Berkeley Earth Temperature Averaging Process

Robert Rohde, Richard Muller (chair), Robert Jacobsen, Saul Perlmutter, Arthur Rosenfeld, Jonathan Wurtele, Don Groom, Judith Curry, Charlotte Wickham

Abstract

A new mathematical framework is presented for producing maps and large-scale averages of temperature changes from weather station thermometer data for the purposes of climate analysis. The method allows the inclusion of short and discontinuous temperature records, so that nearly all digitally archived thermometer data can be used. The framework uses the statistical method known as Kriging to interpolate data from stations to arbitrary locations on the Earth. An iterative weighting process is used to reduce the influence of statistical outliers. Statistical uncertainties are calculated by subdividing the data and comparing the results from statistically independent subsamples using the Jackknife method. Spatial uncertainties from periods with sparse geographical sampling are estimated by calculating the error made when we analyze post-1960 data using similarly sparse spatial sampling. Rather than "homogenize" the raw data, an automated procedure identifies discontinuities in the data; the data is then broken into two parts at those times, and the parts treated as separate records. We apply this new framework to the Global Historical Climatology Network (GHCN) monthly land temperature dataset, and obtain a new global land temperature reconstruction from 1800 to the present. In so doing, we find results in close agreement with prior estimates made by the groups at NOAA, NASA, and at the Hadley Center / Climate Research Unit in the UK. We find that the global land mean temperature increased by 0.89 ± 0.06 C in the difference of the Jan 2000-Dec 2009 average from the Jan 1950-Dec 1959 average (95% confidence for statistical and spatial uncertainties).

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 (including land and oceans) has increased 0.64 ± 0.13 C from 1956 to 2005 at 95% confidence (Trenberth et al. 2007). In a review of temperature changes over land areas, the IPCC summarized four reconstructions of the global land average temperature as having trends ranging from 0.188 ± 0.069 °C / decade to 0.315 ± 0.088 °C / decade over the time interval 1979 to 2005 (Trenberth et al. 2007). However, some of this range reflects methodological differences in how "land average" was defined and over what regions it was computed.

The three major groups that produce ongoing temperature reconstructions are the NASA Goddard Institute of Space Science (NASA GISS), the National Climate Data Center at the National Oceanic and Atmospheric Administration (NOAA NCDC), and the joint project of the UK Meteorological Office Climatic Research Unit and the Hadley Centre at the University of West Anglia (Hadley / CRU). Their annual land-surface temperature histories are presented in Figure 1A, as well as the available uncertainties in Figure 1B. NASA GISS does not publish an uncertainty specific to their land-surface data product. In Figure 1A we show that these groups report a range of best values from 0.81 to 0.93 C when estimating the increase in land temperatures for the 2000s decade relative to the 1950s decade, with reported 95% uncertainties of roughly 0.15 to 0.2 C.

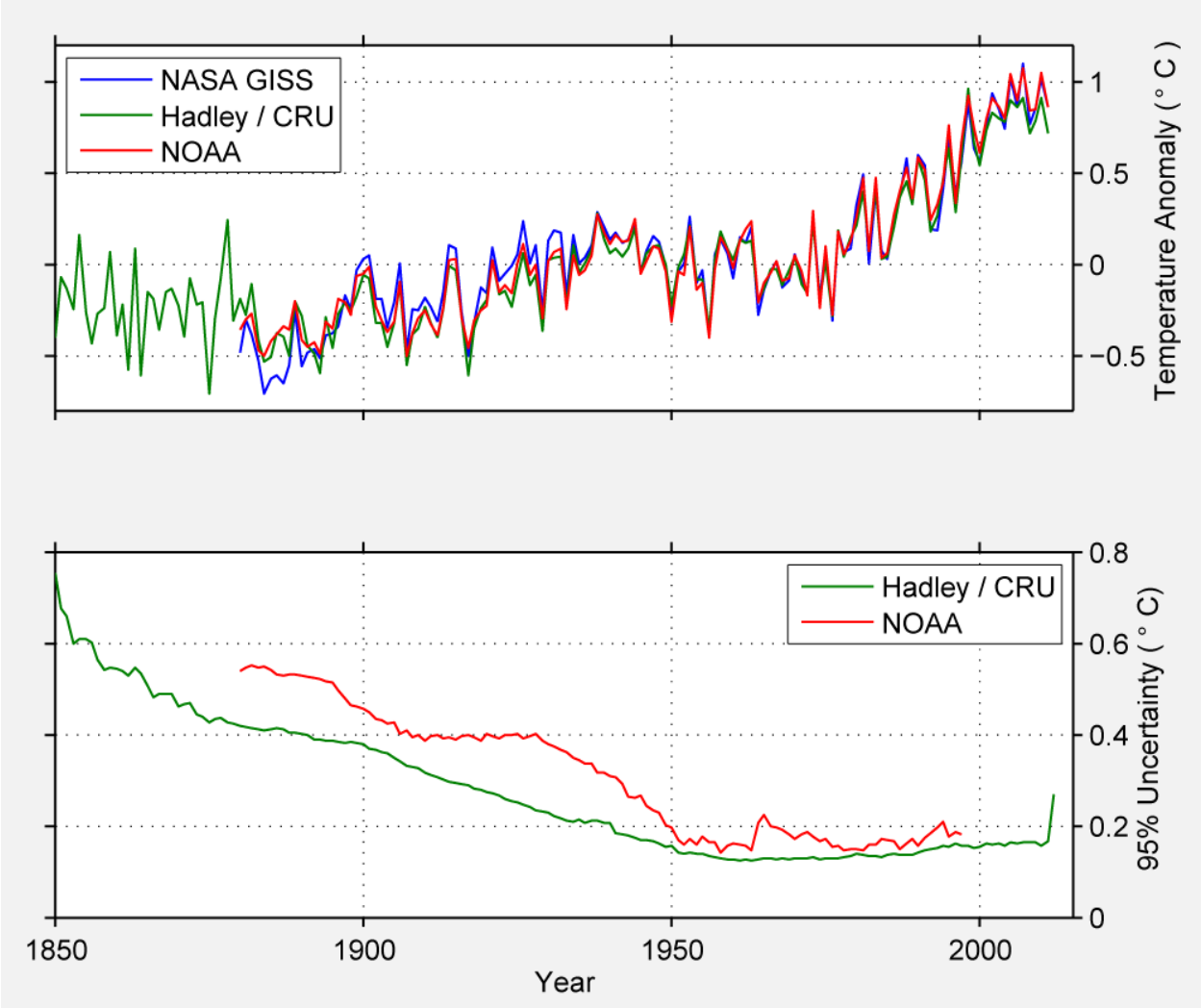

Figure 1: (Top Panel) Comparison of annual land-surface average temperature anomalies for the three major research groups (results updated online, methods described by Brohan et al. 2006 [CRUTEM3]; Smith et al. 2008; Hansen et al. 2010). For this purpose, the Hadley / CRU simple average has been used rather than the more widely cited latitudinal-band weighted average, as the simple average is more similar in methodology and results to the other averages presented here. (Bottom Panel) The 95% percent uncertainty estimate on the annual values provided by Hadley / CRU and NOAA. NASA GISS does not appear to have ever published an uncertainty specific to their land-surface computation, and the most recent available NOAA uncertainty for land-only data (from Smith and Reynolds 2005) terminates in the late 1990s.

During the second half of the twentieth century weather monitoring instruments of good quality were widely deployed, yet the quoted uncertainty on temperature change during this time period is still around 20%. Of the two domains, the uncertainties reported on land averages are often 20-100% larger than ocean uncertainties (Smith and Reynolds 2005, Brohan et al. 2006), though this is somewhat mitigated by the fact that land occupies only 29% of the Earth's surface. The present work will present a

method of significantly reducing the spatial and statistical uncertainty associated with the land-surface temperature calculations

The Berkeley Earth Surface Temperature project was created to develop a new framework for estimating the average Earth surface temperature, and to apply it to perform an independent estimate of the rate of recent global temperature change. Our goals included: A) increasing the size of the data set used to study global climate change, B) bringing different statistical techniques to bear on the problem with a goal of minimizing uncertainties in the resulting averages, and C) reanalyzing 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; rather we describe the framework that we have developed, apply it to a well-known large data set — that of the Global Historical Climate Network (GHCN), compiled and merged by NOAA NCDC. We calculate a new estimate of the Earth's land surface temperature, and then compare our results to those of the prior groups that have done such analysis. We chose to analyze the GHCN dataset largely as a test of our analysis framework that allows us to make a direct comparison with a prior analysis without introducing issues of data selection effects. The Berkeley Earth Project is also developing and merging a much larger data set of temperature records, and analysis of those data will be published separately.

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, one can never capture the ideal global average temperature completely; however, the available data can be used to tightly constrain its value. The land surface temperature average is calculated by including only

land points in the average. It is important to note that these averages count every square kilometer of land equally; the average is not a station average but a land-area weighted average.

In this paper we will give a description of the approach that we use to estimate the global land surface temperature average. In order to be clear and concise, we will be somewhat qualitative in our description; the precise terms and the detailed equations will be provided in the mathematical supplement that accompanies this paper. That supplement also adopts a more standard statistical language, for example, clearly separating estimated values from the true values using the standard "hat" notation of statistics. The details in the supplement will be helpful to those who wish to adopt our methods or reproduce or modify our model. In addition, the computer program that actually implements the model and the parameter calculations has been placed online at BerkeleyEarth.org. It is written in the computer language Matlab, one that we believe is relatively transparent to read even for those who are not familiar with its detailed syntax.

A conceptually simple approach for estimating the land surface temperature average T_{avg} would be to begin by estimating the temperature at every land point on the Earth by using an interpolation based on the existing thermometer data and then averaging the interpolated field to generate the large-scale average. There are many ways to interpolate, but an attractive method is the linear least-squares estimation algorithm known as Gaussian process regression or *Kriging*. [Krige 1951, Cressie 1990, Journel 1989] Kriging uses the covariance between stations to assure that close-packed temperature stations (which can be highly correlated with each other) do not unduly influence the interpolation. Kriging is widely used in both academia and industry, in fields as diverse as geophysical exploration to Real Estate appraisal. If its underlying assumptions are satisfied (e.g. the true field contains normally distributed random variations about a common mean) then Kriging is provably the best linear unbiased estimator of an underlying spatial field.

