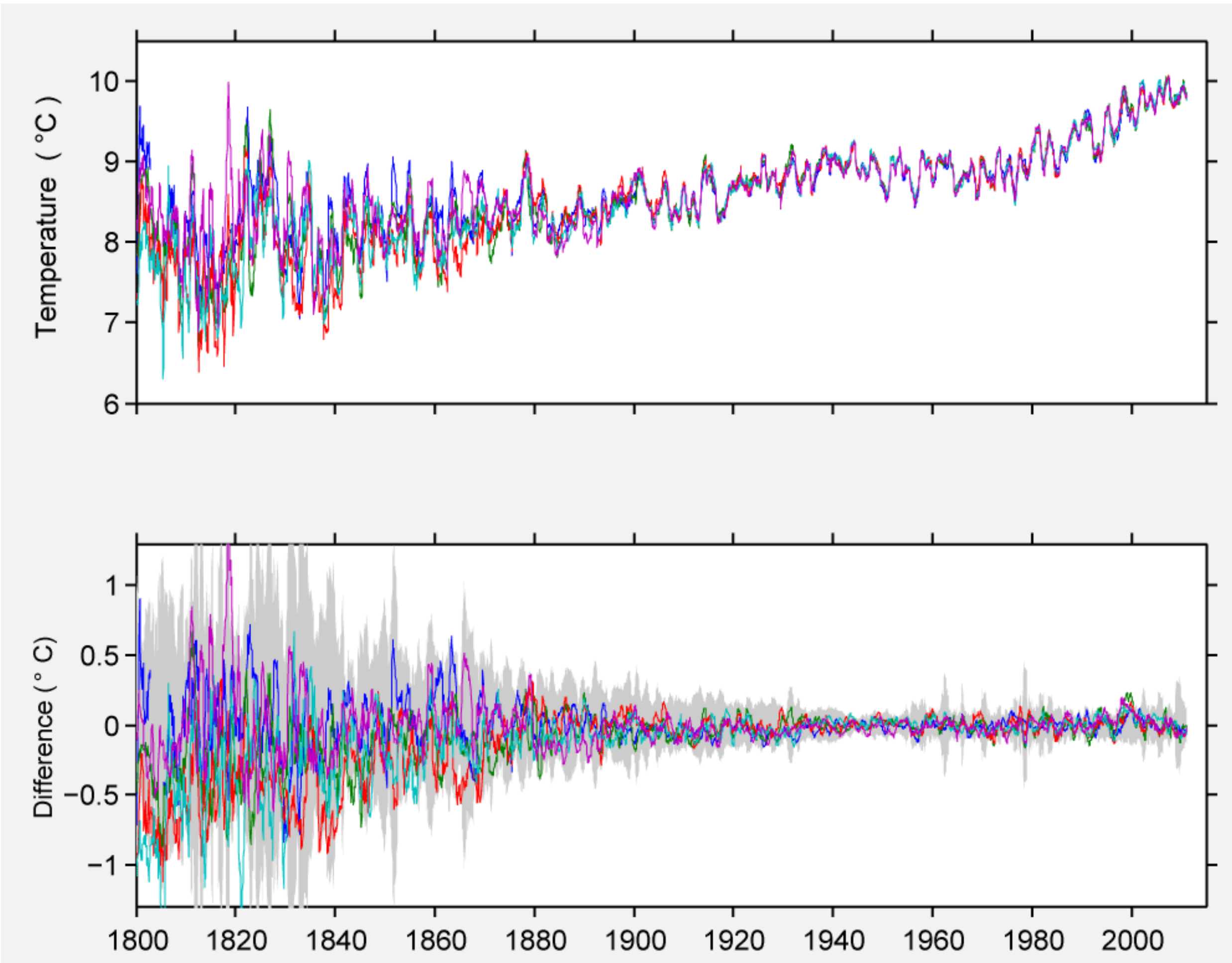
We applied the Berkeley Average methodology to the GHCN monthly data. The results and associated uncertainties are shown in Figure 5. The upper plot shows the 12-month land-only moving average and its associated 95% uncertainty; the lower plot shows the result of applying a 10-year moving average. Applying the methods described here, we find that the average land temperature from Jan 1950 to Dec 1959 was 8.849 ± 0.033 C, and temperature average during the most recent decade (Jan 2000 to Dec 2009) was 9.760 ± 0.041 C, an increase of 0.911 ± 0.042 C. The trend line for the 20th century is calculated to be 0.733 ± 0.096 C/century, well below the 2.76 ± 0.16 C/century rate of global land-surface warming that we observe during the interval Jan 1970 to Aug 2011. (All uncertainties quoted here and below are 95% confidence intervals for the combined statistical and spatial uncertainty). Though it is sometimes argued that global warming has abated since the 1998 El Nino event (e.g. Easterling and Wehner 2009, Meehl et al. 2011), we find no evidence of this in the GHCN land data. Applying our analysis over the interval 1998 to 2010, we find the land temperature trend to be 2.84 ± 0.73 C / century, consistent with prior decades. Meehl et al. (2011) associated the recent decreases in global temperature trends with increased heat flux into the deep oceans. The fact that we observe no change in the trend over land would seem to be consistent with the conclusion that any change in the total global average has been driven solely with oceanic processes.

Figure 5. Result of the Berkeley Average Methodology applied to the GHCN monthly data



