Comparison of Berkeley Earth, NASA GISS, and Hadley CRU averaging techniques on ideal synthetic data

Robert Rohde
Lead Scientist, Berkeley Earth Surface Temperature

1/15/2013

Abstract

This document will examine the accuracy of the land-surface temperature mapping and averaging techniques employed by Berkeley Earth, NASA GISS, and the Hadley Centre / Climatic Research Unit groups. The ability of algorithms to estimate the global properties of weather fields from sparse data is a fundamental limitation on the achievable accuracy of climate reconstructions. Here, these algorithms are tested by creating simulated weather station data from the temperature field of a global climate model (GCM) and then measuring the effectiveness of each method at reproducing the properties of the underlying GCM field. We examine both the ability to estimate the global land average and the typical local error in the mapped field. In nearly all cases, the Berkeley Earth averaging methodology is shown to have greater accuracy at reproducing both the global and local details of the temperature field.

Introduction

There are four major efforts to synthesize the Earth's disparate temperature observations into a coherent picture of our planet's climate history. These efforts are led respectively by NOAA's National Climate Data Center (NOAA NCDC), NASA's Goddard Institute for Space Studies (NASA GISS), a collaboration between the University of East Anglia's Climatic Research Unit and the UK Met Office's Hadley Centre (CRU[1]), and the Berkeley Earth Surface Temperature group. Each group uses different averaging techniques, quality control procedures, homogenization techniques, and datasets.

The current discussion will focus exclusively on the effect of the different averaging and interpolation techniques employed. To do this we will test the accuracy of each averaging technique using synthetic "error-free" data derived from the temperature field of a global climate model. This approach allows us to create simulated station data that is free from the various types of errors and biases that will occur in the real world. By limiting the present discussion to "error-free" data, we can examine the efficacy of the different averaging techniques separate from the consideration of quality control and

[1] Here we choose to use the acronym CRU to emphasize that we are looking at the land-based data product officially known as CRUTEM4. Originally the CRUTEM data products were created by the Climatic Research Unit alone, but for more than a decade these data products have been published in collaboration with the Hadley Centre. A combined data set of land and ocean temperatures, known as HADCRUT, is also produced by this collaboration. The acronym "HadCRU" is frequently used to describe this collaboration, especially in the context of the combined land and ocean data product. However, since the present discussion considers only the land component, we will use the acronym CRU to emphasize that the analysis is referring to the CRUTEM4 methods.

homogenization issues. The actual comparison is made by using our implementations of the GISS and CRU averaging techniques as understood from their papers (Hansen et al. 2010, Jones et al. 2012). We do not currently have a working implementation of the NOAA NCDC method, and hence it will not be included in the present comparison. With their assistance, we may add an analysis of the NOAA method at a future time.

Simulated Temperature Records

In the present analysis the CCSM4 climate model was used.[2] It was chosen because it had the highest spatial resolution for surface air temperature fields (192 latitude steps by 288 longitude steps) among the models that had archived data for the "Past1000" experiment in the Climate Model Intercomparison Project (CMIP5) as of the time the present analysis was started. For the Past1000 experiment, CCSM4 simulated the years 850 to 2006 allowing us to draw on more than 1000 years of climate model output from which simulated station data could be generated. We note that higher resolution climate models do exist, but they are rarely run for the hundreds of years necessary to simulate the three hundred year history of weather observation.

We began the process of creating synthetic data by using the Global Historical Climatology Network monthly dataset (GHCN-M) as a template[3]. This well-established dataset is the foundation of the climate analysis conducted by NOAA and GISS and so is a logical starting point for the present work. The dataset includes 7280 weather stations and will provide a set of times and locations at which the climate model field can be sampled in order to produce synthetic data with a realistic spatial and temporal structure.

Though the Berkeley Average framework can make use of all 7280 time series in the GHCN-M collection, the NASA GISS and CRU methods each have different completeness and normalization requirements for their data, and as a result, not all data is usable by each method.[4] In the GHCN-M network we find 5719 stations where sufficient data exists to compute the baseline climatology required by CRU, and 6278 stations which meet GISS baseline requirements. Collectively, there are 5457 time series that are acceptable to all three methods. In order to isolate the effects of averaging performance, and avoid biases created by the different selection procedures, we have limited our study to these 5457 records that are acceptable to all three methodologies.[5]

---

[2] CCSM4: http://www.cesm.ucar.edu/models/ccsm4.0/

[3] GHCN-M version 3: http://www.ncdc.noaa.gov/ghcnm/v3.php

[4] There is actually some ambiguity in the published account of baseline selection procedures. For example, in Jones et al. 2012, the CRUTEM4 baseline procedure is described in part as "[m]onthly averages for 1961–1990 were calculated from the enhanced station data set, accepting an average if at least 14 years of data are available." However, it is unclear whether "14 years" means (a) any 168 months during the 1961-1990 interval, (b) at least 14 complete calendar years during this interval, (c) at least 14 calendar years each consisting of at least 9 months (i.e. the definition of "complete year" used elsewhere in the same paper), or some other condition. We believe such nuances are ultimately of no consequence.