Let $T(x,t)$ be an estimate for the temperature field at an arbitrary location *x* on the surface of the Earth, based on an interpolation from the existing thermometer data for time *t.* In the simplest approach the average surface land temperature estimate $T_{avg}(t)$ is calculated by integrating over the land area:

$$T_{avg}(t) \equiv \left(\frac{1}{A}\right) \int T(x,t)\, dA \qquad (1)$$

where *A* is the total land area. Note that the average is not an average of stations, but an average over the surface using an interpolation to determine the temperature at every land point. In this approach, there are two other quantities of interest that we call the *local climate* and the *local weather*. We define the local climate $C(x)$ by

$$C(x) \equiv \langle T(x,t) - T_{avg}(t) \rangle_t \qquad (2)$$

where the subscript *t* to the brackets indicates that the average is done for a given location over time. Note that the climate consists of all the variation that depends on location (latitude and longitude) but not on time. The fundamental difference in mean temperature between the North Pole and the Equator will be contained in $C(x)$.

The local weather $W(x,t)$ is defined as the remainder, that is, the difference between the actual temperature record at a given location and what you would estimate from $T_{avg}(t)$ and $C(x)$ alone:

$$W(x,t) \equiv T(x,t) - T_{avg}(t) - C(x) \qquad (3)$$

In this formulation, the weather is the *residual*, the temperature behavior that one can't account for using a simple model that combining only global temperature changes with stable local climate. However, if properly described, the weather includes variability from a variety of true regional scale effects such as moving weather fronts and the influence of El Nino.

This will prove to be a useful decomposition. Put in a more straightforward way, we say that the land temperature is the sum of the average temperature change plus the local (but temporally stable) climate plus the local weather:

$$T(x,t) = T_{avg}(t) + C(x) + W(x,t) \qquad (4)$$

The individual terms by virtue of their definitions satisfy the following important constraints:

$$\langle C(x)\rangle_x = 0 \qquad (5)$$

$$\langle W(x,t)\rangle_x = 0 \qquad (6)$$

$$\langle W(x,t)\rangle_t = 0 \qquad (7)$$

One approach to construct the interpolated field would be to use Kriging directly on the station data to define $T(x,t)$. Although outwardly attractive, this simple approach has several problems. The assumption that all the points contributing to the Kriging interpolation have the same mean is not satisfied with the raw data. To address this, we introduce a baseline temperature b_i for every temperature station *i*; this baseline temperature is calculated in our optimization routine and then subtracted from each station prior to Kriging. This converts the temperature observations to a set of anomaly observations with an expected mean of zero. This baseline parameter is essential our representation for $C(x_i)$. But because the baseline temperatures are calculated solutions to the procedure, and yet are needed to estimate the Kriging coefficients, the approach must be iterative.

Another problem is unreliability of stations; some stations show large differences from nearby stations that are not plausibly related to weather or climate; they could be measurement error or local systematic effects (such poor station siting, or excess heating in an urban environment). To reduce the effects of such stations, we apply an iterative weighting procedure. Weights are applied to the station contributions that affect the Kriging averages, i.e. the contributions made by individual stations towards the estimate of the temperature at a given location. To calculate the weights, we first treat all stations as equal and calculate an initial T_{avg}. Then an automated routine identifies outliers, and a weight is applied to each station that determines how much it contributes to the Kriging average at (x,t), and the Kriging calculation is repeated. The process is iterated until the weights applied to each station converge. No station is omitted, but a poor station could receive a weight as low as 1/26 that of a trusted station (more on this later). Note again that although the weights affect the interpolated temperature estimate for a given location, all square kilometers of land temperature contribute equally to T_{avg}.

In addition to persistent disagreements with nearby stations (as discussed in the previous paragraph), we incorporate a procedure that detects large discontinuities in time in a single station record. These could be caused by undocumented station moves, changes in instrumentation, or just the construction of a building nearby. These discontinuities are identified prior to the determination of the temperature parameters by an automated procedure. Once located, they are treated by separating the data from that record into two sections at the discontinuity time, creating effectively two stations out of one; we call this process the *scalpel*, and we'll discuss it in more detail later in this paper. Other groups typically adjust the two segments to remove the discontinuity; they call this process homogenization. We apply no explicit homogenization; other than splitting, the data is left untouched. Any adjustment needed between the stations will be done automatically as part of the computation of the optimum temperature baseline parameters b_i.

We break the local climate function $C(x)$ into a three subfunctions:

$$C(x) = \lambda(\text{ latitude}) + h(\text{ elevation}) + G(x) \qquad [8]$$

The functions λ and h are adjusted to describe the average behavior of temperature with latitude and elevation. $G(x)$ is the "geographic anomaly", i.e. the spatial variations in mean climatology that can't be explained solely by latitude and elevation. The $G(x)$ will include many large-scale climate patterns, such as the effects on the land of the Gulf Stream and Jet Stream. With appropriate functions chosen for λ and h it is possible to account for 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 are simple functions (polynomials and splines), and are given in the supplement. They include free parameters that are adjusted to give the best fit.

The mathematical model that we use to estimate T_{avg} is the following. We wish to obtain an estimate T_{avg}^{j} of the average Earth's land surface temperature at time t_j. For the period 1800 to 2010, there are 2532 monthly values of T_{avg}^{j}; these are all adjustable parameters that we will fit to the data. In addition, for every temperature station i we have a parameter b_i, the baseline temperature for that station;

these too are adjusted to obtain the best fit. These parameters allows for each station to depart from its expected local mean climatology (the function of latitude and elevation described above). For the GHCN network discussed in this paper, the number of initial stations was 7,280; after scalpeling, they had been broken into 44,840 effective stations that we treated in our analysis as completely independent. Add those to the 2532 values to be determined for T_{avg}^{j} is, and about 18 for the $C(x)$ term, and we have 47,388 parameters. These do not include the weights for each station used to reduce the effects of outliers. Fortunately we are able to apply a mathematical shortcut that allowed us to handle this large number of parameters.

Our conceptual approach is this: the weather term $W(x,t)$ in the temperature expansion Eq. (4) can be interpreted as the residual, the part of the record that is not fit by a simple model of global land temperature change $T_{avg}(t)$ and locally stable climate $C(x)$. $W(x,t)$ represents the mismatch, a measure of the failure to account for the data with the parameters alone. To get the best estimates of the parameters, the values that give the fullest explanation of the data, we can adjust them to minimize the square of the residuals; in other words, we minimize

$$\int W^2(x,t)\, dA \qquad (9)$$

We emphasize again that the integral is over the land area, and not just over the temperature stations. We obtain the weather for the required global land coverage through Kriging interpolation.

To solve this equation we can employ a trick. Mathematically, it can be shown that minimizing the above term is equivalent to solving the following equation:

$$\int W(x,t)\, F(x,t) dA = 0 \qquad (10)$$

where F is a linear combination of the Kriging coefficients; the precise definition of F is given in the mathematical supplement. Its physical interpretation is that it F represents the fraction of the weather field that has been effectively constrained by the data. The important features of F are the following: $0 \le F \le 1$ everywhere; F approaches 1 in the limit of dense sampling (many samples within a correlation length);

and when F is small, the weather term $W(x,t)$ is also small. Given these properties, in the limit of dense sampling we can approximate the equation as:

$$\int W(x,t)dA = 0 \qquad (11)$$

Note that this equation is also true for sparse sampling if the field $W(x,t)$ were to represent the actual weather field rather than our estimate of the field! That's because it is identical to Equation (6). Thus this equation appears to have a robustness that we can exploit. Rather than minimize the term (9), we choose to solve Equation (11). This has the important computational advantage that it isolates the Kriging coefficients so that the integrals can be performed independently for each station (see the supplement). This makes the solution much faster to calculate. It has the disadvantage that it might not give the precise optimal minimization of (9), but it does maintain the natural physical interpretation of the parameters. The added error should be very small in the modern era (post 1960) when the coverage F is nearly complete.

In addition to this procedure, once we have a solution (a set of parameters that satisfies Equation 10) we examine that solution to see if there are spurious contributions, temperature records for particular stations that are unreasonably poor fits to that solution. This is a way of spotting "outliers", records that might derive from poor instruments or mistakes in record keeping. According to the methods of robust data analysis developed by Tukey (1977, such data points should identified and then either discarded or deweighted. We choose to deweight, and we do that in the manner that will be described in our section on outlier weighting. Once the weights are applied to the outlier stations, such stations contribute less to the Kriging estimate; that estimate is redone, and then Equation (10) is solved once more. The process is iterated until the parameters converge.

In our approach, we derive not only the Earth's average temperature record T_{avg}^{j}, but also the best values for the station baseline temperatures b_i. Note, however, that there is an ambiguity; we could arbitrarily add a number to all the T_{avg}^{j} values as long as we subtracted the same number all the baseline temperatures b_i. To remove this ambiguity, in addition to minimizing the weather function $W(x,t)$, we

also minimize the integral of the square of the core climate term $G(x)$ that appears in Equation (8). To do this involves modifying the functions that describe latitude and elevation effects, and that means adjusting the 18 parameters that define λ and h, as described in the supplement. This process does not affect in any way our calculations of the temperature *anomaly*, that is, temperature differences compared to those in a base period (e.g. 1950 to 1980). It does, however, allow us to calculate from the fit the absolute temperature of the Earth land average at any given time, a value that is not determined by methods that work only with anomalies by initially normalizing all data to a baseline period. Note, however, that the uncertainty associated with the absolute temperature value is larger than the uncertainty associated with the changes, i.e. the anomalies. The increased error results from the large range of variations in b_i from roughly 30 C at the tropics to about -50 C in Antarctica, as well as the rapid spatial changes associated with variations in surface elevation. For temperature differences, the $C(x)$ term cancels (it doesn't depend on time) and that leads to much smaller uncertainties for anomaly estimates than for the absolute temperatures.