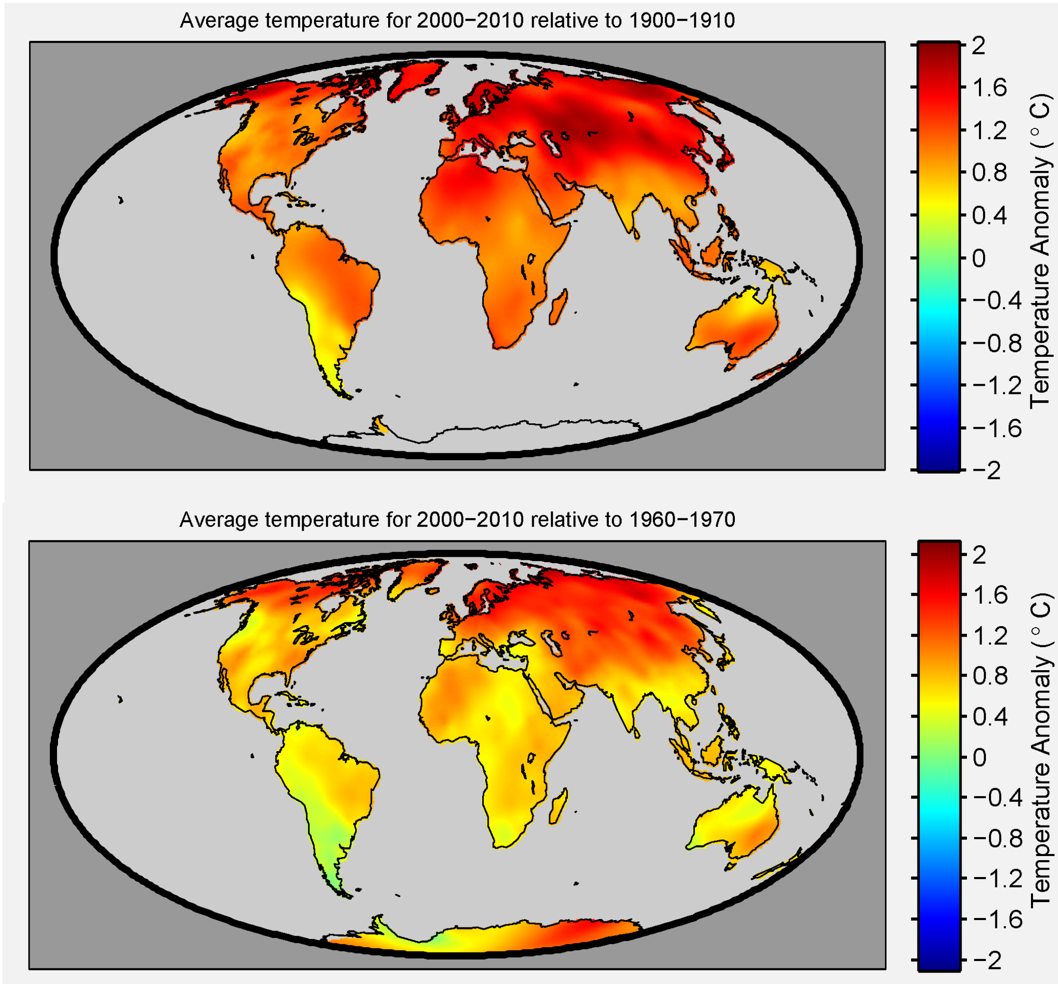
In the section on the sampling method, we discussed the determination of statistical uncertainties by dividing the full data set into five subsamples. In Figure 6 below we show the results of doing this for the GHCN data set. We show this primarily because the sampling method is more intuitive for many people than is the jackknife, and the charts in Figure 6 make it clear why the statistical uncertainties are small. The five completely independent subsamples produce very similar temperature history when processed via the Berkeley Average methodology.

Figure 6. **Five independent temperature reconstructions**



The spatial structure of the climate change during the last century is shown in Figure 7 and found to be fairly uniform, though with greater warming over the high latitudes of North America and Asia, consistent with prior results (Hansen et al. 2010). We also show the pattern of warming since the 1960s, as this is the period during which anthropogenic effects are believed to have been the most significant. Warming is observed to have occurred over all continents, though parts of South America are consistent with no change. No part of the Earth's land surface shows appreciable cooling.

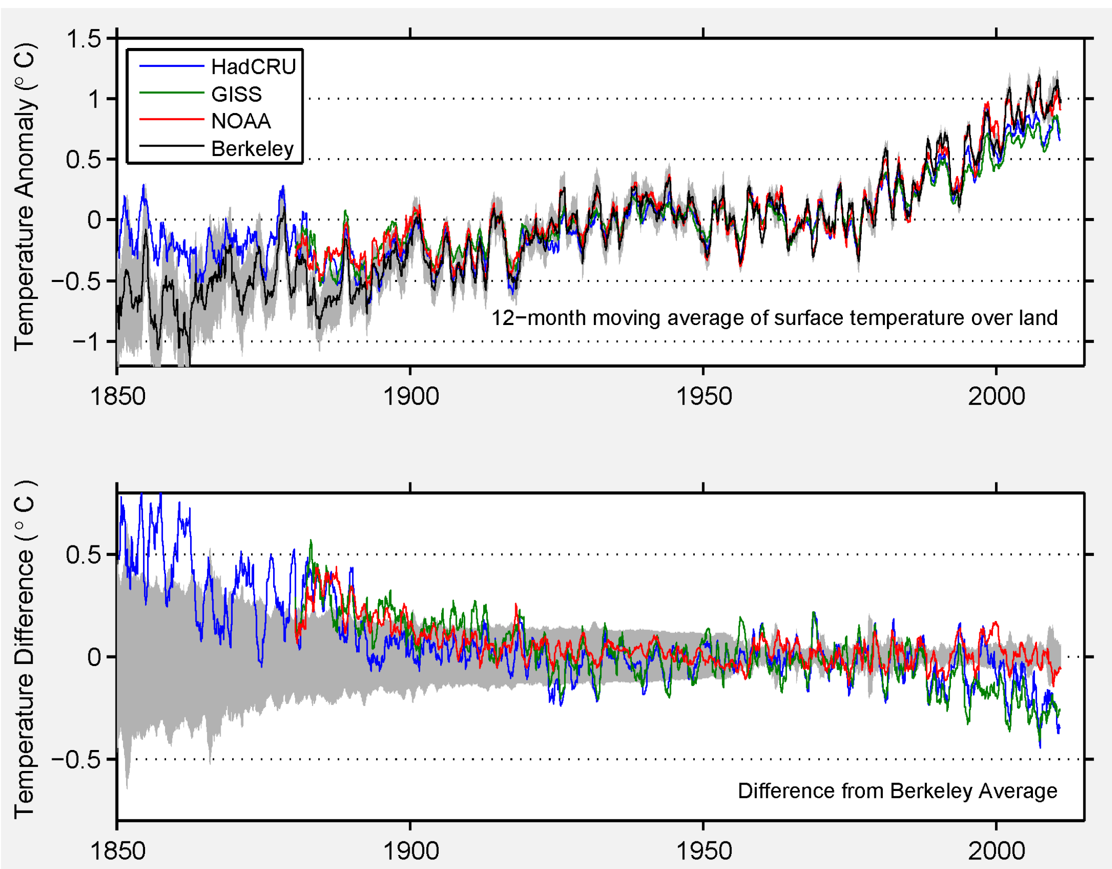
Figure 7. Maps showing the decadal average changes in the land temperature field



In Figure 8, we compare our land reconstruction to the land reconstructions published by the three other groups (results updated online, methods described by Brohan et al. 2006; Smith et al. 2008; Hansen et al. 2010). Overall our global land average is similar to those obtained by these prior efforts. There is some disagreement amongst the three groups, and our result is most similar overall to NOAA's work. The differences apparent in Figure 8 may partially reflect difference in source data, but they probably primarily reflect differences in methodology.