[5] In their ordinary practice, CRU supplements their baseline estimates with information from World Meteorological Organization (WMO) climatologies and other sources. As such, they would be likely to use more than the 5457

The locations of these 5457 stations are shown in Figure 1. The accompanying Figure 2 shows the number of stations active over time. Like the GHCN-M network as a whole, the peak of station activity is in the 1970s followed by a significant decline in station counts as one approaches the present.

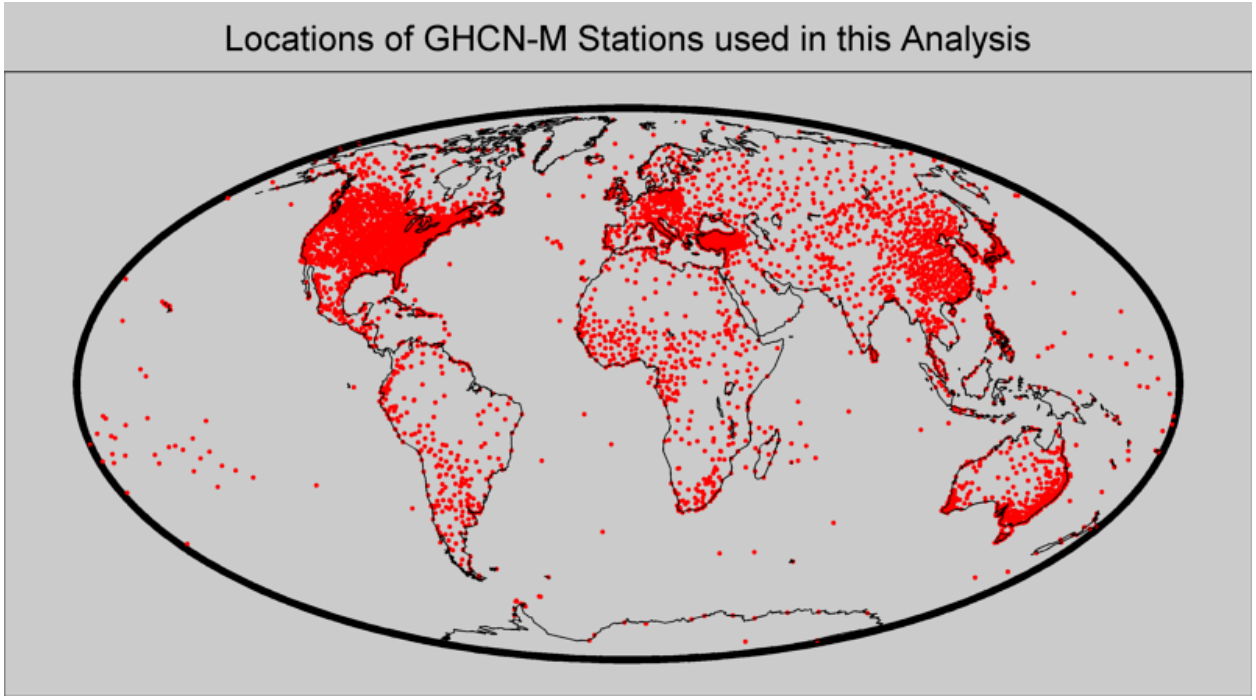

Figure 1: Locations of all GHCN-M stations (5457) that have sufficiently complete records to meet the baseline requirements of all three methodologies considered by this study. The sampling history of these weather stations is used to determine the times and locations at which the synthetic data in this study is constructed.

GHCN-M records with complete reference intervals; however, for simplicity we have chosen to avoid considering such secondary sources, and limit our study to time series with sufficiently complete reference intervals that the baseline may be determined directly.