We make the following additional approximations. Rather than use covariance functions in the Kriging calculation, we use correlation function; this is a good approximation as long as the variance is changing slowly with time. We also assume that the correlation function depends solely on distance, and we estimate it by fitting an empirical function to the correlations observed in the data. The details are described in the mathematical supplement.

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 themselves 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). We estimate that on average 13% of the non-seasonal variance in a typical monthly temperature time series is caused by 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 generally 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, expressed via 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 significant reference or overlap 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 b_i value; however, unnecessary breakpoints introduce unnecessary parameters, and that necessarily increases the overall uncertainty.

There are in general two kinds of evidence that can lead to an expectation of a discontinuity in the data. The first is an examination of station metadata, such as documented station moves or instrumentation changes. For the current paper, the only metadata-based 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 leading to the assumption of 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 break the time series into two segments at that point, 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 the case appropriate for weather records, where spatially correlated data are widely available, has generally been of a more specialized interest. 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 is 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 using an average of the neighboring stations weighted by the square of the correlation coefficient with the target station.

This is compared to the station's time series, and a breakpoint is introduced at places where there is an abrupt shift in mean larger than 4 times the standard error of the mean. This empirical technique results in approximately 1 cut for every 13.5 years of record, which is somewhat more often than the change point 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.

Outlier Weighting

The next potential problem to consider is the effect of outliers, i.e. single data points that vary greatly from the expected value as determined by the local average. We identify outliers by defining the difference, $\Delta_i(t_j)$, between a temperature station's reported data ${T_i}^j$ and the expected value at that same site from our T_{avg} solution:

$$\Delta_i(t_j) = {T_i}^j - b_i - T^j_{avg} - {W_i}^A_j \qquad (12)$$

where ${W_i}^A_j$ approximates the effect of constructing the weather field without the influence of the *i*-th station. (See the supplementary material for the details.) The root-mean-square deviation from average of $\Delta_i(t_j)$ for our analysis with the GHCN data set turns out to be e = 0.62 C. An *outlier* is now defined as a data point for which $|\Delta_i(t_j)| \geq 2.5e$. For the Kriging process, we then apply a deweighting factor to such outliers:

$$w_{ij} = \frac{2.5e}{|\Delta_i(t_j)|} \qquad \text{for outliers} \qquad (13)$$

$$w_{ij} = 1 \qquad \text{otherwise}$$

Equation (13) thus specifies a downweighting term to be applied for point outliers (single temperature measurements at a single site) that are more than 2.5e from the expected value, based on the

interpolated field. This choice of target threshold $2.5e$ was selected with the expectation that it would leave most of the data unweighted. If the underlying data fluctuations were normally distributed, we would expect this process to deweight 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 deweight 2.9% of the data, so the actual weight adjustment rate can be expected to be intermediate between 1.25% and 2.9% for the typical station time series. Of course the goal is not to suppress legitimate data, but rather to limit the impact of erroneous outliers. In defining the weighting function, we downweight data rather than eliminate it; this choice helps to ensure numerical stability.

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 at the same location.

For each station, the average misfit for the entire time series can be expressed as:

$$\varepsilon_i^2 = \frac{1}{N}\sum_j \min\left\{\left(\Delta_i(t_j)\right)^2, 25e^2\right\} \quad [15]$$

Here, the "min" is introduced to avoid giving too great a weight to the most extreme outliers when judging the average reliability of the series; N is the number of terms in the sum. A metric of relative reliability is then defined as:

$$\varphi_i = \frac{2e^2}{e^2 + \varepsilon_i^2} \quad [16]$$

This metric φ_i is used as an additional deweighting factor for each station i. Due to the limits on outliers imposed in the outlier equation (15), this metric has a range between 2 and 1/13, effectively

allowing a "perfect" station to receive up to 26 times the score of a "terrible" station. This functional form was chosen due to several desirable qualities. First, the typical record is expected to have a reliability factor near 1, with poor records being more severely downweighted than good records are enhanced. Using an expression 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 functional forms with similar properties were also considered, and we found that the construction of global temperature time series was largely insensitive to the details of how the downweighting of inconsistent records was handled.

After defining this reliability factor, it is necessary to incorporate this information into the spatial averaging process, i.e. by adjusting the Kriging coefficients. We did this in a way (described in more detail in the supplementary material) that assures that if all of the records in a given region have a similar value of the reliability factor φ_i, then they will all receive a similar weight, regardless of the actual numerical value of φ_i. This behavior is important as some regions of the Earth, such as Siberia, tend to have broadly lower values of φ_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 φ_i, then the individual stations will still receive approximately 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 Tavg.

The determination of the weighting factors is accomplished via an iterative process that seeks convergence. The iterative process generally requires between 10 and 60 iterations to reach the chosen convergence threshold of having no changes greater than 0.001 C in Tavg between consecutive iterations.

Implicit in the discussion of station reliability considerations are several assumptions. Firstly, the local weather field constructed from many station records is assumed be a better estimate of the underlying temperature field than any individual record was. 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 the local comparison

methods to fail. Secondly, it is assumed 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 here. By breaking records into multiple pieces on the basis of metadata changes and/or empirical discontinuities, it creates the opportunity to assess the reliability of each fragment individually. A detailed comparison and contrast of our results with those obtained using other approaches that deal with inhomogeneous data will be presented elsewhere.

Uncertainty Analysis

There are 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 b_i and T_{avg}^{j}, and due to the fact that the data points T_i^j may not be an accurate reflections of the true temperature changes at location 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 spatial coverage.

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 averaging approach on multiple data sets that differ in a way that helps isolate the 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 to execute since it is rare that one can easily construct data sets that isolate the experimental variables without introducing other confounding variations, such as changes in spatial coverage. We will not provide any such analysis of such experiments in this paper; however, additional analysis done by this group is provided as supplemental information (Wickham et al., 2012; Muller et al., 2012). These studies find that objective measures of station quality and urbanization show little or no residuals that survived the deweighting and scalpeling procedures. 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 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 the Hadley/CRU team, 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 do not attempt to consider these additional proposed uncertainties, which will be discussed once future papers have examined the various contributing factors individually.

Statistical Uncertainty

The statistical uncertainty calculation in the current averaging process is intended to capture the portion of uncertainty introduced into due to the noise and other factors that may prevent the basic data from being an accurate reflection of the climate at the measurement site. In order to empirically estimate the statistical uncertainties on the global mean temperature time series T_{avg}, we apply a systematic resampling method known as the "jackknife" method (Miller 1974, Tukey 1958, Quenouille 1949). In this section we give a qualitative overview; more mathematical details are given in the supplement.

This approach is different from the approaches that have been commonly used in the past for historic temperature estimation. 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 a similar amount of data uncertainty to all measurements, even though we would expect that in practice some time series are more accurate than others.

The approach presented here considers the statistical uncertainty quantification from a top-down direction. At its core, this means measuring how sensitive the result is to changes in the available data, by creating subsets of the data and reproducing the analysis on each subset. This allows us to 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. It also helps avoids the possible tendency of scientists to be overly "conservative" in their error estimates, and thereby overestimating the uncertainty in their final result.

The "jackknife" method in statistics is a well-known approach for empirical estimation of uncertainties that can minimize bias in that estimate that might otherwise occur from a small number of events (Quenoille 1949, Tukey 1958, Miller 1974). For climate analysis this feature is important because

there exist regions of the world that are poorly sampled during interesting time periods (e.g. much of the world during much of the 19th century).

The jackknife method is applied in the following way. From the population of stations, we construct eight overlapping subsets, each consisting of 7/8th's of the stations, with a different and independent 1/8th removed from each group. The data from each of these subsets is then run through the entire Berkeley Average machinery to create 8 new estimates, $T_{avg}^{k}(t_j)$, of the average global land temperature vs. time. Following Quenoille and Tukey, we then create a new set of 8 "effectively independent" temperature records $T_{avg}^{k\dagger}(t_j)$ by the jackknife formula

$$T_{avg}^{k\dagger}(t_j) = 8\,T_{avg}^{k}(t_j) - 7\,T_{avg}(t_j) \quad [1]$$

where $T_{avg}(t_j)$ is the reconstructed temperature record from the full (100%) sample. Then the uncertainty estimate for the full sample is the standard error on the mean of the effectively independent samples:

$$\sigma_{jackknife}(t_j) = \frac{\sqrt{\sum_k \left(T_{avg}^{k\dagger}(t_j) - \langle T_{avg}^{k\dagger}(t_j)\rangle\right)^2}}{N} \quad [18]$$

For our example, N = 8. The typical statistical uncertainties estimated using the jackknife are consistent with expectations based on Monte Carlo tests (which we will describe shortly). As the jackknife constructs each temperature average in its sample using a station network that is nearly complete, it is much more robust against spatial distribution biases than simpler sampling techniques where only a small fraction of the stations might be used.

A brief comment should be made here that in computing the subsampled temperature series, $T_{avg}^{k}(t_j)$, the outlier and reliability adjustment factors are recomputed for each sample. This means the process of generating $T_{avg}^{k}(t_j)$ is not entirely linear, and consequently the jackknife estimate in equation [18] is not analytically guaranteed to be effective. However, in the present construction the deviations from linearity are expected to be small since most adjustment factors are approximately 1. This observation, plus the validation by Monte Carlo tests, appear sufficient to justify the use of the jackknife

technique. One could ensure linearity by holding outlier and reliability adjustment factors fixed; however, this would necessarily lead to an underestimate of the statistical uncertainty and require a separate estimate be made of the uncertainty associated with the weighting procedures.

