# Berkeley Earth Temperature Averaging Process — supplement on statistical and mathematical methods

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

This supplement is meant for those who wish to duplicate the methods that we describe in the main paper. Perhaps even more helpful is the detailed computer code, which is available online at www.BerkeleyEarth.org. This code is written in the programming language Matlab, one that we consider relatively transparent even to those who have not studied this particular language. Because this supplement goes into great detail, we will use a more formal notation than we used in the descriptive paper. Unfortunately this notation, to include all the subtleties, works best if it is not identical to the notation used in the main paper; that paper was written for clarity; this supplement is written to make it easier for a programmer to duplicate. In addition, we will adopt the standard notation used in statistics, in which a hat above a quantity signifies that it is an estimate of a parameter, not the true value.

Let $T(\vec{x},t)$ be the true global temperature field in space and time, a field that we will never know but that we will try to estimate. Define the decomposition:

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

These quantities are similar to the ones we defined in the main paper. Note that we are using the quantity $\theta(t)$ to represent the same quantity that we designated $T_{avg}$ in that paper. $\theta(t)$ refers to the true value of the global average, while we will subsequently use a hat notation, $\hat{\theta}(t)$, to indicate our estimate of the average. In order to make the decomposition unique, we specify the following additional constraints:

$$\int_{\text{Earth's surface}} C(\vec{x}) d\vec{x} = 0,$$

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

$$\int_{\text{Earth's surface}} W(\vec{x}, t) dt = 0, \text{ for all locations } \vec{x}$$

In the main paper we derived these from a simple Kriging model. In this supplement we simply describe them as constraints on the temperature function.

Given this decomposition, $\theta(t)$ 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. We will estimate the global temperature field by simultaneously constraining all three pieces of $T(\vec{x}, t)$ using the available data. A summary of all the key symbols may be found in the Appendix at the end of this supplement.

As this study is based solely on the use of land-based temperature data, the spatial integrals in equation [2] shall be restricted to only the Earth's land surface. As a result, we 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, the algorithm described below shall 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 break $C(\vec{x})$ into a number of additional components:

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

Here $\lambda$ 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. The $G(\vec{x})$ will include many large-scale climate patterns, such as the effects of the

Gulf Stream. With appropriate models for $\lambda$ 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 $\lambda$, $h$, 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, let:

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

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. Further:

$$b_i \approx C(\vec{x}_i) \quad [5]$$

The difference between the expected climatology at each station and apparent baseline, $b_i$, can often be 1-2 C due to misreported station locations, instrumental biases, and local-scale features not captured by the climatology field. This motivates us to measure the apparent baseline at each station as part of our analysis procedure, rather than relying on a smooth climatology to extract this information. Rather, the analysis procedure presented here does the reverse. The smooth climatology field is inferred from the apparent baselines at each station.

As mentioned earlier, for each of the parameters and fields discussed we adopt a "hat" notation, e.g. $\hat{\theta}(t_j)$, $\hat{b}_i$, to denote values that are estimated from data and 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. The minimization might 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

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

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. The weights, $\omega_{i,j}$, are initially set all equal to 1. However, as discussed in the section on outlier detection, a small fraction of these weights are adjusted to remove apparent outliers.

To determine the other fields, it is instructive to consider the properties that one would expect the "weather" field, $\widehat{W}(\vec{x}_i, t_j)$, to have. To begin, it should have (at least approximately) zero mean over space and time in accordance with equation [2]. Secondly, the 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 interpolation technique provides best linear unbiased estimator of an underlying spatial field. This technique is also sometimes known as Gaussian Process Regression.

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

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

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

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

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

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

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

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

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. We will adopt an approach to estimating the covariance matrix that preserves the natural spatial considerations provided by Kriging, but also shares characteristics with the local averaging approach adopted by NASA 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:

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

Where $a_1 \ldots a_N$ denotes the collection of stations active at time $t_j$. This leads to the Kriging-based description of the weather:

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

Like the $\omega_{i,j}$ in equation [7], the $\xi_{i,j}$ are adjustment factors described in the section on dealing with outliers, these factors are set to 1 initially and remain near 1 in most cases. Thus the “weather” field is constructed as a spatially-weighted linear combination of the fluctuations in the data $d_i(t_j)$ relative to the global trend $\hat{\theta}(t_j)$ and the station’s baseline $\hat{b}_i$.

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, the Kriging estimate of the field will necessarily approach the underlying field exactly as the density of data increases. This feature of Kriging contrasts with the NASA GISS and Hadley / CRU averaging approaches which smooth over fine structure.