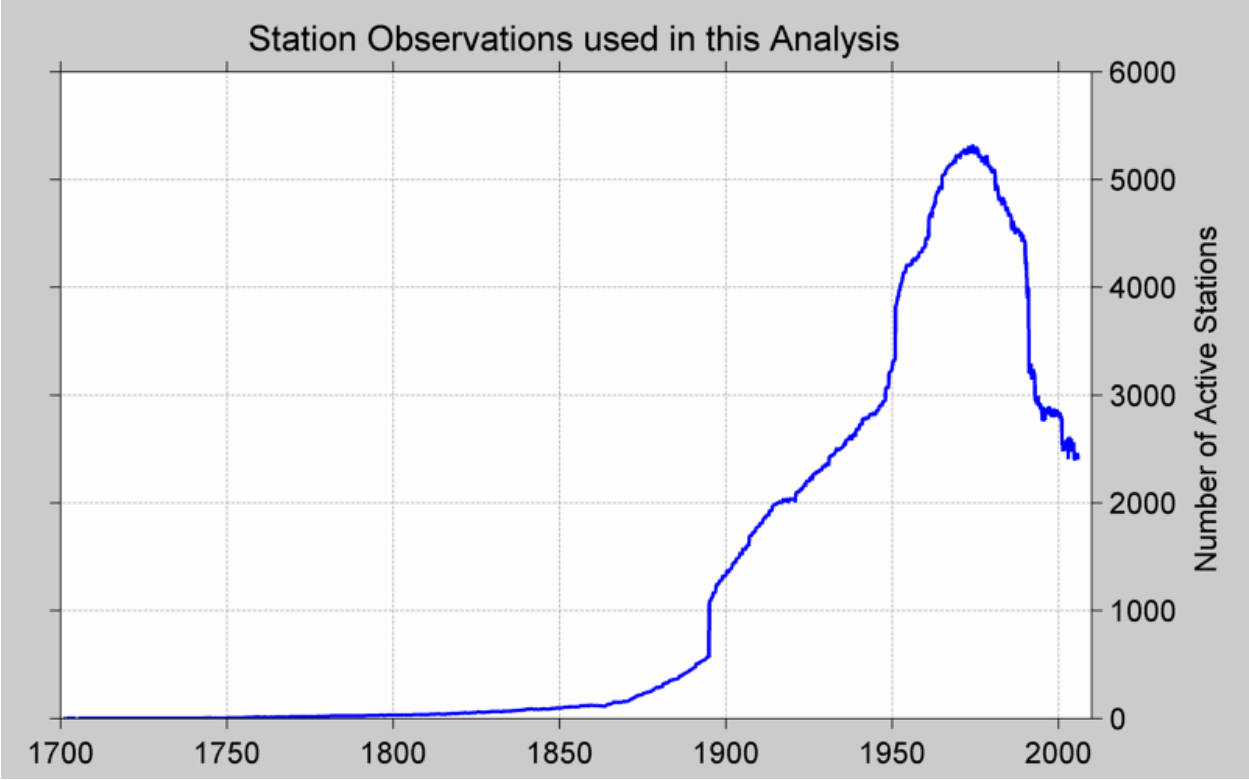

Figure 2: The number of observations vs. time for the 5457 GHCN-M stations considered in this study.

Using the locations and observation histories of these 5457 GHCN-M stations, it was possible to sample from the CCSM4 climate model's air surface temperature field in order to create simulated station histories. The earliest observation in our dataset occurred in 1701, meaning approximately 300 years of model output are needed to generate a complete set of simulated observations. Rather than perform this sampling only once, we created 50 sets of simulated data by using randomly chosen time offsets within the 1150 year history of the CCSM4 model run. For example, if an offset of 300 years were chosen, then model year 1600 would be treated as if it were the observation year 1900. This allows us to create 50 different sets of simulated weather from which to attempt reconstruction. Obviously since the model run is only 1150 years long, these different simulations often overlap to an extent; however, we do not believe that the overlaps will impact our conclusions. Only whole numbers of years were chosen as offsets in order to preserve any seasonal behavior.

In order to draw conclusions about the effectiveness of the different averaging techniques, we will apply Berkeley Earth, CRU, and GISS style averaging to each of these simulated data sets and compare the resulting global averages and mapped fields to the properties of the original GCM data set. As the simulated data is intrinsically free from any noise or bias, we have omitted any parts of the respective algorithms associated with quality control or homogenization. In addition, the averaging techniques generally require data that has had its seasonality removed, generally by subtracting an average seasonal cycle from each time series. To further reduce differences, we used the true seasonality in the GCM field as the basis for removing seasonality from each simulated time series so that slight differences in the handling of seasonality in the three algorithms would not affect our conclusions.

Because CRU has the poorest spatial resolution of the methods considered (5°x5° latitude by longitude), analysis of the mapped fields will be made after downscaling the other results to this same resolution.

Land Field Completeness

To begin, we will look at the completeness of the generated fields. Specifically, we examine what fraction of the Earth's land surface is estimated by each method as a function of time. This is shown in Figure 3. At essentially all times, Berkeley Earth estimates temperature anomalies over a larger fraction of Earth's land area than the NASA GISS method, which in turn estimates a larger fraction of land area than the CRU method. Since each method is utilizing identical data, this plot shows how willing each algorithm is to extrapolate the available observations to regions with no nearby observations. It is not surprising that CRU is the most limiting since it uses data only in the same 5°x5° grid box where the data is reported. The other two techniques each allow data to be extrapolated over more than 1000 km (equivalent to more than 10° at the equator). Of course, the fact that such extrapolations are made says nothing about their accuracy, which will be addressed in further sections.