We performed over 10,000 Monte Carlo simulations in order to investigate the relative reliability of the jackknife method as well as other sampling techniques not ultimately used here. For each of these simulations, a toy temperature model of the Earth was constructed consisting of 100 independent climate regions. We simulated noisy data for each region, using a spatial sampling distribution that was chosen to mimic the distribution of the real data. So, for example, some regions had many sites, while other regions had only 1 or 2. The calculation showed that the jackknife method gave a consistently accurate measure of the true error (known since in the toy model by construction) while simpler sampling processes that used non-overlapping subsets tended to consistently underestimated the true error due to spatial biasing.

A practical disadvantage of the jackknife method is that it requires that 7/8 of the data be analyzed in 8 different groups. Since the time to run the program is roughly proportional to the square of the data set size, this means that calculating the uncertainties takes roughly 8 x $(7/8)^2$ = 6 times as long as to calculate the initial time series.

Spatial Uncertainty

Spatial uncertainty measures the error likely to occur due to incomplete sampling of land surface areas. The primary technique we applied to estimate this uncertainty is empirical. The sampled area available at past times is superimposed over recent time periods, and we calculate the error that would be incurred in measuring the modern temperature field given only that limited sample area. For example, if we knew only 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, that exists at each time and superimposing it on the nearly complete temperature anomaly fields that exist at later times, specifically $1960 \leq t_j \leq 2010$ when spatial land coverage approached 100%. Our uncertainty estimate is then the

root-mean-square deviation of the difference. The details are given in the mathematical supplement, as are some of the alternatives we tried.

GHCN Results

As a test of our framework, we applied it to a relatively large set of data that had previously been analyzed by NOAA, the 7280 weather stations in the Global Historical Climatology Network (GHCN) monthly average temperature data set developed by Peterson and Vose [1997] and Menne and Williams (2009). Our analysis used the non-homogenized data set, with none of the NOAA corrections for inhomogeneities included; rather, the scalpel method was applied to break records at the discontinuities. Using the scalpel, the original 7,280 data records were broken into 44,840 record fragments. Of the 37,561 cuts, 6,115 were based on gaps in record continuity longer than 1 year and the rest were found by our empirical method. As the raw data was used, a pre-filtering process was also used to remove a small number of bad data points in the raw data. These include values exceeding 60 C, records filled with zeros, extremely abrupt swings in temperature, and a few other indicators of presumptively bad data. We note that such bad points are often already flagged in the GHCN (or removed when considering quality controlled samples); however, we chose to implement our own similar checks to those existing in the GHCN in order to make our analysis more portable to other datasets and not simply rely on GHCN quality control processes. In total 1.1% 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 in the present paper a decision was made to avoid such pre-existing corrections and examine the ability of the full technique to work with data that was not previously homogenized. A further 0.2% of data was eliminated after cutting and filtering because the resulting record fragment was either too short to process (record length less than 6 months) or it occurred at a time with fewer than 5 total stations active.

The median length of a temperature time series processed by the Berkeley Average was only 5.9 years. Further, the inner 50% range for station record lengths was 2.3 to 10.8 years, and only 4.5% of

records were longer than 30 years. This compares to GHCN data before the scalpel was applied where 72% of the time series are longer than 30 years and the median length is nearly 50 years. As already stated, the current climate change analysis framework is designed to be very tolerant of short and discontinuous records which will allow it to work with a wide variety of data.

Figure 1 shows the station locations used by GHCN, the number of active stations vs. time, and the land area sampled vs. time. The sudden drop in the number of stations around 1990 is largely a result of the methodology used in compiling the GHCN dataset. The present GHCN monthly dataset generally only accepts records from stations that explicitly issue a monthly summary report; however, many stations have stopped reporting monthly results and only report daily ones. Despite this drop, Figure 4(c) shows that the coverage of the Earth's land surface remained around 94%, reflecting the broad distribution of the stations that did remain.

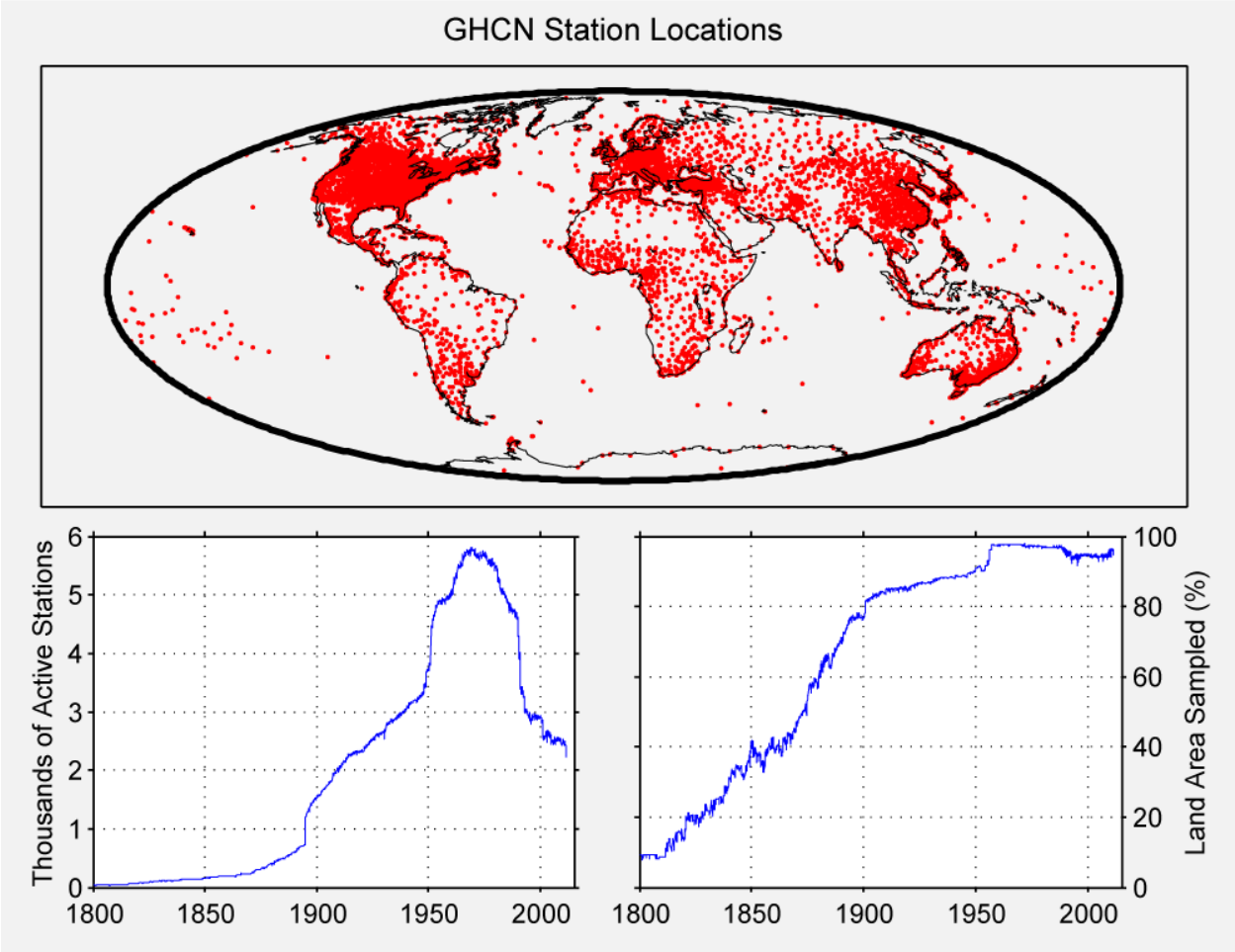

Figure 2. (Upper) Station locations for the 7280 temperature stations in the Global Historical Climatology Network (GHCN) Monthly dataset. (Lower Left) Number of active stations over time. (Lower Right) Percentage of the Earth's land area sampled by the available stations versus time, calculated as explained in the text. The sudden rise in land area sampled during the mid 1950s corresponds to the appearance of the first temperature records on Antarctica.

The results from applying the Berkeley Average methodology to the GHCN monthly data 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 9.290 ± 0.032 C, and temperature average during the most recent decade (Jan 2000 to Dec 2009) was 10.183 ± 0.047 C, an increase of 0.893 ± 0.063 C. The trend line for the 20th century as a whole is calculated to be 0.696 ± 0.099 C/century, well below the 2.74 ± 0.24 C/century rate of global land-surface warming that we observe during the interval Jan 1970 to Nov 2011. All uncertainties quoted here and in the following

discussion are 95% confidence intervals for the combined statistical and spatial uncertainty. In addition, the uncertainty associated with the absolute normalization, discussed below, is omitted unless explicitly stated otherwise.