A further modification is made by assuming that $R(\vec{a},\vec{b})$ can be approximated as $R(d)$, where $d = |\vec{a} - \vec{b}|$ denotes the distance between $\vec{a}$ and $\vec{b}$. This allows a parameterization the correlation field as a simple function of one variable, though doing so admittedly neglects differences in correlation that might be related to spatially varying factors such as latitude, altitude, and local vegetation, etc. The correlation function is parameterized using the “spherical” correlation function (Kitandis 1997, Isaaks and Srivastava 1989):

$$R(d) = \begin{cases} \alpha \left(1 - \frac{d}{d_{max}}\right)^2 \left(1 + \frac{d}{2\, d_{max}}\right) + \mu & \text{for all } d < d_{max} \\ 0 & \text{otherwise} \end{cases} \quad [14]$$

The free parameters $\alpha, d_{max},$ and $\mu$ are determined by fitting this functional form to a reference data set created by randomly selecting 500,000 pairs of stations that have at least ten years of overlapping data, and measuring the correlation of their non-seasonal temperature fluctuations as function of distance. 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 functional form of equation [14] is one a small number of families of analytic functions known to describe valid correlation matrices over a sphere (Huang et al. 2011). The properties of a correlation matrix require that it must always be positive definite for all possible sets of points in the space being considered, which restricts the functional forms that $R(d)$ can have. The functional form presented here was chosen because it gave the best fit to the reference data among known families of valid functions.

The small constant term $\mu$ in equation [14] 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 by rescaling the rest of equation [14] by $1/(1 - \mu)$ to compensate near $d = 0$. This allows us to treat stations at distances greater than $d_{max}$ 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 ($R(1000\text{ km})^2 = 0.24$), and non-trivial up to ~1800 km ($R(1800\text{ km})^2 = 0.05$) from each site. It should be emphasized that this is the correlation structure of the monthly average field. On shorter timescales the correlation length will also generally be shorter, and it is only by choosing to work with monthly data that we are able to observe a field which is relatively smooth over such substantial distances.