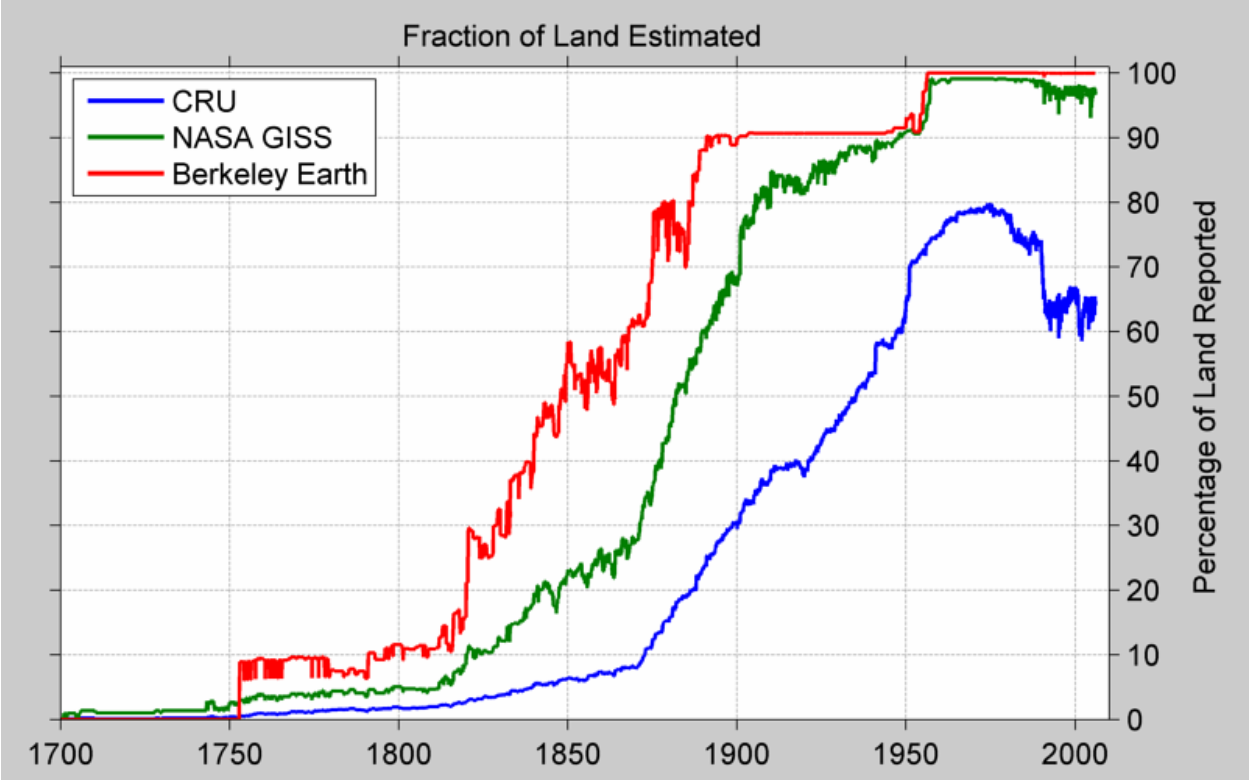

Figure 3: Comparison of the fraction of Earth's land surface over which each method reports a value. Since each method is using identical data, this figure shows the willingness of each algorithm to extrapolate. At all times, CRU has a significant number of empty grid cells, while GISS and Berkeley Earth fill out nearly the entire globe during the 20th century. The step-change in coverage circa 1955 corresponds to the initiation of weather monitoring efforts in Antarctica.

Accuracy of Global Land Averages

We will begin our examination of the efficacy of the different methods by considering how well they each reproduce the global land average. In Figure 4 we show the typical error in reproducing the 12-month moving average of global land surface temperatures. This is found by comparing the global land average in each of the 50 simulated data sets to the corresponding true land average of the GCM field and taking the standard deviation of the respective differences across all 50 simulations. In these plots, we show the two standard deviation level, corresponding roughly to a 95% confidence level. We have chosen to focus on the time interval 1850 to present because CRU and GISS generally do not consider reconstructions before this time to be reliable.[6]