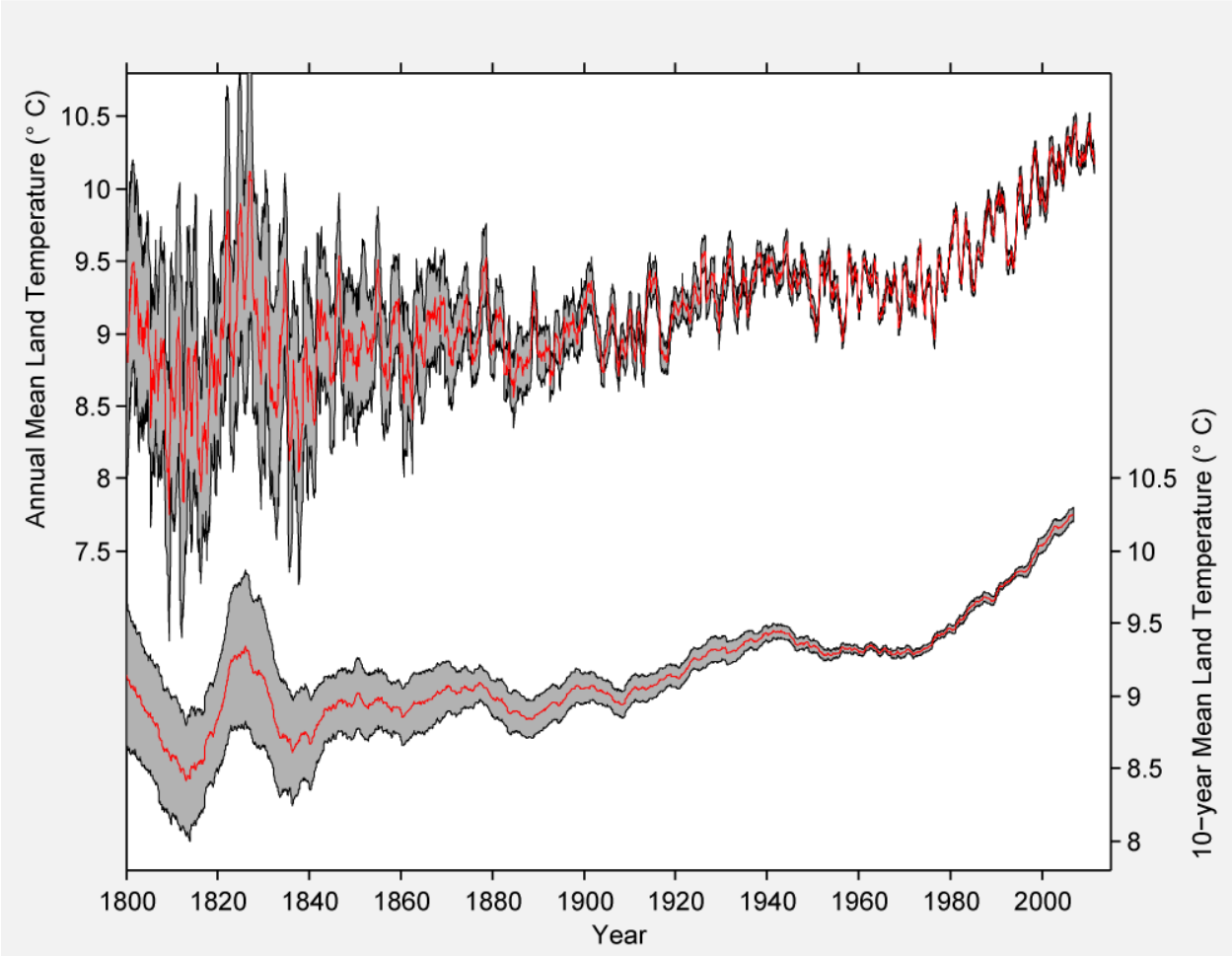

Figure 3. Result of the Berkeley Average Methodology applied to the GHCN monthly data. Top plot shows a 12-month land-only moving average and associated 95% uncertainty from statistical and spatial factors. The lower plot shows a corresponding 10-year land-only moving average and 95% uncertainty. Our 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.

In Figure 8, the land reconstruction is compared to 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 the Berkeley Earth global land average is consistent with the results obtained by these prior efforts. The differences apparent in Figure 8 may partially reflect differences 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 the 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, NASA GISS uses GHCN as the source for ~85% of the time series in their analysis. The remaining 15% of NASA 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. The Hadley/CRU collaboration maintains a separate data set from GHCN for their climate analysis work though approximately 60% of the GHCN stations also appear in their data set.

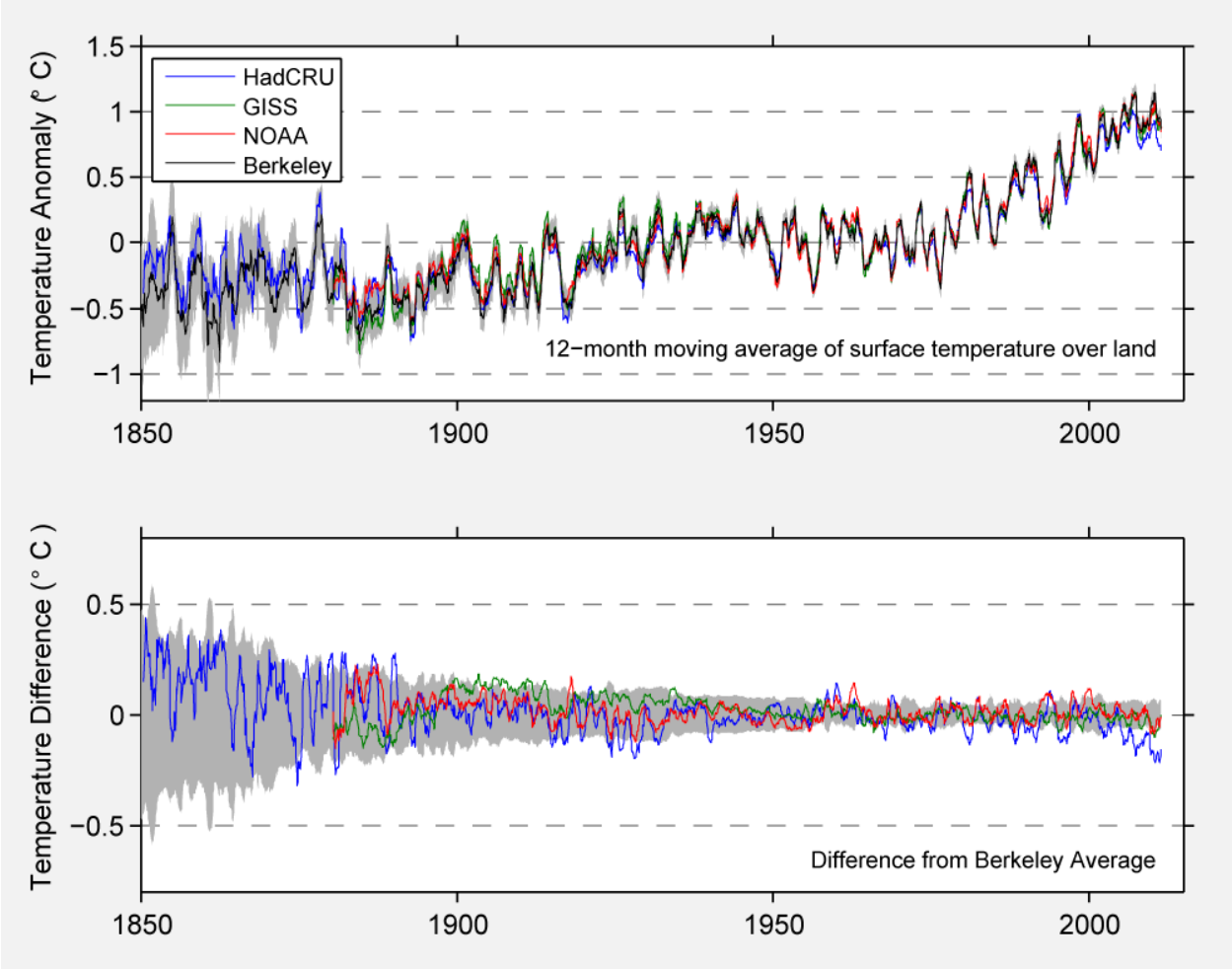

Figure 8. Comparison of the Berkeley Average based on the GHCN data set to existing land-only averages reported by the three major temperature groups. The upper panel shows 12-month moving averages for the four reconstructions, and a gray band corresponding to the 95% uncertainty range on the Berkeley average. The lower panel shows each of the prior averages minus the Berkeley average, as well as the Berkeley average uncertainty. Hadley/CRU collaboration has a systematically lower trend than the other groups for the most recent portion of the land-only average. Berkeley is very similar to the prior results during most of the range. After considering the uncertainties associated with the other reconstructions (not shown) it is likely that the Berkeley result is consistent with the other results nearly everywhere. For all curve the interval 1951-1980 was defined to be zero.

The uncertainty limits we obtained are plotted on the results shown in the preceding figures by the grey bands. The uncertainties are remarkably small, particularly in the recent past, but remain relatively small even in the interval 1800 to 1850, a time period that was not previously reported by other groups. Recall that the statistical uncertainties are estimated by subdividing the data into smaller groups and then intercomparing them, and the spatial sampling uncertainties are estimated by re-running the data set for the modern era with the same limited sampling available for the earlier data. Thus our uncertainties

are empirical, estimated by the behavior of the data, rather than theoretical based on estimates of initial uncertainties. If our analysis technique had severe biases (e.g. ignoring data that should have been included, or failing to deweight spurious data) then the uncertainties estimated using our empirical method would be larger.

Because of the relatively small values on the uncertainties, it is worthwhile to look in some detail at the actual empirical calculation of these uncertainties. We described how a determination of statistical uncertainties can be made by using the jackknife method. Figure 6 shows the results of applying a conceptually simpler subsample approach for the GHCN data set. Five completely independent subsamples were constructed each containing a random 20% of the stations, and these were then processed via the Berkeley Average methodology. The agreement in Figure 6 of the averages constructed from these independent subsamples make it clear that the current averaging procedure is robust against noise on individual data and the inclusion / exclusion of individual stations, and provides a reasonable estimate of the uncertainty arising from these factors.