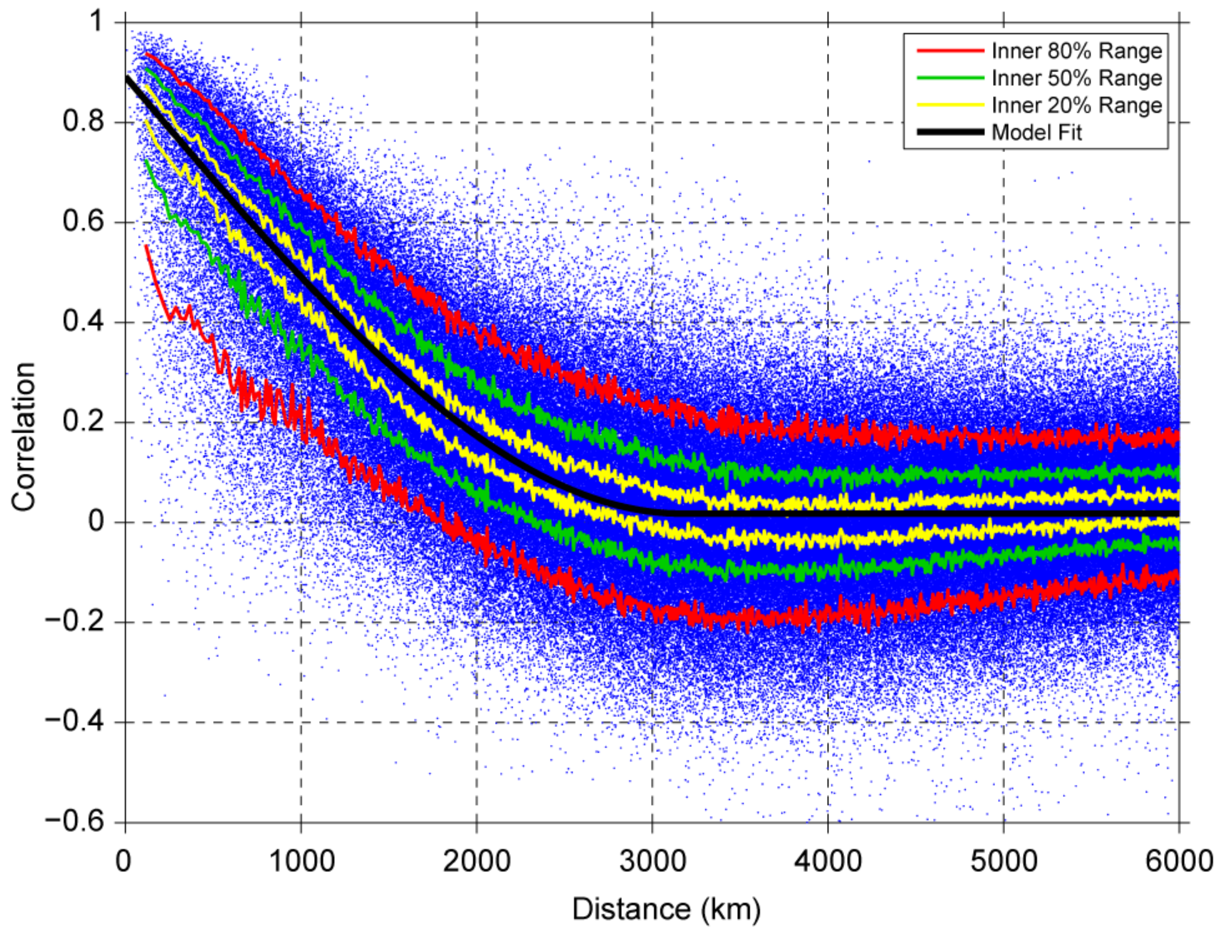


**Figure 1.** Mean correlation versus distance curve constructed from 500,000 pair-wise comparisons of station temperature records. Each station pair was selected at random, and the measured correlation was calculated after removing seasonality and with the requirement that they have at least 10 years 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.

Based on the data, the best fit values in equation [14] were $\alpha$ = 0.8741, $d_{max}$ = 3163.5 km, and $\mu$ = 0.0180. These were the values we used in the Berkeley Earth temperature reconstruction method.

Though these values were used in the averaging model, the global scale results were found to be quite insensitive to the specific parameter choices in the correlation function. Experiments where $d_{max}$ was adjusted by large factors (e.g. +100% or -50%) were conducted and the changes in the global annual average were generally smaller than or similar to the uncertainties arising from other factors. This suggests that the Berkeley averaging method is relatively insensitive to the details of $R(d)$. This is not surprising given that the separation distance between stations is often much less than the effective correlation length.

In Figure 3 we show similar fits to Figure 2 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 which had reduced correlation at short distances. This band is one of several that included relatively few temperature stations as it spans much of the Atlantic Ocean, and so this deviation might be primarily a statistical fluctuation. However, the deviations observed in Figure 3 for latitude bands are more meaningful. The 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. Though not shown, we also find that the East-West correlation length is about 18% greater than the North-South correlation length. This is consistent with the fact that weather patterns primarily propagate along East-West bands.

The variations discussed above, though non-trivial, are relatively modest for most regions (except perhaps at the equator). As previously noted, when considering large-scale averages the Kriging process described here is largely insensitive to the details of the correlation function, so it is expected that small changes in the correlation structure with location or orientation can be safely ignored. Hence, the current construction applies only the simple correlation function given by equation [14]. However, developing an improved correlation model that incorporates additional spatial variations is a likely topic for future research.

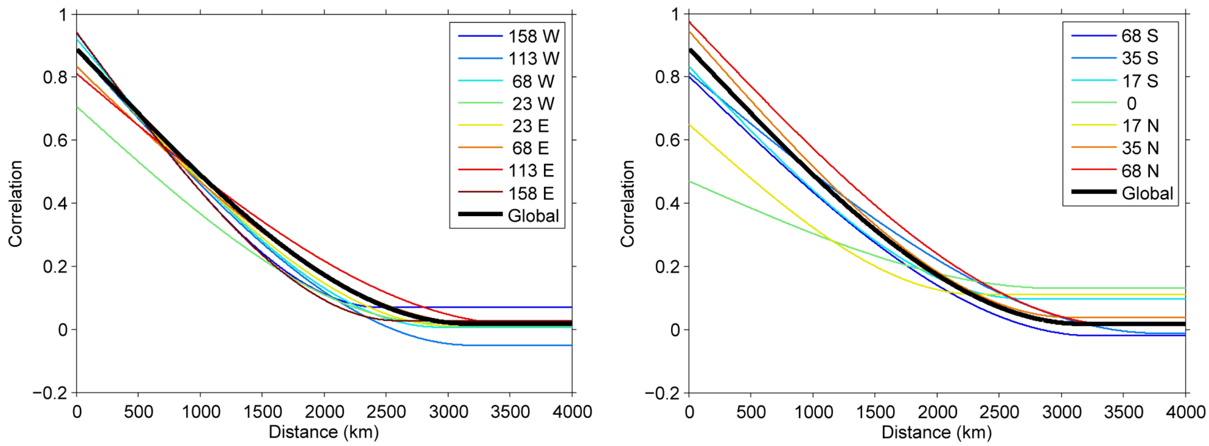


Figure 2. Correlation versus distance fits, similar to **Figure 2**, but using only stations selected from portions of the Earth. The Earth is divided into eight longitudinal slices (Left) or seven latitudinal slices (Right), with the slice centered at the latitude or longitude appearing in the legend. 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.