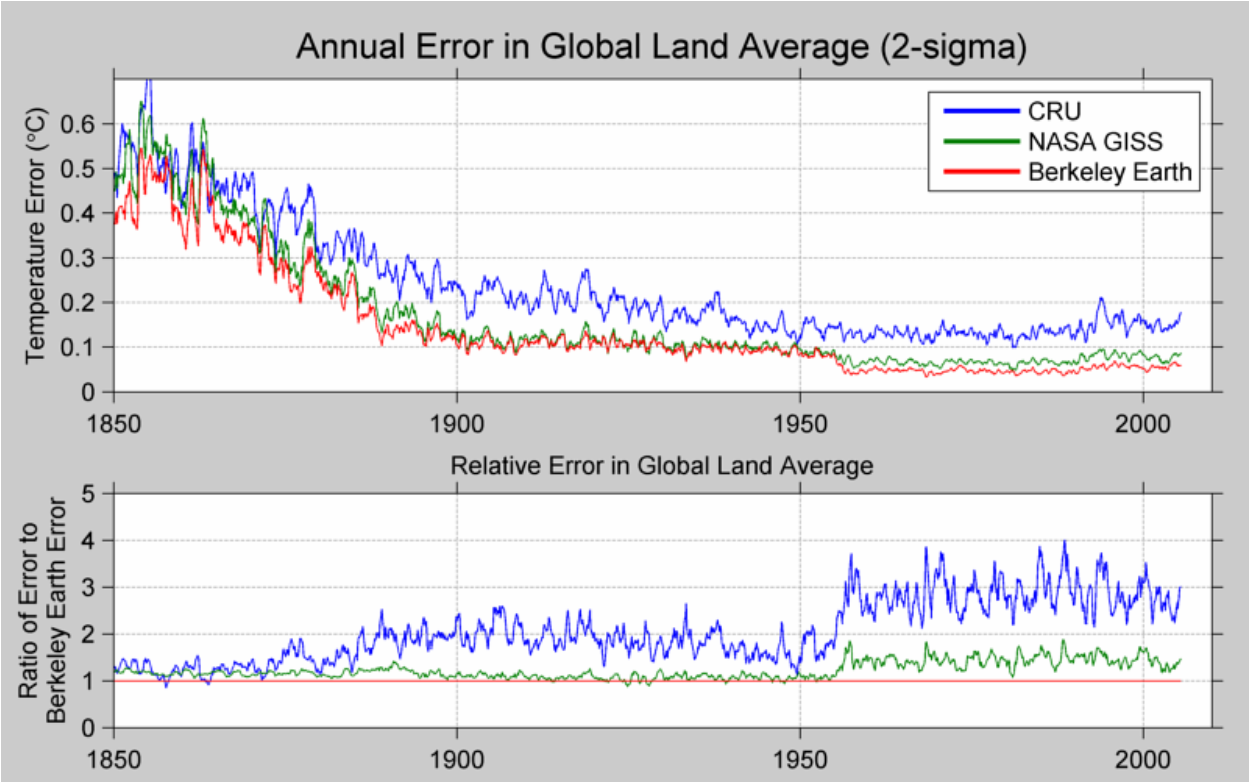

Figure 4: (Top) Typical error in the global land average reconstructions produced from our 50 simulated data sets. This is computed for the 12-month moving average, and expressed as the two standard deviation level for the calculated errors. (Bottom) The relative error associated with the various reconstruction methods expressed as a ratio to the Berkeley Earth error.

Here we find that Berkeley Earth is at least as accurate as the other methods, and often significantly superior. From 1850 to 1900, Berkeley has slightly greater accuracy than GISS, and a somewhat greater advantage over CRU. From 1900 to about 1955, GISS and Berkeley have a similar level of performance. During this period the largest factor limiting accuracy for both methods appears to be the absence of

[6] CRU begins their reconstruction in 1850. NASA GISS begins in 1880. In both cases, the averaging methodologies have no problem extending further backward in time. In the case of our "error-free" data simulations it is easy to consider doing early period reconstructions; however, in the real world issues of data quality become particularly important for the sparse early data. In the very early period, issues of uncertainty and homogeneity analysis must be considered which are beyond the scope of the present discussion.

any stations in Antarctica. After weather stations are introduced to Antarctica in the 1950s, the accuracy of the Berkeley methods typically improves upon GISS by 20 to 50%. During the 20th century, the error associated with the CRU method is substantially larger than the other two methods, often producing errors 50% to 250% greater than the Berkeley methodology. This is consistent with our findings in other papers that the Berkeley Earth uncertainties are significantly lower than CRU uncertainties. We believe the large errors associated with the CRU method are mainly caused by the incompleteness of their reconstructed temperature field (i.e. Figure 3).

In Figure 5 we show a similar estimate of the error associated with measuring decadal averages of the land-surface temperature. As in Figure 4, Berkeley Earth is often slightly more accurate than GISS in determining the decadal average temperatures and several times more accurate than CRU.