The GHCN dataset used in the current analysis overlaps strongly with the data used by other groups. The GHCN was developed by NOAA and is the sole source of the land-based weather station data in their temperature reconstructions (but does not include the ocean data also used in their global temperature analyses). In addition, GISS uses GHCN as the source for ~85% of the time series in their analysis. The remaining 15% of GISS stations are almost exclusively US and Antarctic sites that they have added / updated, and hence would be expected to have somewhat limited impact due to their limited geographic coverage. HadCRU maintains a separate data set from GHCN for their climate analysis work though approximately 60% of the GHCN stations also appear in HadCRU.

Figure 8. Comparison of the Berkeley Average to existing land-only averages reported



The GISS and HadCRU work produce lower land-average temperature trends for the late part of the 20^{th} century. In this regard, our analysis suggests a degree of global land-surface warming during the anthropogenic era that is consistent with prior work (e.g. NOAA) but on the high end of the existing range of reconstructions. We note that the difference in land average trends amongst the prior groups has not generally been discussed in the literature. In part, the spread in existing land-only records may have received little attention because the three groups have greater agreement when considering global averages that include oceans (Figure 1). We strongly suspect that some of the difference in land-only averages is an artifact of the different approaches to defining "land-only" temperature analyses. Our analysis and that produced by NOAA explicitly construct an average that only considers temperature values over land. However, that is not the only possible approach. The literature suggests that the GISS "land-only" data product may be generated by measuring the "global" temperature fields using only data reported over land. In this scenario temperature records in coastal regions and on islands would be extrapolated over the oceans to create a "global" field using only land data. Whether or not this approach was actually used is unclear from the literature, but it would result in an overweighting of coastal and oceanic stations. This would in turn lead to a reduction in the calculated "land" trend in a way that is qualitatively consistent with the difference observed in Figure 8.

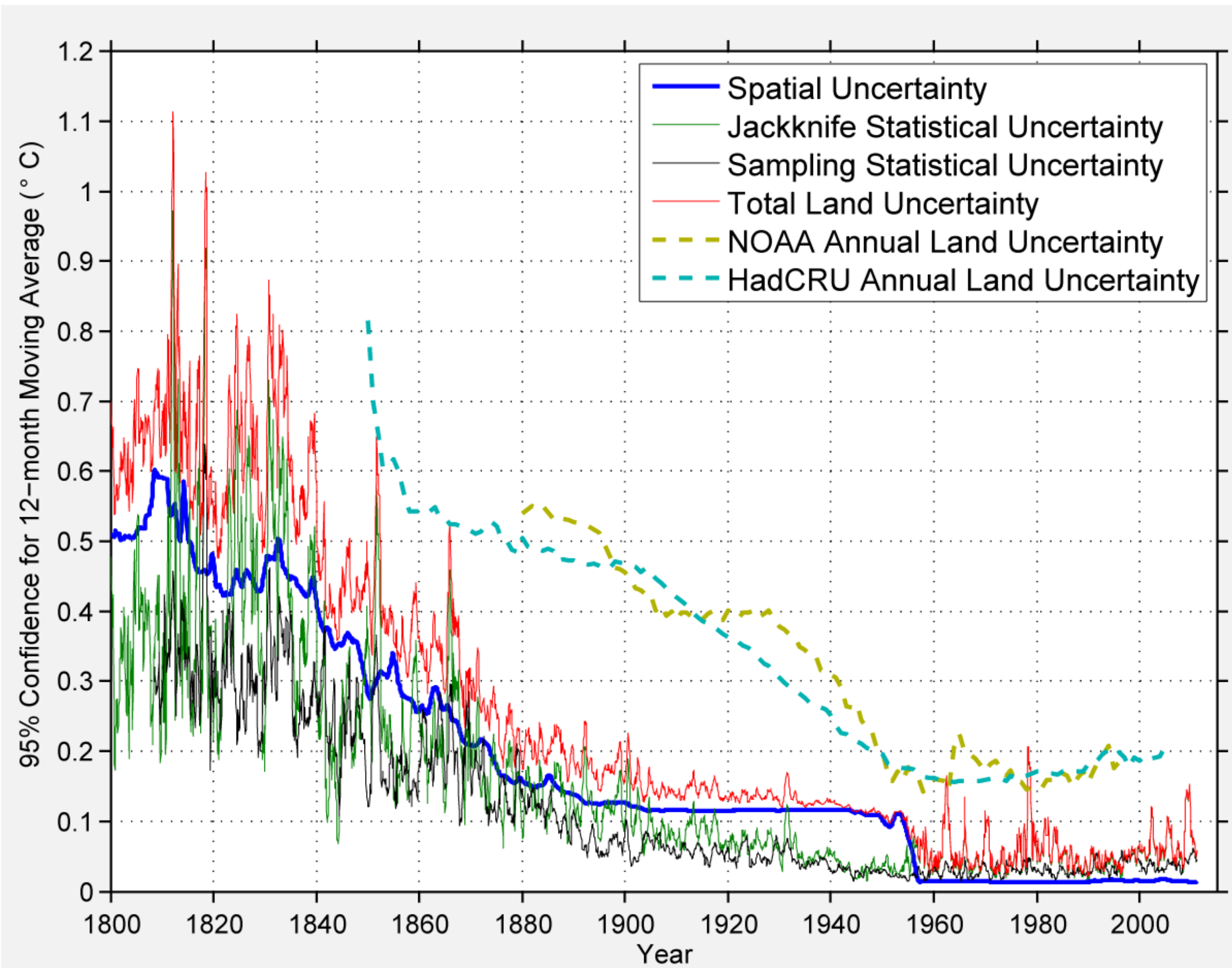
Though we are similar to NOAA for most of the 20^{th} century, we note that we have somewhat lower average temperatures during the period 1880-1930. This gives us a slightly larger overall trend for the 20^{th} century than any of the three groups. Most of that difference comes from the more uncertain early period. In previous work, it has been argued that instrumentation changes may have led to an artificial warm bias in the early 1900s (Folland et al.