We note that the correlation in the limit of zero distance, $R(0) = 0.874$, has a natural and important physical interpretation. It is an estimate of the correlation that one expects to see between two typical weather stations 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 $1 - R(0) = 12.6\%$ of the non-seasonal variation in the typical station record is caused by noise processes that are 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 noise for the typical month at a typical station is ~0.49 C at 95% confidence. It must be emphasized that such estimates of noise incorporate all the variations that exists between stations, including those attributable to different instrumentation, different measurement procedures, different instrumental settings or microclimates, etc. Further, this estimate is also influenced by both historical and regional differences in the way temperature has been measured. Prior studies on the reproducibility of temperature observations using consistent instrumentation have generally reported much greater precision, e.g. +/- 0.06 C (Folland et al. 2001), so it is likely that most of the noise we report here is due to differences in instrumentation and measurement approaches. For example, a station that reports mean temperature by calculating the simple of average of max and min extremes may vary considerably from stations that average data recorded hourly, or via other processes. Our results suggest that estimates based on identical instrumentation and methods fail to capture most of the noise that actually exists in the historical weather observing system. However, others authors generally assign a large uncertainty to the homogenization process (e.g. 0.8 C in Folland et al. 2001). We suspect that the large uncertainty they associate with homogenization essentially captures much the same short-term noise that we observe.

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 $\frac{\iint R(\vec{x})^2 d\vec{x}}{\iint 1 d\vec{x}} = 0.43\%$ of the Earth's temperature field; equivalently, 235 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 235 stations includes the effect of measurement noise. Removing this consideration, we would find that the underlying monthly mean temperature field has approximately 185 independent degrees of freedom. It should be emphasized that such estimates apply to the monthly mean field, which is inherently much smoother than the daily average or instantaneous field. In practice though, quality control and bias correction procedures will substantially increase the number of records required for adequate constraint of uncertainties.

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:

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

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 the correlation matrix is positive definite, it follows that:

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

Here $F(\vec{x}, t_j)$ 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 estimate is

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

it follows that in the limit of dense coverage, $F(\vec{x}, t_j) \rightarrow 1$, the temperature estimate at $\vec{x}$ depends only on the local data $d_i(t_j)$, while in the limit of no coverage, $F(\vec{x}, t_j) \ll 1$, the temperature field at $\vec{x}$ is simply assumes the same value as the global average of the data. For diagnostic purposes it is also useful to define:

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

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:

$$\int \widehat{W}(\vec{x}, t_j)^2\, d\vec{x} \quad [19]$$

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

$$\int \widehat{W}(\vec{x}, t_j)\, F(\vec{x}, t_j)\, d\vec{x} = 0 \quad [20]$$

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

$$\int \widehat{W}\left(\vec{x}, t_j\right) d\vec{x} = 0 \qquad [21]$$

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 $\widehat{W}\left(\vec{x}, t_j\right)$ also becomes 0 in this limit. We choose to use equation [21] as our fundamental constraint equation rather than equation [20]. This implies that our solution will be similar but not identical to the ordinary Kriging solution in the spatial mode; however, taking this approach 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.

Once $\widehat{W}\left(\vec{x}, t_j\right)$ has been found via equation [21], it is possible to evaluate $\int \widehat{W}\left(\vec{x}, t_j\right) F\left(\vec{x}, t_j\right) d\vec{x}$ as an estimate of the error introduced via this approximation. Given the relationship to the minimization constraint, these integrals provide a direct estimate of the error this approximation introduces in $\hat{\theta}\left(t_j\right)$. In the case of the GHCN data, the difference created by using equation [21] rather than equation [20] is nearly always much less than the uncertainty arising from other effects, and hence does not appear to be significant.