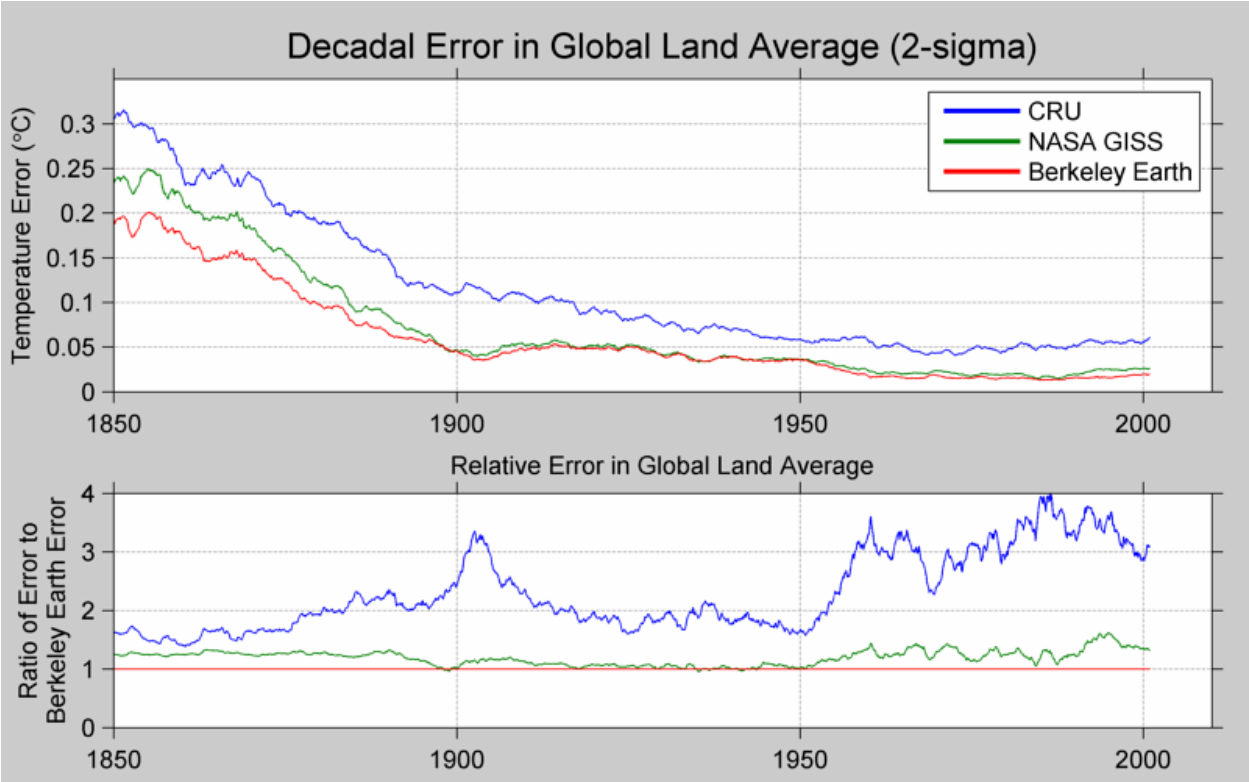

Figure 5: Same as Figure 4, but for 10-year moving averages of the land surface data rather than the 12-month moving averages presented in Figure 4.

Figure 6 shows similar results to Figures 4 and 5 expressed at a monthly level. Here we observe that there is a large degree of seasonal variability. Even though the average seasonal cycle has been removed, there is still higher spatial variability during Northern Hemisphere winter than in any other season. This limits the ability to accurately reconstruct the field, and hence larger reconstruction errors occur during Northern Hemisphere winter than any other season. To our knowledge the uncertainty estimates published by groups other than Berkeley Earth generally don't explicitly declare a seasonal variation in their ability to perform these reconstructions; nonetheless, such seasonal noise is present to varying degrees in all three reconstructions studied. Seasonal variations in error are included in the uncertainties reported by Berkeley Earth. Beyond that, the pattern observed is similar to the pattern

observed for the annual and decadal averages. In general, the performance of Berkeley Earth slightly surpasses GISS and substantially surpasses CRU.

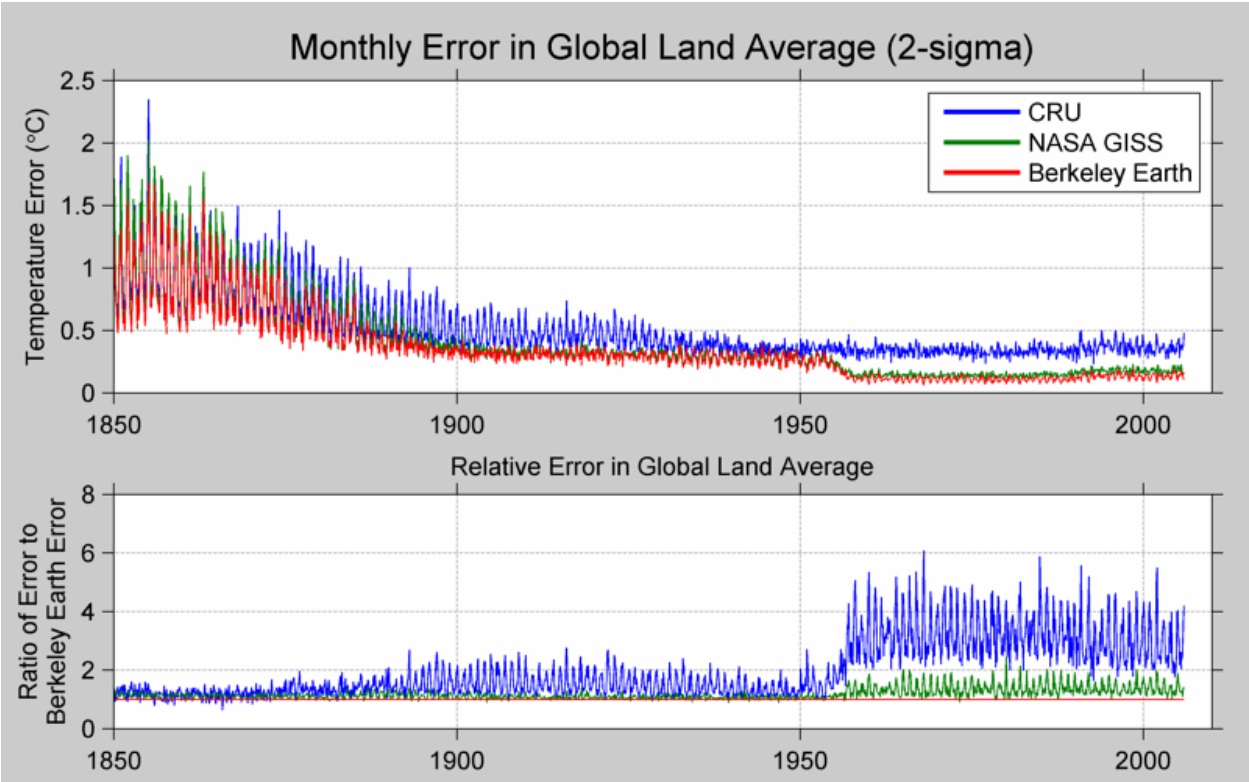

Figure 6: Same as Figures 4 and 5, though showing monthly values rather than moving averages. Here we observe that the error has a substantial seasonal component. This is associated with seasonal differences in spatial variability.