2001, Parker 1994). To the degree that our reconstruction from that era is systematically lower than prior work (Figure 8) it could be that our methods are more resistant to biases due to those instrumental changes.

As is shown in Figure 5, we extend our record all the way back to 1800, including 50 more years than HadCRU and 80 more years than NOAA and GISS. We feel this extension is justifiable though obviously, any such reconstruction will have large uncertainties. Our analysis technique suggests that temperatures during the 19th century were approximately constant (trend 0.20 ± 0.25 C/century) and on average 1.48 ± 0.13 C cooler than the interval 2000-2009. Circa 1820 there is a negative temperature excursion that happens to roughly coincide with both the 1815 eruption of Mount Tambora and the Dalton Minimum in solar activity. The Mount Tambora eruption was the largest eruption in the historical era and has been blamed for creating the "year without a summer" (Oppenheimer 2003; Stothers 1984). It was preceded by an additional large eruption in 1809 (Wagner and Zorita 2005). The Dalton Minimum in solar activity from circa 1790 to 1830 includes the lowest 25 year period of solar activity during the last 280 years, but this is considered to have produced only minor cooling during this period, while volcanism was the dominant source of cooling (Wagner and Zorita 2005). Though the uncertainties are very large, the fact that this temperature excursion is well-established in the historical record and motivated by known climate forcings gives us confidence than the ~1820 excursion is a reflection of a true climate event. However, we will note that our early data is heavily biased towards North America and Europe, so we cannot draw conclusions about the regional versus global extent of the event.

As discussed above, the uncertainty in our result is conceptually divided into two parts, the "statistical uncertainty" which measures how well the temperature field was constrained by data in regions and times where data is available, $F(\vec{x}, t_j) \approx 1$, and the "spatial uncertainty" which measures how much uncertainty has been introduced into the temperature average due to the fact that some regions are not effectively sampled, $F(\vec{x}, t_j) \approx 0$. These uncertainties for the GHCN analysis are presented in Figure 9.

Figure 9. The 95% uncertainty on the Berkeley Average and the component spatial and jackknife statistical uncertainties for 12-month moving land averages



The two types of uncertainty tend to covary. This reflects the reality that station networks historically developed in a way that increasing station density (which helps statistical uncertainties) tended to happen at similar times to increasing spatial coverage (which helps spatial uncertainties). Overall, we estimate that the total uncertainty in the 12-month land-surface average from these factors has declined from about 0.7 C in 1800 to about 0.06 C in the present day.

The step change in spatial uncertainty in the early 1950s is driven by the introduction of the first weather stations to Antarctica during this time. Though the introduction of weather stations to Antarctica eliminated the largest source of spatial uncertainty, it coincidentally increased the statistical uncertainty during the post-1950 period. The Antarctic continent represents slightly less than 10% of the Earth's land area and yet at times has been monitored by only about dozen weather stations. To the extent that these records disagree with each other they serve as a large source of statistical noise. An example of this occurred in 1979 (see Figure 9) when an uncertainty of a couple degrees regarding the mean temperature of Antarctica led to an uncertainty of ~0.2 C for the whole land-surface.

Since the 1950s, the GHCN has maintained a diverse and extensive spatial coverage, and as a result the inferred spatial uncertainty is low. However, we do note that GHCN station counts have decreased precipitously from a high of 5883 in 1969 to about 2500 at the present day. This decrease has primarily affected the density of overlapping stations while maintaining broad spatial coverage. As a result, the statistical uncertainty has increased somewhat. We note again that the decrease in station counts is essentially an artifact of the way the GHCN monthly data set has been constructed. In fact, the true density of weather monitoring stations has remained nearly constant since the 1960s, and that should allow the "excess" statistical uncertainties shown here to be eliminated once a larger number of stations are considered in a future paper.

A comparison of our uncertainties to those reported by HadCRU and NOAA (Figure 9) is warranted (comparable figures for GISS are not available). Over much of the record, we find that our uncertainty calculation yields a value 50-75% lower than these other groups. As the sampling curves demonstrate (Figure 6), the reproducibility of our temperature time series on independent data is extremely high which allows us to feel justified in concluding that the statistical uncertainty is very low. This should be sufficient to estimate the uncertainty associated with any unbiased sources of random noise affecting the data. Similarly, the concordance of the analytical and empirical spatial uncertainties gives us confidence in those estimates as well.

In comparing the results we must note that curves by prior groups in Figure 9 include an extra factor they refer to as "bias error" by which they add extra uncertainty associated with urban heat islands and systematic changes in instrumentation (Brohan et al. 2006; Smith and Reynolds 2005). As we do not include comparable factors, this could explain some of the difference. However, the "bias" corrections being used cannot explain the bulk of the difference. HadCRU reports that the inclusion of "bias error" in their land average provides a negligible portion of the total error during the period 1950-2010. This increases to about 50% of the total error circa 1900, and then declines again to about 25% of the total error around 1850 (Brohan et al. 2006). These amounts, though substantial, are still substantially less than the difference between our uncertainty estimates and the prior estimates. We therefore conclude that our techniques can estimate the global land-based temperature with considerably less spatial and statistical uncertainty than prior efforts.

The assessment of bias / structural uncertainties may ultimately increase our total uncertainty, though such effects will not be quantified here. As mentioned previously, in one of our other submitted papers (Wickham et al.) we conclude that the residual effect of urbanization on our temperature reconstruction is probably close to zero nearly everywhere. In addition, the scalpel technique, baseline adjustments, and reliability measures should be effective at reducing the impact of a variety of biases. As such, we believe that any residual bias in our analysis will also be less than previous estimates. However, further analysis of our approach is needed before we can decide how effective our techniques are at eliminating the full range of biases.