Given equations [13] and [21] it follows that:

$$\hat{\theta}\left(t_j\right) = \frac{\sum_{i=1}^{N} \xi_{i,j} \left(\int S_{a_i}\left(\vec{x}, t_j\right) d\vec{x}\right)\left(d_{a_i}\left(t_j\right) - \hat{b}_{a_i}\right)}{\sum_{i=1}^{N} \xi_{i,j} \left(\int S_{a_i}\left(\vec{x}, t_j\right) d\vec{x}\right)} \qquad [22]$$

Combined with equation [7] this 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 subtraction from 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:

$$C(\vec{x}_i) = \lambda(\vec{x}_i) + h(\vec{x}_i) + G(\vec{x}_i) \approx \hat{b}_i \qquad [23]$$

We parameterize $h(\vec{x})$ as a simple quadratic function of elevation:

$$h(\vec{x}_i) = \beta\left(\text{Elevation}(\vec{x}_i)\right) + \gamma\left(\text{Elevation}(\vec{x}_i)\right)^2 \qquad [24]$$

Where $\beta$ and $\gamma$ are parameters to be determined. Similarly $\lambda(\vec{x}_i)$ is parameterized as cubic spline function of the $cos\left(\text{Latitude}(\vec{x}_i)\right)$ with 16 knots. These 16 free parameters, as well as the free parameters related to elevation are determined empirically as part of a Kriging process used in the construction of $G(\vec{x}_i)$. The Kriging formulation for $G(\vec{x}_i)$ is the same as that developed for $W(\vec{x}_i, t_j)$ above except that

$$\begin{pmatrix} B_1(\vec{x}) \\ \vdots \\ B_N(\vec{x}) \end{pmatrix} = \begin{pmatrix} \frac{1+(n_1-1)R(0)}{n_1} & R(\vec{x}_1,\vec{x}_2) & \cdots & R(\vec{x}_1,\vec{x}_N) \\ R(\vec{x}_2,\vec{x}_1) & \frac{1+(n_2-1)R(0)}{n_2} & & R(\vec{x}_2,\vec{x}_N) \\ \vdots & & \ddots & \vdots \\ R(\vec{x}_N,\vec{x}_1) & R(\vec{x}_N,\vec{x}_2) & \cdots & \frac{1+(n_N-1)R(0)}{n_N} \end{pmatrix}^{-1} \begin{pmatrix} R(\vec{x},\vec{x}_1) \\ \vdots \\ R(\vec{x},\vec{x}_N) \end{pmatrix} \qquad [25]$$

$$\hat{G}(\vec{x}) = \sum_{i=1}^{N} B_i(\vec{x})\left(\hat{b}_i - \hat{\lambda}(\vec{x}) - \hat{h}(\vec{x})\right) \qquad [26]$$

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 gives greater weight to values of $\hat{b}_i$ that are more precisely constrained by longer records. The free parameters for $\hat{\lambda}(\vec{x})$ and $\hat{h}(\vec{x})$ are then constrained minimizing:

$$\int \hat{G}(\vec{x})^2\, d\vec{x} \qquad [27]$$

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. The characteristics of the fit for $C(\vec{x}_i)$ using GHCN data is shown in Figure 3.