Lastly, we want to make a special note about the CRU land average. Unlike other groups, CRU currently creates their land average by first averaging the hemispheres separately and then estimating the global land average as 2/3 times the Northern Hemisphere average plus 1/3 times the Southern Hemisphere average.[7] This is in contrast to directly weighting each populated grid cell by its land area when averaging. Using the synthetic data we can evaluate whether pre-averaging the hemispheres is beneficial. In Figure 7, we compare CRU's hemisphere weighing to a more direct grid cell weighted average using the error in the 12-month moving average as a measure of performance. The result is somewhat ambiguous. From roughly 1850 to 1900 the CRU hemispheric weighting does reduce the overall uncertainty. From around 1900 to 1950, the two approaches are comparable. However, from roughly 1950 to the present day, we find that CRU's preferred technique is actually worse than weighting and averaging the occupied grid cells directly.

[7] The Northern Hemisphere contains 67.2% of Earth's land, which is approximately the 2/3 weight that CRU uses.

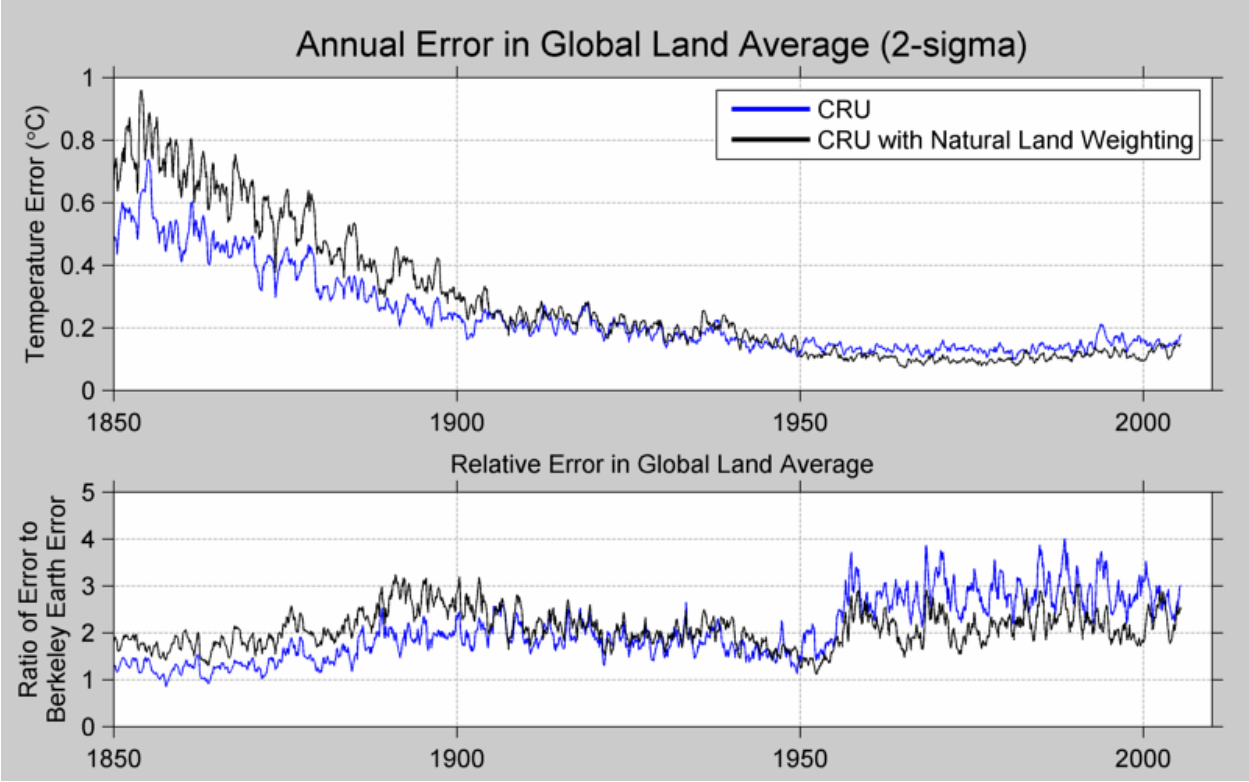

Figure 7: Comparison of the current CRU land average technique using hemisphere weighting to an average where each grid cell is individually weighted by the land area it contains. We find that CRU's current method is superior during the 19th century, but generally increases the error during the latter half of the 20th century.

Mapping Accuracy

Having examined the ability of these methods to reconstruct the global land average, we now look at their accuracy in reproducing the variations in the local field. Because the different averaging methods populate different number of grid cells from the available data, we will do this analysis in two parts. First, we will look at the average error in grid cells where all three methods report a value. Subsequently, we will look at the error in measuring grid cells reported by both Berkeley Earth and GISS.