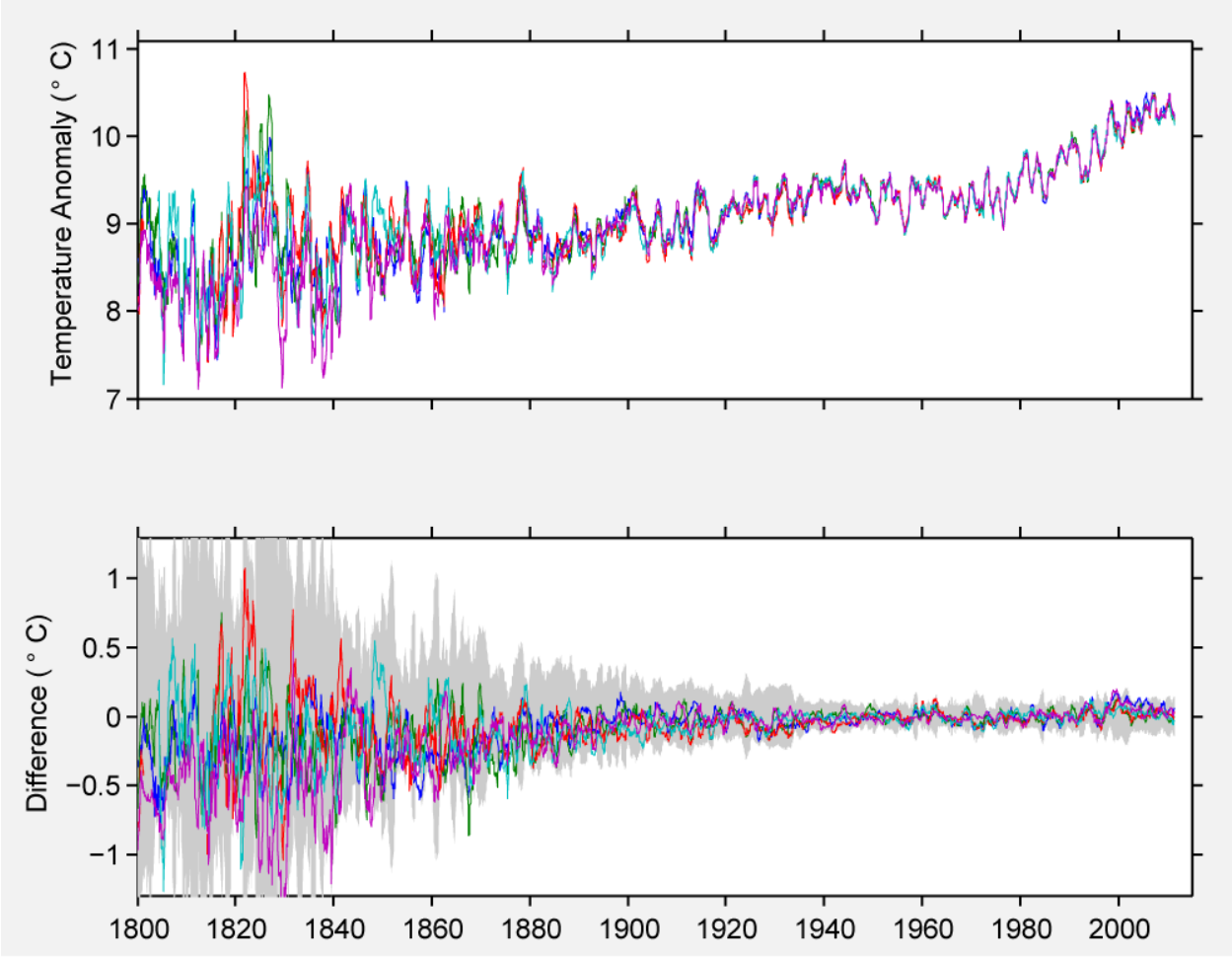

Figure 6. Five independent temperature reconstructions each derived from a separate 20% of the GHCN stations. These stations are statistically independent, although they ultimately sample the same global climate. The upper figure shows the calculation of the temperature record based on five independent subsamples. The lower plot shows their difference from the 100% result, and the expected 95% uncertainty envelope relative to zero difference. The uncertainty envelope used here is scaled by $\sqrt{5}$ times the jackknife calculated statistical uncertainty. This reflects the larger variance expected for the 20% samples.

To understand the spatial uncertainty, we show in the top panels of Figure 7 the sampled regions available in the years 1800 and 1860. At future times, it is possible to compare the apparent average over these sampled regions with the average calculated over all land areas. In the middle panels of Figure 7, time series capturing this difference are shown. In the case of the year 1800 sample region, the global average quickly diverges as new stations are added, but nonetheless it remains within a characteristic difference of the estimated global mean during the entire record. The difference trend for the sample region from the year 1860 is slower to diverge since the expansion in coverage does not occur as quickly.

However, it also remains within a characteristic bound. Evidence such as this is used as described in the supplement to estimate the spatial uncertainty.

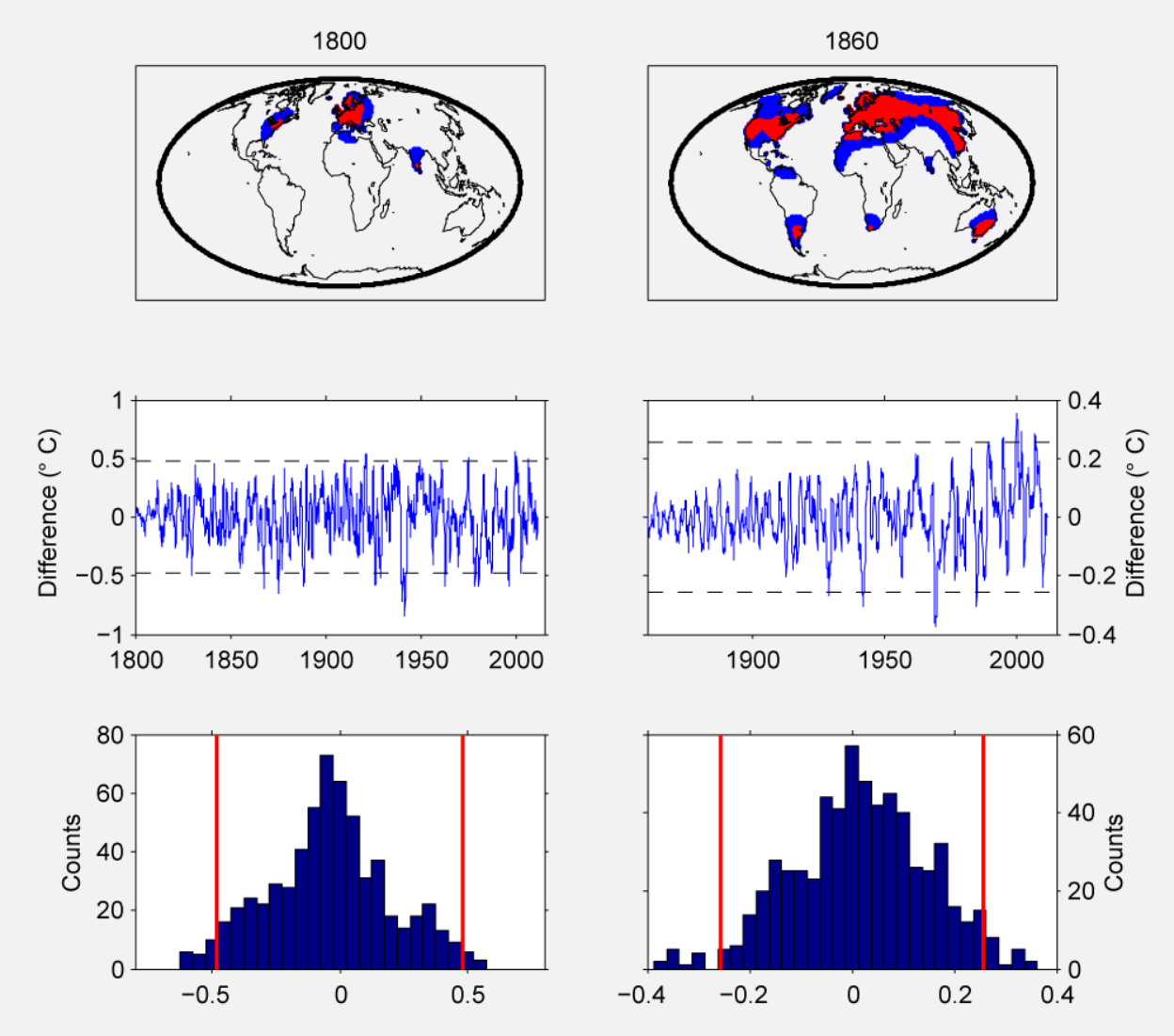

Figure 4: Panels A and D shows the spatial coverage available in 1800 and 1860 respectively. Red indicates greater than 80% local completeness, while blue indicates greater than 40% to 80% completeness. The metric for measuring sample coverage is explained in the supplement. These coverage patterns are applied to the temperature field at latter times to estimate the effect of incomplete spatial sampling. Panels B and E show the difference between the full land average at each latter time and the average that would obtained by using only the coverage regions indicated in Panels A and C, respectively. This difference is shown here using the 12-month moving average, and horizontal bars indicate the 95% uncertainty level. Panels C and F shows histograms of the fluctuations in panels B and E over the interval 1960 to 2010. Only about 5% of months exceed the 95% threshold, as expected.

As is shown in Figure 5, the current record is extended all the way back to 1800, including 50 more years than Hadley/CRU group and 80 more years than NOAA and NASA GISS. We feel this extension is justifiable; although the uncertainties are large, there is interesting and statistically significant

structure that can be seen. The analysis technique suggests that temperatures during the 19th century were approximately constant (trend 0.18 ± 0.45 C/century) and on average 1.27 ± 0.21 C cooler than the interval 2000-2009. Circa 1815 there is a negative temperature excursion that happens to roughly coincide with both a period of major volcanism as well as the Dalton Minimum in solar activity. Two very large volcanic eruptions occurred about this time: a large unidentified event in 1809 (evidence seen in ice cores; Wagner and Zorita 2005), and the Tambora eruption in 1815 — the largest eruption in the historical era, blamed for creating the "year without a summer" (Oppenheimer 2003; Stothers 1984). 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 forcing 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 nature of the excursion.

Our empirical uncertainty estimates are remarkably small in the period 1800 to 1850, a time when there were often no measurements whatsoever in the Southern Hemisphere. Our calculation assumes that the regional fluctuations in the Earth's climate system during the entire study interval have been similar in scale to those observed in the reference period 1960 to 2010. For example, we note that the combined temperature anomaly over Eastern North America and Europe stayed within 0.5 C of the global land average anomaly 95% of the time during the 20th century. The temperature time series presented here should adequately capture the temperature fluctuations at early times in the well-sampled regions (i.e. Europe and Eastern North America), however, we rely on the spatial uncertainty calculation to further estimate of how different the sample region may have been from the whole globe.