We should also comment on the relatively large uncertainties in Figure 9 compared to those in Figure 1. These imply that the other groups believe past ocean temperatures have been much more accurately constrained than land-based temperatures. This conclusion is stated more explicitly at Smith and Reynolds 2005, Brohan et al. 2006.

In considering the very earliest portions of our reconstruction, we should note that our uncertainty analysis may be appreciably understating the actual uncertainty. This can occur for two principle reasons. Firstly, the uncertainty attributed to spatial undersampling is based primarily on the variability and spatial structure of climate observed during the latter half of the twenty century. For example, our approach assumes that the difference between temperatures in the Southern Hemisphere and temperatures in Europe remain similar in magnitude and range of variation in the past as they are today. The plausibility of this assumption is encouraged by the relative uniformity of climate change during the 20th century, as shown in Figure 7. However, this assumption could turn out to be overly optimistic and result in an under (or over) estimation of the natural climate variation in other parts of the world. Secondly, as the number of stations gets low the potential for additional systematic biases increases. The statistical error measurement technique essentially tests the internal consistency of the data. The more the data disagrees amongst itself, the larger the estimated statistical error. This is adequate if older measurement technology is simply more prone to large random errors. However, this technique cannot generally capture biases that occur if a large fraction of the records erroneously move in the same direction at the same time. As the number of available records becomes small, the odds of this occurring will increase. This is made more likely every time there is a systematic shift in the measurement technology being employed.

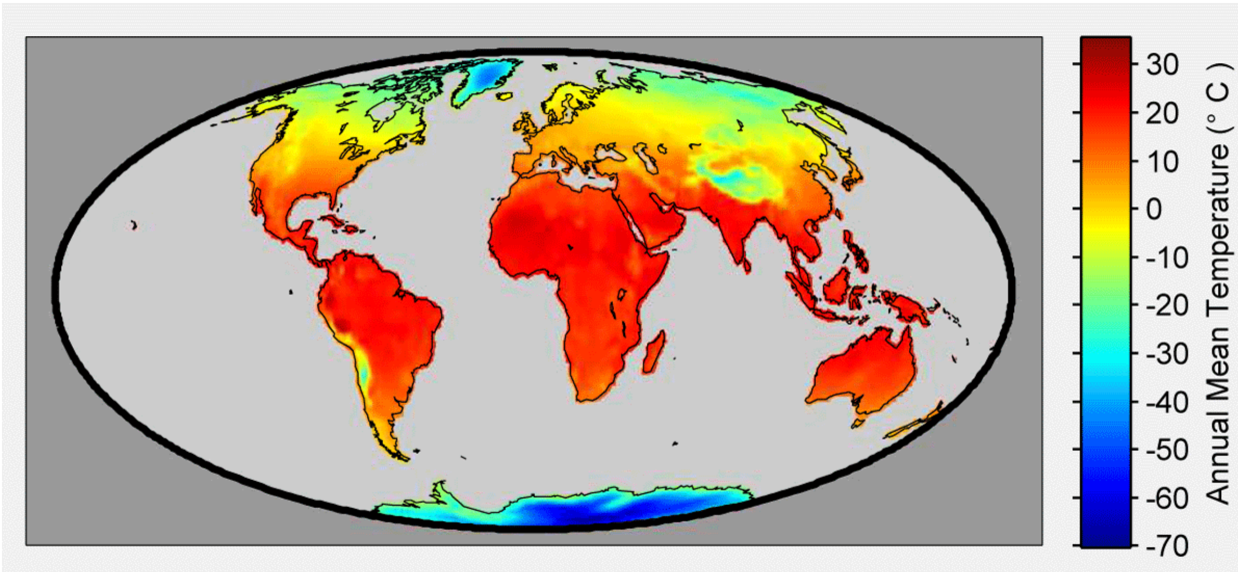
13. Climatology

Earlier in this paper, we defined the local temperature at position and time $\vec{x}_i, t_j$ to be given by

$$T(\vec{x}_i, t_j) = \theta(t_j) + C(\vec{x}_i) + W(\vec{x}_i, t_j)$$

where $\theta(t_j)$ is the global average temperature plotted in Figure 5, $W(\vec{x}_i, t_j)$ is the "weather field" that we estimated using equation 12. The remaining term $C(\vec{x}_i)$ is the approximately time-invariant long-term mean temperature of a given location, often referred to as the *climatology*. In our construction we treat this via equation [3] a function of latitude, altitude, and a smoothed local average calculated using equation [24]. As mentioned earlier, the latitude and altitude components account for about 95% of the structure. A map of the climatology $C(\vec{x}_i)$ is shown in Figure 10. We found the global land average from 1900 to 2000 to be about 8.90 ± 0.48 C, which is broadly consistent with the estimate of 8.5 C provided by Peterson et al. (2011). The Berkeley Average analysis process is somewhat unique in that it produces a global climatology and estimate of the global mean temperature as part of its natural operations, rather than discarding this information as the three other groups generally do.

Figure 10. A map of the derived Climatology term



14. Discussion

In this paper we described a new approach to global temperature reconstruction. We used spatially and temporally diverse data exhibiting varying levels of quality and constructed a global index series that yields an estimate of the mean surface temperature of the Earth. We employ an iteratively reweighted method that simultaneously determines the history of global mean land-surface temperatures and the baseline condition for each station, as well as making adjustments based on internal estimates of the reliability of each record. The approach uses variants of a large number of well-established statistical techniques, including a generalized fitting procedure, Kriging, and the jackknife method of error analysis. Rather than simply excluding all short records, as was done by prior Earth temperature analysis groups, we designed a system that allows short records to be used with appropriate – but non-zero – weighting whenever it is practical to do so. This method also allows us to exploit discontinuous and inhomogeneous station records without prior "adjustment", by breaking them into shorter segments at the points of discontinuity.