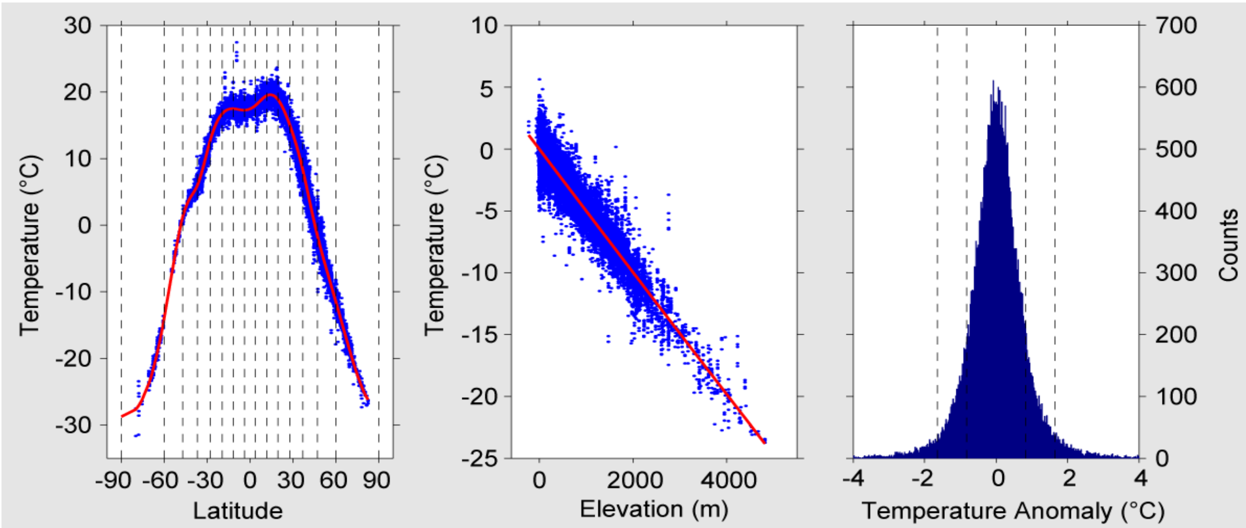


**Figure 3: Shows the characteristics of the climatology fit using GHCN data. (Left) Latitude dependent component of the fit, $\hat{\lambda}(\vec{x})$, shown as a red curve and compared to the data expressed as $\hat{b}_i - \hat{h}(\vec{x}_i) - \hat{G}(\vec{x}_i)$. The position of spline knots are indicated with vertical lines. The slight bump circa 15 N is associated with the elevated temperatures across the Sahara. (Center) Elevation dependent component of the fit $h(\vec{x})$ expressed as a red curve and compared to the data expressed as $\hat{b}_i - \hat{\lambda}(\vec{x}_i) - \hat{G}(\vec{x}_i)$. (Right) The misfit residuals $C(\vec{x}_i) - \hat{b}_i$, with one and two sigma standard deviations indicated as vertical lines. The far outliers on the residual plot are most likely to be caused by stations whose latitude or elevation was significantly misreported.**

This final normalization described here has the effect of placing the mean temperature, $\hat{\theta}(t_j)$, on an absolute scale such that these values are a true measure of mean temperature and not merely a measure of a temperature anomaly. However, the uncertainty associated with this normalization is often larger than the uncertainty associated with other parts of the temperature time series estimate. This occurs due to the large range of variations in $\hat{b}_i$ 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. However, one can ignore the uncertainty in the climatology, when considering temperature difference since

$$T(\vec{x}, t_2) - T(\vec{x}, t_1) = \theta(t_2) + W(\vec{x}, t_2) - \theta(t_1) - W(\vec{x}, t_1) \quad [28]$$

is independent of $C(\vec{x})$.

## Outlier Weighting

In this section we discuss the mathematics of our handling of 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, $\Delta_i(t_j)$, between a temperature station's reported data and the expected value at that same site:

$$\Delta_i(t_j) = d_i(t_j) - \hat{b}_i - \hat{\theta}(t_j) - \widehat{W}^{\dagger}(\vec{x}_i, t_j) \tag{29}$$

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:

$$\widehat{W}^{\dagger}(\vec{x}_i, t_j) = \widehat{W}(\vec{x}_i, t_j) - S_i(\vec{x}_i, t_j)(d_i(t_j) - \hat{b}_i - \hat{\theta}(t_j)) \tag{30}$$

The scale of the typical measurement error ($e \approx 0.62$ C) is estimated from:

$$e^2 = \frac{\sum_{i,j}\left(\Delta_i(t_j)\right)^2}{\sum_{i,j} 1} \tag{31}$$

The outlier weight adjustment, originally mentioned in equation [7], is then defined as

$$\omega_{i,j} = \begin{cases} 1 & \text{if } \left(\Delta_i(t_j)\right)^2 \leq (2.5e)^2 \\ 2.5e/\left|\Delta_i(t_j)\right| & \text{otherwise} \end{cases} \tag{32}$$

Equation [32] thus specifies a downweighting term to be applied for point outliers that are more than $2.5e$ from the expected value based on the interpolated field.

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 [32], 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.

## 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, the average misfit for the entire time series can be expressed as:

$$\varepsilon_i^2 = \frac{\sum_j \min\{\left(\Delta_i(t_j)\right)^2, 25e^2\}}{\sum_j 1} \quad [33]$$

We introduce the "min" function to avoid giving too great a weight to the most extreme outliers when judging the average reliability of the series. A metric of relative reliability is then defined as:

$$\varphi_i = \frac{2e^2}{e^2 + \varepsilon_i^2} \quad [34]$$

Due to the limits on outliers imposed in equation [33], this metric has a range between 1/13 and 2, 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, but 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, 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 the spatial adjustment coefficients, originally mentioned in equation [13], to be:

$$\xi_{i,j} = \frac{\varphi_i \, \omega_{i,j}}{\left(\sum_m \varphi_m S_m(\vec{x}, t_j)\right) + \left(1 - F(\vec{x}, t_j)\right)} \quad [35]$$

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

$$\hat{\theta}(t_j) + \widehat{W}(\vec{x}, t_j) = \left(1 - F(\vec{x}, t_j)\right) \hat{\theta}(t_j) + \sum_i S_i(\vec{x}, t_j)\left(d_i(t_j) - \hat{b}_i\right) \quad [36]$$

combined with 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 $\xi_{i,j} S_i(\vec{x}, t_j)$ is equal to $S_i(\vec{x}, t_j)$ if all the reliability factors, $\varphi_i$, and outlier weights, $\omega_{i,j}$, 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 [35] would lead to a weighting scheme that is numerically unstable.