To look for possible changes in the structure with time, we show the spatial structure of the climate change during the last century Figure 7. The structure is fairly uniform, though with greater

warming over the high latitudes of North America and Asia, consistent with prior results (e.g. 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.

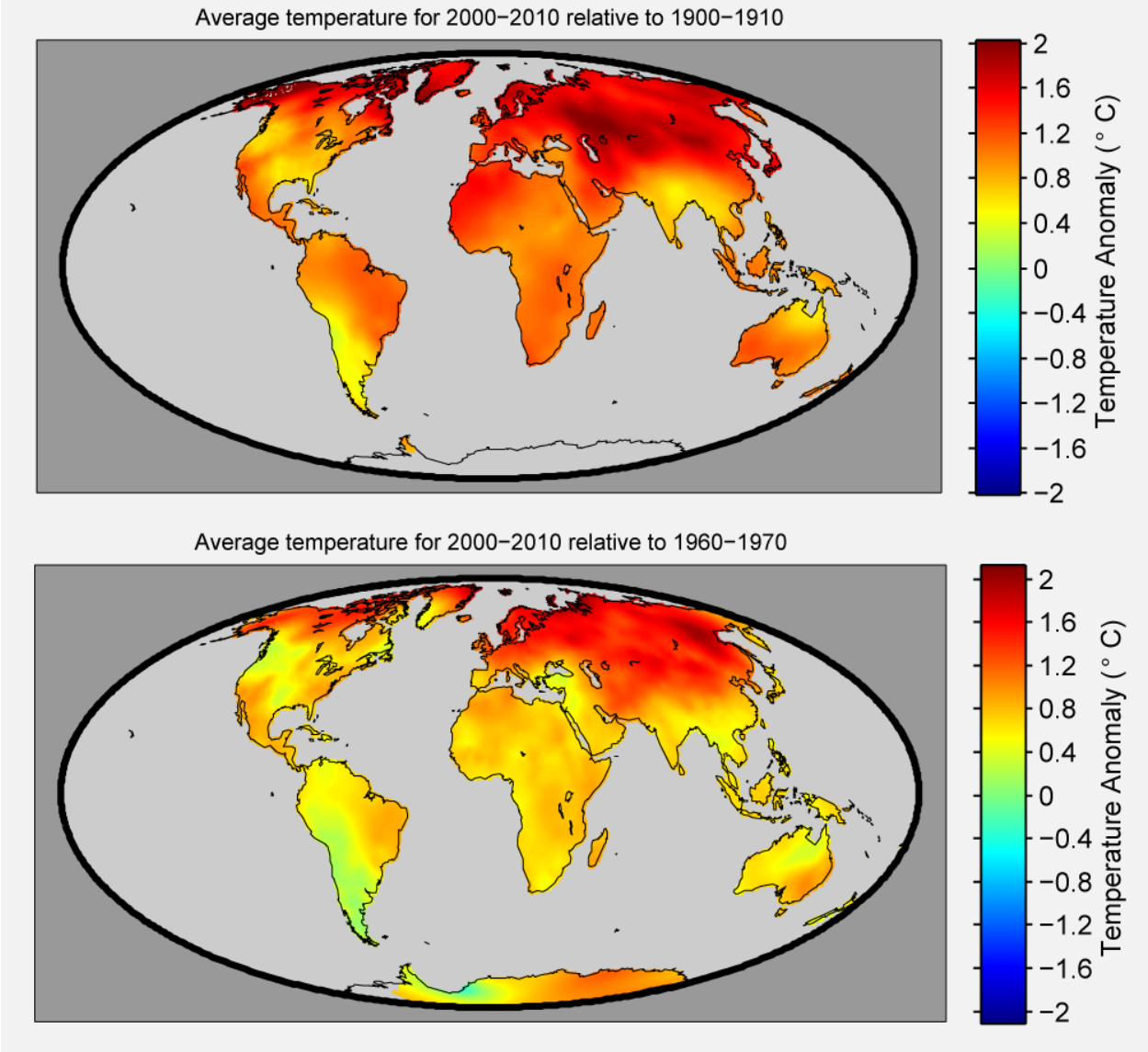

Figure 7. Maps showing the decadal average changes in the land temperature field. In the upper plot, the comparison is drawn between the average temperature in 1900 to 1910 and the average temperature in 2000 to 2010. In the lower plot, the same comparison is made but using the interval 1960 to 1970 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.

Discussion of Relative Uncertainty Estimates

As discussed above, the uncertainty in the current results are 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, 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. These uncertainties for the GHCN analysis are presented in Figure 10.

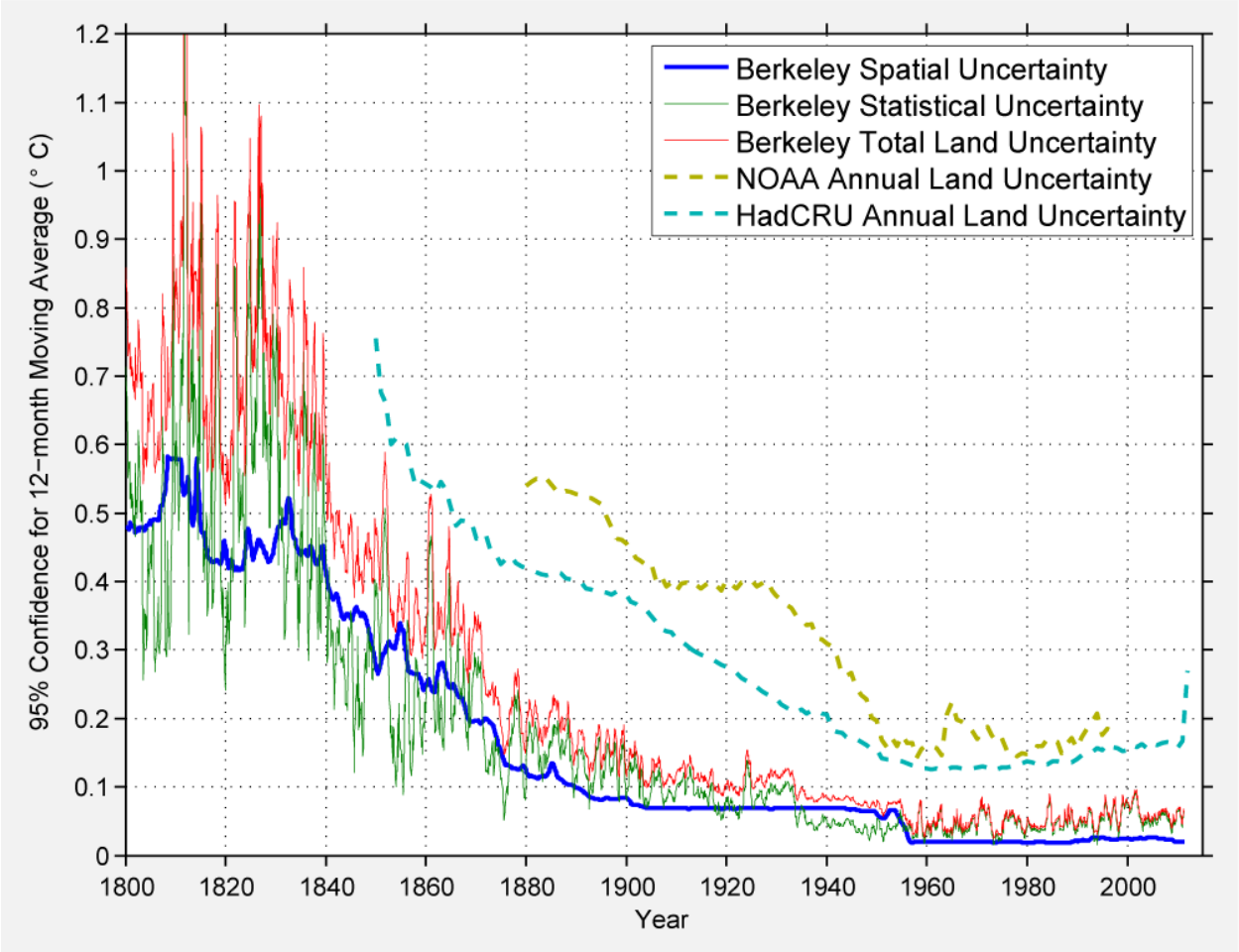

Figure 10. The 95% uncertainty on the Berkeley Average (red line) and the component spatial (blue) and jackknife statistical (green) uncertainties for 12-month moving land averages. From 1900 to 1950, the spatial uncertainty is dominated by the complete lack of any stations on the Antarctic continent. From 1960 to present, the largest contribution to the statistical uncertainty is fluctuations in the small number of Antarctic temperature stations. For comparison, the land-only 95% uncertainties for Hadley/CRU collaboration and NOAA are presented. As discussed in the text, in addition to spatial and statistical considerations, the Hadley/CRU collaboration 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.

Note that the two types of uncertainty tend to co-vary. 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). The step change in spatial uncertainty in the middle 1950s is driven by the introduction of the first weather stations to Antarctica. 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 nearly 9% of the Earth's land area and yet GHCN provides fewer than 40 stations from Antarctica, and only 3 from the interior. To the extent that these few available records disagree with each other they can serve as a relatively large source of statistical noise.

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 since the 1960s. Again, the decrease in station counts is essentially an artifact of the way the GHCN monthly data set has been constructed. In fact, once one includes daily as well as monthly monitoring reports, the true density of weather 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.

Over much of the record, the Berkeley uncertainty calculation yields a value 50-75% lower than that reported by other groups. As the sampling curves demonstrate (Figure 6), the reproducibility of the temperature time series on independent data is extremely high which justifies 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.