It is an important feature of this method that the entire discussion of spatial interpolation has been conducted with no reference to gridded data sets at all. The fact that our approach can, in principle, avoid gridding allows us to avoid a variety of noise and bias that can be introduced by gridding. That said, the integrals required by equation [2] will in general need to be computed numerically, and per equation [12] require the solution of a large number of matrix inverse problems. In the current paper, the numerical integrals were computed based on a 15,984 element equal-area array. Note that using an array for a numerical integration is qualitatively different from the gridding used by other groups. There are no sudden discontinuities, for example, depending on whether a station is on one side of a grid point or another, and no trade-offs to be made between grid resolution and statistical precision. We estimate that the blurring effects of the gridding methods used by HadCRU and GISS each introduce an unaccounted for uncertainty of approximately ~0.02 C in the computation of annual mean temperature. Such a gridding error is smaller than the total ~0.05 C uncertainties these groups report during the modern era, but not so small as to be negligible. The fact that the resolution of our calculation can be expanded without excess smoothing or trade offs for bias correction allows us to avoid this problem and reduce overall uncertainties. In addition, our approach could be extended in a natural way to accommodate variations in station density; for example, high data density regions (such as the United States) could be mapped at higher resolution without introducing artifacts into the overall solution.

We tested the method by applying it to the GHCN data based from 7280 stations used by the NOAA group. However, we used the GHCN raw data base without the "homogenization" procedures that were applied by NOAA which included adjustments for documented station moves, instrument changes, time of measurement bias, and urban heat island effects, for station moves. Rather, we simply cut the record at time series gaps and places that suggested shifts in the mean level. Nevertheless, the results that we obtained were very close to those obtained by NOAA using the same data and their full set of homogenization procedures. Our results did differ, particularly in recent years, from the analyses reported by the other two groups (NASA GISS and HadCRU). In the older periods (1860 to 1940), our statistical methods allow us to significantly reduce both the statistical and spatial uncertainties in the result, and they allow us to suggest meaningful results back to 1800. We note that we have somewhat lower average temperatures during the period 1880-1930 than found by the prior groups, and significantly lower temperatures in the period 1850 to 1880 than had been deduced by the HadCRU group. We also see evidence suggesting that temperature variability on the decadal time scale is lower now than it was the in the early 1800s. One large negative swing, around 1820, is coincident with both the eruption of Mt. Tambora and the Dalton Minimum in solar activity.

In another paper, we will report on the results of analyzing a much larger data set based on a merging of most of the world's openly available digitized data, consisting of data taken at over 39,000 stations, more than 5 times larger than the data set used by NOAA.

Acknowledgements

We are very grateful to David Brillinger for his guidance, key suggestions, and many discussions that helped lead to the averaging method presented in this paper. This work was done as part of the Berkeley Earth project, organized under the auspices of the Novim Group (www.Novim.org). We thank many organizations for their support, including the Lee and Juliet Folger Fund, the Lawrence Berkeley National Laboratory, the William K. Bowes Jr. Foundation, the Fund for Innovative Climate and Energy Research (created by Bill Gates), the Ann and Gordon Getty Foundation, the Charles G. Koch Charitable Foundation, and three private individuals (M.D., N.G. and M.D.). More information on the Berkeley Earth project can be found at www.BerkeleyEarth.org.

References

1. Arguez, Anthony, Russell S. Vose, 2011: The Definition of the Standard WMO Climate Normal: The Key to Deriving Alternative Climate Normals. *Bull. Amer. Meteor. Soc.*, **92**, 699–704.
2. 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, *J. Geophys. Res.*, *111*, D12106, doi:10.1029/2005JD006548.
3. Cressie, Noel. "The Origins of Kriging." *Mathematical Geology*, Vol. 22, No. 3, 1990.
4. David R. Easterling, Briony Horton, Philip D. Jones, Thomas C. Peterson, Thomas R. Karl, David E. Parker, M. James Salinger, Vyacheslav Razuvayev, Neil Plummer, Paul Jamason and Christopher K. Folland. "Maximum and Minimum Temperature Trends for the Globe" *Science*. Vol. 277 no. 5324 pp. 364-367.
5. Davis, R. A., T. C. M. Lee, and G. A. Rodriguez-Yam, 2006: "Structural break estimation for nonstationary time series models." J. Amer. Stat. Assoc., 101, 223–239.
6. Easterling, D. R. & Wehner, M. F. (2009) "Is the climate warming or cooling?" Geophys. Res. Lett. 36, L08706.
7. Folland, C. K., et al. (2001), Global temperature change and its uncertainties since 1861, *Geophys. Res. Lett.*, 28(13), 2621–2624, doi:10.1029/2001GL012877.
8. 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**, 957-966, doi:10.1126/science.213.4511.957
9. Hansen, J., R. Ruedy, J. Glascoe, and Mki. Sato, 1999: GISS analysis of surface temperature change. *J. Geophys. Res.*, **104**, 30997-31022, doi:10.1029/1999JD900835.
10. Hansen, J., R. Ruedy, Mki. Sato, and K. Lo, 2010: Global surface temperature change. *Rev. Geophys.*, **48**, RG4004, doi:10.1029/2010RG000345.

11. Hansen, J.E., and S. Lebedeff, 1987: Global trends of measured surface air temperature. *J. Geophys. Res.*, **92**, 13345-13372, doi:10.1029/JD092iD11p13345.
12. Hinkley, D. V. (1971), “Inference about the change-point from cumulative sum tests,” Biometrika, 58 3, 509-523.
13. Jones, P. D., P. Ya. Groisman, M. Coughlan, N. Plummer, W.-C. Wang and T. R. Karl, Assessment of urbanization effects in time series of surface air temperature over land, *Nature*, 347, 169-172, 1990.
14. Jones, P. D., and A. Moberg (2003), Hemispheric and Large-Scale Surface Air Temperature Variations: An Extensive Revision and an Update to 2001, J. Clim., 16, 206–23.
15. Jones, P.D., T.M.L. Wigley, and P.B. Wright. 1986. Global temperature variations between 1861 and 1984. *Nature*322:430-434.
16. Journel, A. G. *Fundamentals of geostatistics in five lessons*. American Geophysical Union, 1989; 40 pages.
17. Klein Tank, A. M. G., G. P. Können, 2003: Trends in Indices of Daily Temperature and Precipitation Extremes in Europe, 1946–99. J. Climate, 16, 3665–3680.
18. Krige, D.G, *A statistical approach to some mine valuations and allied problems at the Witwatersrand*, Master's thesis of the University of Witwatersrand, 1951.
19. L. V. Alexander, X. Zhang, T. C. Peterson, J. Caesar, B. Gleason, A. M. G. Klein Tank, M. Haylock, D. Collins, B. Trewin, F. Rahimzadeh, A. Tagipour, K. Rupa Kumar, J. Revadekar, G. Griffiths, L. Vincent, D. B. Stephenson, J. Burn, E. Aguilar, M. Brunet, M. Taylor, M. New, P. Zhai, M. Rusticucci, and J. L. Vazquez-Aguirre (2006) “Global observed changes in daily climate extremes of temperature and precipitation,” *Journal of Geophysical Research*, v. 111, D05109.
20. Meehl, Gerald A.; Arblaster, Julie M.; Fasullo, John T.; Hu, Aixue; Trenberth, Kevin E., (2011) Model-based evidence of deep-ocean heat uptake during surface-temperature hiatus periods. *Nature Climate Change*. 2011/09/18/online