It is important to note that equation [35] merely says that the estimate of the local weather $\widehat{W}(\vec{x}, t_j)$ should give proportionally greater weight to more reliable records. However, if all of the records in a given region have a similar value of the reliability factor $\varphi_i$, then they will all receive a similar weight, $\xi_{i,j}$, regardless of the actual numerical value of $\varphi_i$. This behavior is important as some regions of the Earth, such as Siberia, tend to have broadly lower values of $\varphi_i$ due to the high variability of local weather conditions. However, as long as all of the records in a region have similar values for $\varphi_i$, 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 $\hat{\theta}(t_j)$.

As noted above, the formulation of equation [35] 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 $\xi_{i,j} S_i(\vec{x}, t_j)$.

The determination of the weighting factors $\omega_{i,j}$ and $\xi_{i,j}$ is accomplished via an iterative process that seeks convergence. Similarly, it is computationally efficient to use the value of $\widehat{W}(\vec{x}_i, t_j)$ from the prior iteration when computing $\hat{b}_i$ via equation [7]. As described in the main text, 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 $\hat{\theta}(t_j)$ between consecutive iterations.

Implicit in the discussion of station reliability considerations are several assumptions. Firstly, the local weather field constructed from many station records, $\widehat{W}(\vec{x}, t_j)$, 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, this approach cannot 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 in the main text 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

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 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 papers submitted by our group (Wickham et al. submitted;

Muller et al. submitted) find that objective measures of station quality and urbanization have little or no 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.

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.

## Statistical Uncertainty – Jackknife Method

The "jackknife" method developed by Quenoille and Tukey (Tukey 1958, Quenoille 1949, Miller 1974) is a superior sampling method that can largely avoids problems associated with spatial biasing when subsampling, and is the primary technique we apply for calculating statistical uncertainty. This method is traditionally used when the number of data points is too small to give a good result using ordinary sampling. Given the fact that temperature reconstructions use thousands of stations, each often having hundreds of data points, it may seem surprising that this method would prove important. However, despite the large set of data, there are always times and regions that are sparsely sampled.

The jackknife method is used in the following way. Given a set of stations, eight overlapping subpopulation are constructed each consisting of 7/8th of the data, with a different and independent 1/8th removed from each group. The data from each of these subsamples 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

$$\hat{\theta}_k^\dagger(t_i) = 8\,\hat{\theta}_k(t_j) - 7\,\hat{\theta}(t_j) \quad [37]$$

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:

$$\sigma_{\text{jackknife}}(t_j) = \frac{\sqrt{\sum_k \left(\hat{\theta}_k^\dagger(t_j) - \langle \hat{\theta}_k^\dagger(t_j) \rangle\right)^2}}{\sum_k 1} \quad [38]$$

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. In addition, the number of samples can be easily increased without worrying that the network would become too sparse.

A brief comment should be made here. In computing the subsampled temperature series, $\hat{\theta}_k(t_j)$, the outlier and reliability adjustment factors $\omega_{i,j}$ and $\xi_{i,j}$ are recomputed for each sample. This means the process of generating $\hat{\theta}_k(t_j)$ is not entirely linear, and consequently the jackknife estimate in equation [39] is not analytically guaranteed to be effective. However, in the present construction the deviations from linearity are expected to be small since most adjustment factor will be 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 $\omega_{i,j}$ and $\xi_{i,j}$ 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.

## Spatial Uncertainty

Spatial uncertainty measures the amount of error that is likely to occur due to incomplete sampling of land surface areas. The primary technique applied in this case is empirical. The sampled area available at past times is superimposed over recent time periods, and one is able to calculate the error that would be incurred in measuring the modern temperature field given only that limited sample area. 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 $1960 \leq t_j \leq 2010$ 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:

$$\tau(t_j, t_m) = \frac{\int F(\vec{x}, t_j)\widehat{W}(\vec{x}, t_m)d\vec{x}}{\int F(\vec{x}, t_j)d\vec{x}} \quad [39]$$

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

$$\sigma_{\text{spatial}}(t_j) = \sqrt{\frac{\sum_{t_m=1960}^{2010}\left(\tau(t_j, t_m) - \tau(t_m, t_m)\right)^2}{\sum_{t_m=1960}^{2010} 1}} \quad [40]$$