In comparing the results one must note that curves by prior groups in Figure 10 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 Berkeley Earth does not include comparable factors, this could explain some of the difference. However, the "bias" corrections being used cannot explain the bulk of the difference in estimated uncertainty. The Hadley/CRU collaboration 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 less than the difference between the Berkeley Earth uncertainty estimates and the prior estimates. The present techniques estimate the global land-based temperature with considerably less apparent spatial and statistical uncertainty than prior efforts.

A large portion of the difference in uncertainty may be related to systematic overstatement of the uncertainty present in the prior work, as a result of the prior groups being conservative in their error estimates. For example, when grid cells are missing in the Hadley/CRU collaboration reconstruction, they are assumed to be completely unknown (and hence contribute large uncertainty), even though populated cells may exist nearby. Ignoring the relationships between grid cells make the calculations easy to understand but can lead to overstating the uncertainties. Similarly, processes that consider each grid cell individually, are likely to do a poorer job of identifying homogeneity breaks or establishing relative alignments amongst records. Such factors will also increase the apparent uncertainty. The results of prior groups actually agree significantly more closely than would be expected given their stated uncertainties and different averaging approaches. This could be evidence that prior uncertainties were overstated, though some of the similarity in results is undoubtedly also attributable to the overlapping datasets being used.

In considering the Berkeley Earth uncertainties, it must be acknowledged that adding some degree of bias error may ultimately be necessary. The integrated reliability assessment procedures and the use of the scalpel are expected to significant reduce the potential impact of many forms of bias by

detecting local stations that are inconsistent with regional averages, and allowing iterative adjustments to be made. The effectiveness of these procedures will be addressed via further publications. Finally, we note that as the number of stations gets low there is an increased potential for systematic bias, as could 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 can increase.

Climatology

In Equation 4, we defined the local temperature at position and time $\vec{x}_i, t_j$ to be given by

$$T(x,t) = T_{avg}(t) + C(x) + W(x,t)$$

where $C(\vec{x}_i)$ is the approximately time-invariant long-term mean temperature of a given location, sometimes referred to as the *climatology*. A map of the $C(\vec{x}_i)$ that we obtain from our fit is shown in Figure 11. The global land average from 1900 to 2000 is 9.35 ± 1.45 C, broadly consistent with the estimate of 8.5 C provided by Peterson et al. (2011). This large uncertainty in the normalization is not included in the shaded bands that we put on our T_{avg} plots, as it only affects the absolute scale and doesn't affect relative comparisons. In addition, most of this uncertainty is due to the presence of only three GHCN sites in the interior of Antarctica, which leads the algorithm to regard the absolute normalization for much of the Antarctic continent as poorly constrained. Preliminary work with more complete data from Antarctica and elsewhere suggests that additional data can reduce this normalization uncertainty by an order of magnitude without changing the underlying algorithm. 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.

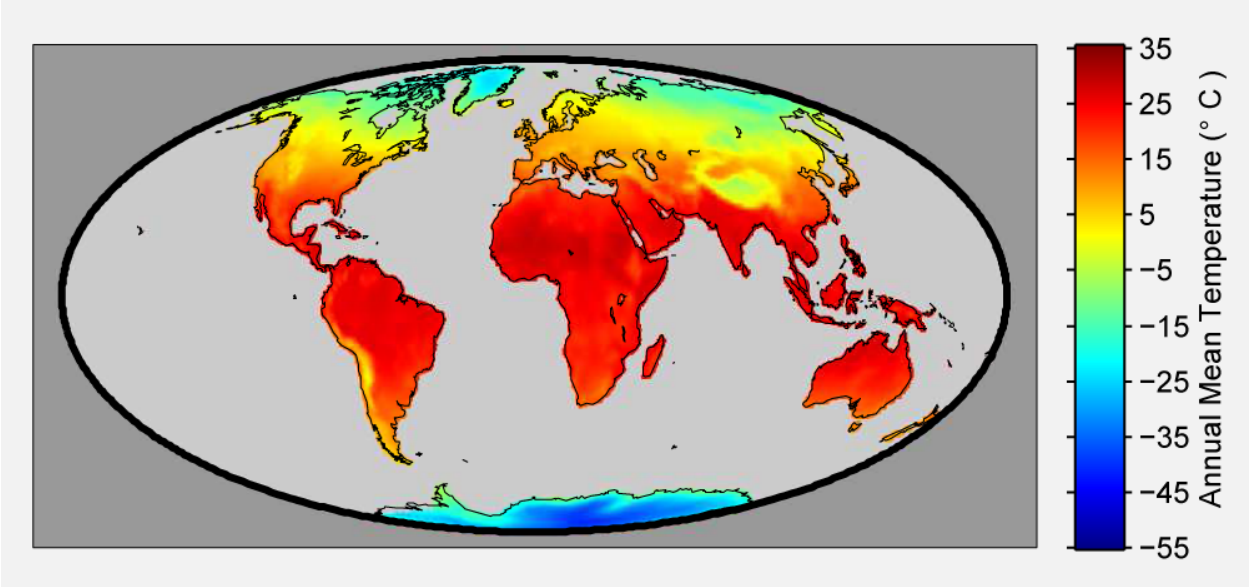

Figure 12. A map of the derived Climatology term, $\boldsymbol{C}(\vec{\mathbf{x}}_\mathbf{i})$. 95% of the variation is accounted for by altitude and latitude. Departure from this is evident in Europe and in parts of Antarctica.

The temperature field is nearly symmetric with latitude, though with a positive excess around the Sahara, and asymmetry between the Arctic and Antarctic. The change of surface temperature with elevation is found to be almost exactly linear, even though our fit included a quadratic term. The cumulative effect is that one can predict the mean temperature of the typical station based solely on its location to within ±1.6 C at 95% confidence. However, there are also long outliers on both sides of the fit, which may indicate stations that are inaccurately located in either latitude or altitude. A comparison of these results to climatology maps produced by ERA40 (Uppala et al. 2005) find these results to be very similar, except in regions of rapidly changing topography such as the Andes or Himalaya, where differences of several degrees can occur. Given that neither the current method nor the ERA40 climate models can fully resolve rapidly varying topography, it isn't immediately clear which system is likely to be more accurate in those regions.

Discussion

This paper describes a new approach to global temperature reconstruction. Spatially and temporally diverse weather data exhibiting varying levels of quality were used to construct an estimate of

the mean land-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 avoiding short records, as is necessary for most prior 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 a grid. This fact allows us, in principle, to avoid a variety of noise and bias effects that can be introduced by gridding. 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 must be made between grid resolution and statistical precision.

That said, the integrals required to compute *Tavg* will in general need to be computed numerically, and computation of the Kriging coefficients 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. 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 dataset created by the NOAA group, using the raw data 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). Instead, 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 prior groups, who used the same or similar data and full homogenization procedures. 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. The temperature variability on the decadal time scale is lower now than it was the in the early 1800s. One large negative swing, between 1810 and 1820, is coincident with both volcanic eruptions at that time (including Mt. Tambora) and the Dalton Minimum in solar activity.

We chose to analyze the NOAA dataset largely as a test that allows us to make a direct comparison with a prior analysis, without introducing issues of data selection effects. 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 35,000 stations, more than 5 times larger than the data set used by NOAA.

Acknowledgements

We thank David Brillinger for many helpful conversations and key suggestions that helped lead to the averaging method presented in this paper. We also thank Zeke Hausfather, Steve Mosher, and Stephen Hodgart for useful comments on this work. 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. Alexander L. V., 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.
2. 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.
3. 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.
4. Cressie, Noel. "The Origins of Kriging." *Mathematical Geology*, Vol. 22, No. 3, 1990.
5. 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.
6. 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.
7. Easterling, D. R. & Wehner, M. F. (2009) "Is the climate warming or cooling?" Geophys. Res. Lett. 36, L08706.

8. 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.
9. 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
10. 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.
11. Hansen, J., R. Ruedy, Mki. Sato, and K. Lo, 2010: Global surface temperature change. *Rev. Geophys.*, **48**, RG4004, doi:10.1029/2010RG000345.
12. Hansen, J.E., and S. Lebedeff, 1987: Global trends of measured surface air temperature. *J. Geophys. Res.*, **92**, 13345-13372, doi:10.1029/JD092iD11p13345.
13. Hinkley, D. V. (1971), "Inference about the change-point from cumulative sum tests," Biometrika, 58 3, 509-523.
14. 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.
15. 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.
16. Jones, P.D., T.M.L. Wigley, and P.B. Wright. 1986. Global temperature variations between 1861 and 1984. *Nature*322:430-434.
17. Journel, A. G. *Fundamentals of geostatistics in five lessons*. American Geophysical Union, 1989; 40 pages.
18. 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.
19. 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.

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 (2012) "Earth Atmospheric Land Surface Temperature and 1 Station Quality", included as supplementary material; submitted to 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. (1977), Exploratory Data Analysis (Addison-Wesley, New York, 688 pages)
38. Tukey, J.W. (1958), “Bias and confidence in not quite large samples”, *The Annals of Mathematical Statistics*, 29, 614.
39. Uppala, S.M., Kållberg, P.W., Simmons, A.J., Andrae, U., da Costa Bechtold, V., Fiorino, M., Gibson, J.K., Haseler, J., Hernandez, A., Kelly, G.A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R.P., Andersson, E., Arpe, K., Balmaseda, M.A., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B.J., Isaksen, L., Janssen, P.A.E.M., Jenne, R., McNally, A.P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N.A., Saunders, R.W., Simon, P., Sterl, A., Trenberth, K.E.,

Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J. 2005: The ERA-40 re-analysis. Quart. J. R. Meteorol. Soc., 131, 2961-3012.doi:10.1256/qj.04.176

40. 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
41. 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.
42. Wickham, Charlotte, Robert Rohde, Richard Muller, Robert Jacobsen, Saul Perlmutter, Arthur Rosenfeld, Jonathan Wurtele, , Don Groom, Judith Curry (2012) "Influence of Urban Heating on the Global Temperature Land Average Using Rural Sites Identified from MODIS Classifications", included as supplementary material; submitted to JGR.
43. 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.