21. Menne M.J., C.N. Williams Jr., and R.S. Vose (2009), The United States Historical Climatology Network Monthly Temperature Data – Version 2. *Bull. Amer. Meteor. Soc.*, *90,* 993-1007

22. Menne, M.J., and C.N. Williams, Jr. (2009), Homogenization of temperature series via pairwise comparisons. *J. Climate*, **22,** 1700–1717.

23. Miller, Rupert (1974), “The Jackknife – A review,” *Biometrika*, v. 61, no. 1, pp. 1-15.

24. Muller, Richard A, Judith Curry, Donald Groom, Robert Jacobsen, Saul Perlmutter, Robert Rohde, Arthur Rosenfeld, Charlotte Wickham, Jonathan Wurtele (submitted) “Earth Atmospheric Land Surface Temperature and 1 Station Quality” JGR.

25. Oke, T.R. (1982), The energetic basis of the urban heat island. *Quarterly Journal of the Royal Meteorological Society*, V. 108, no. 455, p. 1-24.

26. Oppenheimer, Clive (2003). "Climatic, environmental and human consequences of the largest known historic eruption: Tambora volcano (Indonesia) 1815". *Progress in Physical Geography* **27** (2): 230–259.

27. Page, E. S. (1955), “A test for a change in a parameter occurring at an unknown point,” *Biometrika*, 42, 523-527.

28. Parker, D. E., (1994) “Effects of changing exposure of thermometers at land stations,” *International Journal of Climatology*, v. 14, no. 1, pp 1-31.

29. Peterson, T.C., and R.S. Vose, 1997: An overview of the Global Historical Climatology Network temperature database. Bulletin of the American Meteorological Society, 78 (12), 2837-2849.

30. Peterson, Thomas C., Katharine M. Willett, and Peter W. Thorne (2011) “Observed changes in surface atmospheric energy over land,” GEOPHYSICAL RESEARCH LETTERS, VOL. 38, L16707, 6 PP.

31. Quenoille, M. H. (1949), “Approximate tests of correlation in time-series,” *Journal of the Royal Statistical Society* B 11, p. 68-84.

32. Smith and Reynolds, 2005: A global merged land air and sea surface temperature reconstruction based on historical observations (1880–1997). *J. Climate,* **18,** 2021–2036.

33. Smith, T. M., et al. (2008), Improvements to NOAA's Historical Merged Land-Ocean Surface Temperature Analysis (1880-2006), J. Climate, 21, 2283-2293.

34. Stothers, Richard B. (1984). "The Great Tambora Eruption in 1815 and Its Aftermath". *Science* 224 (4654): 1191–1198.

35. Trenberth, K.E., P.D. Jones, P. Ambenje, R. Bojariu, D. Easterling, A. Klein Tank, D. Parker, F. Rahimzadeh, J.A. Renwick, M. Rusticucci, B. Soden and P. Zhai, 2007: Observations: Surface and Atmospheric Climate Change. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

36. Tsay, R, S.. (1991) Detecting and Modeling Non-linearity in Univariate Time Series Analysis. Statistica Sinica 1:2,431-451.

37. Tukey, J.W. (1958), "Bias and confidence in not quite large samples", *The Annals of Mathematical Statistics*, 29, 614.

38. Vose, C. N. Williams Jr., T. C. Peterson, T. R. Karl, and D. R. Easterling (2003), An evaluation of the time of observation bias adjustment in the U.S. Historical Climatology Network. *Geophys. Res. Lett., 30,* 2046, doi:10.1029/2003GL018111

39. Wagner, Sebastian and Eduardo Zorita (2005) "The influence of volcanic, solar and CO2 forcing on the temperatures in the Dalton Minimum (1790–1830): a model study," Climate Dynamics v. 25, pp. 205–218.

40. Wickham, Charlotte, Judith Curry, Don Groom, Robert Jacobsen, Richard Muller, Saul Perlmutter, Robert Rohde, Arthur Rosenfeld, Jonathan Wurtele (submitted) "Influence of Urban Heating on the Global Temperature Land Average Using Rural Sites Identified from MODIS Classifications", JGR.

41. Zhang, Xuebin, Francis W. Zwiers, Gabriele C. Hegerl, F. Hugo Lambert, Nathan P. Gillett, Susan Solomon, Peter A. Stott & Toru Nozawa, (2007) "Detection of human influence on twentieth-century precipitation trends" *Nature* 448, 461-465.

Figure Captions

Figure 1. (Upper panel) Comparison of the global annual averages of the three major research groups, plotted relative to the 1951-1980 average. (Lower panel) The annual average uncertainty at 95% confidence reported by each of the three groups. NASA reports an uncertainty at only three discrete times, shown as solid dots, while the other two groups provide continuous estimates of the uncertainty.

Figure 2. Mean correlation versus distance curve constructed from 500,000 pair-wise comparisons of station temperature records. Each station pair was selected at random, and the measured correlation was calculated after removing seasonality and with the requirement that they have at least 10 years of overlapping data. Red, green, and yellow curves show a moving range corresponding to the inner 80, 50, and 20% of data respectively. The black curve corresponds to the modeled correlation vs. distance reported in the text. This correlation versus distance model is used as the foundation of the Kriging process used in the Berkeley Average.