Ideally $F(\vec{x}, t_j)$ would be identically 1 during the target interval $1960 \leq t_j \leq 2010$ 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:

$$\sigma_{\text{spatial}}(t_j) \approx \sqrt{\int\int\left(1 - \frac{F(\vec{x}, t_j)}{\bar{F}(t_j)}\right)\left(1 - \frac{F(\vec{y}, t_j)}{\bar{F}(t_j)}\right)\hat{V}(\vec{y})\hat{V}(\vec{x})R(\vec{x}, \vec{y})d\vec{x}d\vec{y}} \quad [41]$$

Where $R(\vec{x}, \vec{y})$ is the correlation function estimated in equation [14], $\bar{F}(t_j)$ 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:

$$H(\vec{x}, t_j) = \begin{cases} F(\vec{x}, t_j) & \text{if } F(\vec{x}, t_j) \geq 0.4 \\ 0 & \text{otherwise} \end{cases} \quad [42]$$

$$\hat{V}(\vec{x}) = \sqrt{\frac{\sum_j H(\vec{x}, t_j) \left( \frac{\widehat{W}(\vec{x}, t_j)}{F(\vec{x}, t_j)} \right)^2}{\sum_j H(\vec{x}, t_j)}} \quad [43]$$

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 [42] tends to be 30-50% smaller than the result of equation [41]. This is probably because the linearized error propagation formula in equation [42] and the approximate correlation function estimated in equation [14] do not capture enough of the structure of the field for this application, and hence the formulation of uncertainty in equation [41] is likely to be superior. At late times both estimates of the uncertainty due to spatial incompleteness tend be far lower than the statistical uncertainty. 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 [41] is partially limited due to incomplete sampling during the target interval. To compensate for this we add a small analytical correction, determined via equation [42] in the computation of our final spatial uncertainty estimates at regions with incomplete sampling. This correction is essentially negligible except at late times.

## 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. Huang, Chunfeng, Haimeng Zhang, and Scott M. Robeson (2011), “On the Validity of Commonly Used Covariance and Variogram Functions on the Sphere,” *Math Geoscience*, DOI 10.1007/s11004-011-9344-7.

14. Isaaks EH, Srivastava RM (1989) Applied geostatistics. Oxford University Press, New York

15. 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.

16. 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.

17. Jones, P.D., T.M.L. Wigley, and P.B. Wright. 1986. Global temperature variations between 1861 and 1984. *Nature*322:430-434.

18. Journel, A. G. *Fundamentals of geostatistics in five lessons*. American Geophysical Union, 1989; 40 pages.

19. Kitandis PK (1997) *Introduction to geostatistics*. University of Cambridge Press, Cambridge

20. 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.

21. 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.

22. 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.

23. 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

24. 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

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

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

27. 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.

28. 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.

29. 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.

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

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

32. 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.

33. 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.

34. Quenoille, M. H. (1949), “Approximate tests of correlation in time-series,” *Journal of the Royal Statistical Society* B 11, p. 68-84.
35. 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.
36. Smith, T. M., et al. (2008), Improvements to NOAA's Historical Merged Land-Ocean Surface Temperature Analysis (1880-2006), J. Climate, 21, 2283-2293.
37. Stothers, Richard B. (1984). "The Great Tambora Eruption in 1815 and Its Aftermath". *Science* 224 (4654): 1191–1198.
38. 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.
39. Tsay, R, S.. (1991) Detecting and Modeling Non-linearity in Univariate Time Series Analysis. Statistica Sinica 1:2,431-451.
40. Tukey, J.W. (1958), “Bias and confidence in not quite large samples”, *The Annals of Mathematical Statistics*, 29, 614.
41. 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
42. 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.

43. 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.
44. 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.

**APPENDIX**

Symbols used in the Berkeley Average mathematical supplement.

| | |
|---|---|
| $t$ | 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 $t_j$ |
| $\omega_{i,j}$ | the reliability adjusted weight associated with data point $d_i(t_j)$ |
| $\varphi_i$ | the relative reliability of the *i*-th station |
| $\xi_{i,j}$ | the reliability adjusted 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 $\vec{x}_b$ |
| $\mathrm{Cov}(\vec{x}_a, \vec{x}_b)$ | the covariance in temperature between locations $\vec{x}_a$ and $\vec{x}_b$ |
| $\sigma_i^2$ | the variance of the temperature record at the *i*-th station |
| $\Delta_i(t_j)$ | 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.