Figure 3. Correlation versus distance fits, similar to Figure 2, but using only stations selected from portions of the Earth. The Earth is divided into eight longitudinal slices (Left) or seven latitudinal slices (Right), with the slice centered at the latitude or longitude appearing in the legend. In each panel, the global average curve (Figure 2) is plotted in black. All eight longitudinal slices are found to be similar to the global average. For the latitudinal slices, we find that the correlation is systematically reduced at low latitudes. This feature is discussed in the text.

Figure 4. (Upper) Station locations for the 7280 temperature stations in the Global Historical Climatology Network Monthly dataset. (Lower Left) Number of active stations over time. (Lower Right) Percentage of the Earth's land area sampled by the available stations versus time, calculated as explained in the text. The transition during the mid 1950s corresponds to the appearance of the first temperature records on Antarctica.

Figure 5. Result of the Berkeley Average Methodology applied to the GHCN monthly data. Top plot shows a 12-month land-only moving average and associated 95% uncertainty from statistical and spatial factors. The lower plot shows a corresponding 10-year land-only moving average and 95% uncertainty. This plot corresponds to the parameter $\theta(t_j)$ in Equation 5. Our plotting convention is to place each value at the middle of the time interval it represents. For example, the 1991-2000 average in the decadal plot is shown at 1995.5.

Figure 6. Five independent temperature reconstructions each derived from a separate 20% of the GHCN stations. The upper figure shows the calculation of the temperature record based on five independent subsamples. The lower plot shows their difference from the 100% result, and the expected 95% uncertainty envelope. The uncertainty envelope used here is scaled by $\sqrt{5}$ times the statistical uncertainty reported for the complete Berkeley Average analysis. This reflects the larger variance expected for the 20% samples.

Figure 7. Maps showing the decadal average changes in the land temperature field. In the upper plot, the comparison is drawn between the average temperature in 1900 to 1910 and the average temperature in 2000 to 2010. In the lower plot, the same comparison is made but using the

interval 1960 to 1970 as the starting point. We observe warming over all continents with the greatest warming at high latitudes and the least warming in southern South America.

Figure 8. Comparison of the Berkeley Average to existing land-only averages reported by the three major temperature groups. The upper panel shows 12-month moving averages for the four reconstructions, and a gray band corresponding to the 95% uncertainty range on the Berkeley average. The lower panel shows each of the prior averages minus the Berkeley average, as well as the Berkeley average uncertainty. As noted in the text, there is a much larger disagreement among the existing groups when considering land-only data than when comparing the global averages (**Figure 1**). HadCRU and GISS have systematically lower trends than Berkeley and NOAA. In part, this is likely to reflect differences in how "land-only" has been defined by the three groups. Berkeley is very similar to the NOAA result during the twentieth century and slightly lower than all three groups during the 19th century.

Figure 9. The 95% uncertainty on the Berkeley Average (red line) and the component spatial (blue) and jackknife statistical (green) uncertainties for 12-month moving land averages. For comparison the sampling statistical uncertainty is also shown (black), though it does not contribute to the total. From 1900 to 1950, the spatial uncertainty is dominated by the complete lack of any stations on the Antarctic continent. From 1960 to present, the statistical uncertainty is largely dominated by fluctuations in the small number of Antarctic temperature stations. For comparison, the land-only 95% uncertainties for HadCRU and NOAA are presented. As discussed in the text, in addition to spatial and statistical consideratios, the HadCRU and NOAA curves include additional estimates of "bias error" associated with urbanization and station

instrumentation changes that we do not currently consider. The added "bias error" contributions are small to negligible during the post 1950 era, but this added uncertainty is a large component of the previously reported uncertainties circa 1900.

Figure 10. A map of the derived Climatology term, $C(\vec{\mathbf{x}}_i)$. 95% of the variation is accounted for by altitude and latitude. Departure from this is evident in Europe and in parts of Antarctica.

APPENDIX

Symbols used in the Berkeley Average method.

t	the time
t_j	the j-th time step (i.e. month)
$\vec{x}$	an arbitrary position on the surface of the earth
$\vec{x}_i$	the position of the i-th station on the surface of the earth
$T(\vec{x}, t)$	the true temperature at location $\vec{x}$ and time t
$\hat{T}(\vec{x}, t)$	the estimated temperature at location $\vec{x}$ and time t
$d_i(t_j)$	the measured temperature time series (e.g. "data") at the i-th station and j-th time step
$\theta(t)$	the global mean temperature time series
$C(\vec{x})$	the long-term average temperature as a function of location ("climatology")
$W(\vec{x}, t)$	spatial and temporal variations in $T(\vec{x}, t)$ not ascribed to $\theta(t)$ or $C(\vec{x})$ (e.g. the "weather")
$\lambda(\vec{x})$	the temperature change as a function of latitude
$h(\vec{x})$	the temperature change as a function of surface elevation
$G(\vec{x})$	the variations in $C(\vec{x})$ not ascribed to $h(\vec{x})$ or $\lambda(\vec{x})$, i.e. the geographical anomalies in the mean temperature field.
$\hat{b}_i$	the baseline temperature of the i-th station
$S_i(\vec{x}, t_j)$	the initial spatial weight of the i-th station at location $\vec{x}$ and time t_j
$S_i^*(\vec{x}, t_j)$	the adjusted spatial weight of the i-th station at location $\vec{x}$ and time t_j
ω_i	the reliability weight associated with the i-th station
e	the mean local misfit between a temperature record and the interpolated

	field
$F(\vec{x}, t_j)$	a measure of the completeness of the sampling at location $\vec{x}$ and time t_j
$\bar{F}(t_j)$	a measure of the completeness of the sampling across all land at time t_j
$B_i(\vec{x})$	the baseline spatial weighting factor for the i-th station at location $\vec{x}$
$R(\vec{x}_a, \vec{x}_b)$	the expected spatial correlation in temperature between locations $\vec{x}_a$ and $\vec{x}_b$
$C(\vec{x}_a, \vec{x}_b)$	the covariance in temperature between locations $\vec{x}_a$ and $\vec{x}_b$
σ_i^2	the variance of the temperature record at the i-th station
$O_{i,j}$	the outlier weight associated with data point $T_i(t_j)$
$\Delta_i(t_j)$	the difference between data point $d_i(t_j)$ and the estimated value of the temperature field at the same location and time.

Table 1: Summary of the primary symbols used to describe the Berkeley Earth averaging method.