In Figure 8, we show the average error in reconstructing the 12-month moving average at the typical 5°x5° grid cell populated by all three groups. Here we find that Berkeley provides the most accurate estimate of each grid cell. However, unlike the previous section, we find that the CRU method provides the second best average with GISS lagging significantly behind. As discussed in a separate document, the GISS method uses distance weighted averaging out to 1200 km, and this has the general effect of blurring out fine details. We believe it is likely that this blurring limits the ability of the GISS technique to accurately capture local details. If the local error in GISS's reconstructions are being generated primarily by this blurring effect it also helps to explain why the addition of more data throughout the 20th century did little to improve the local accuracy of GISS's reconstructions. Ultimately though, we note that Berkeley Earth improves on both of these techniques in measuring the local structure. During

the 20th century, local errors associated with the CRU method appear to be about 50% larger than Berkeley Earth, while GISS errors are 100 to 150% larger.

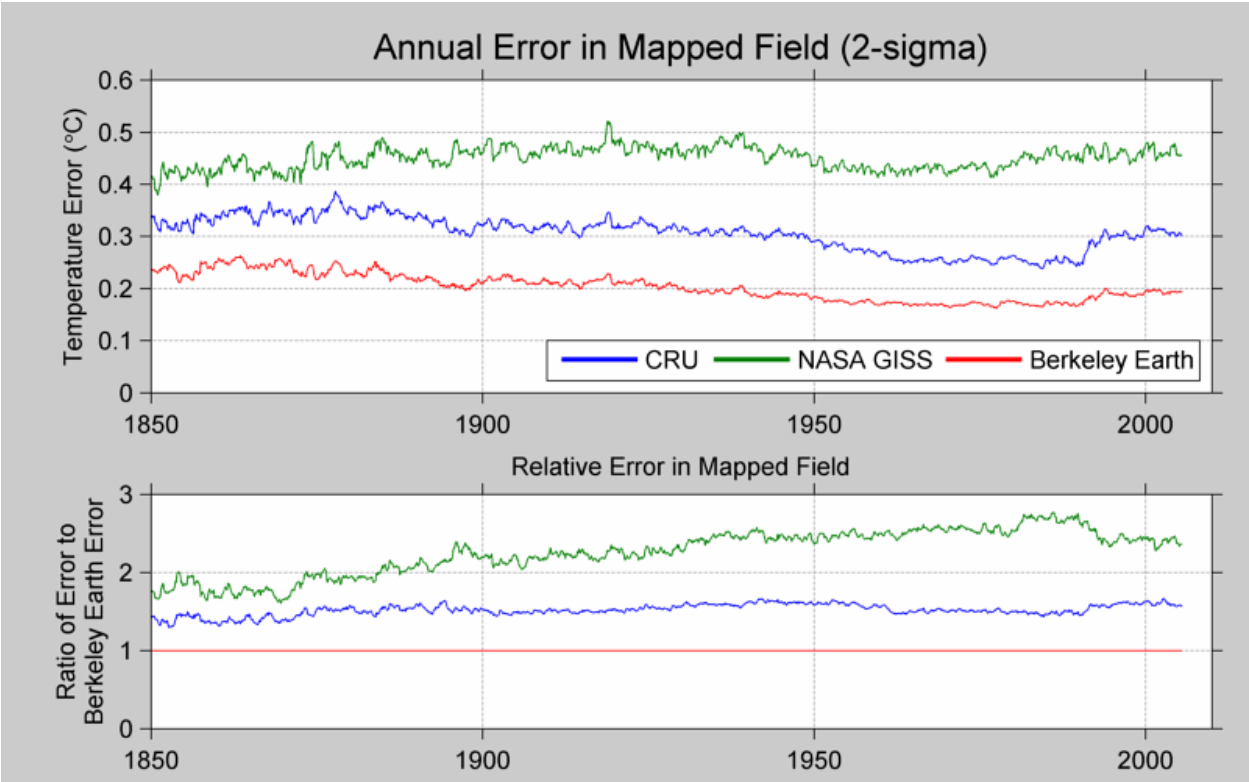

Figure 8: Average error in the 12-month moving average estimated at the typical grid cell, stated at the 2 standard deviation level. The error is computed on a 5°x5° grid using at each time only those grid cells where all three methodologies are able to report a value.

In Figure 9, we provide the same information as in Figure 8, but calculating the typical error in estimates of the 10-year moving average at each populated grid cell. As before, Berkeley Earth is usually superior to CRU, which is usually superior to GISS. However, at this longer time scales we can also observe artifacts associated with the GISS and CRU baseline processes. As a baseline, GISS defines the period 1951-1980 to have zero mean at every location. This approach to defining a baseline period, though often used, has the side effect that it introduces slight distortions into the reconstructed field. Specifically it artificially suppresses spatial variation in the field during the baseline period while also slightly increasing apparent variation outside the baseline interval. This makes it easier for GISS to reproduce averages within the baseline interval, while also creating corresponding slight increases in error outside the baseline period. CRU shows a corresponding effect for their baseline interval, 1961 to 1990. These side effects of using a fixed baseline period, though often of no real consequence, can become important when examining certain statistical properties of the field such as long-term averages, or the frequency of extreme events. Since Berkeley Earth does not use this type of a normalization process, there is no specific interval that is favored in the Berkeley reconstruction.

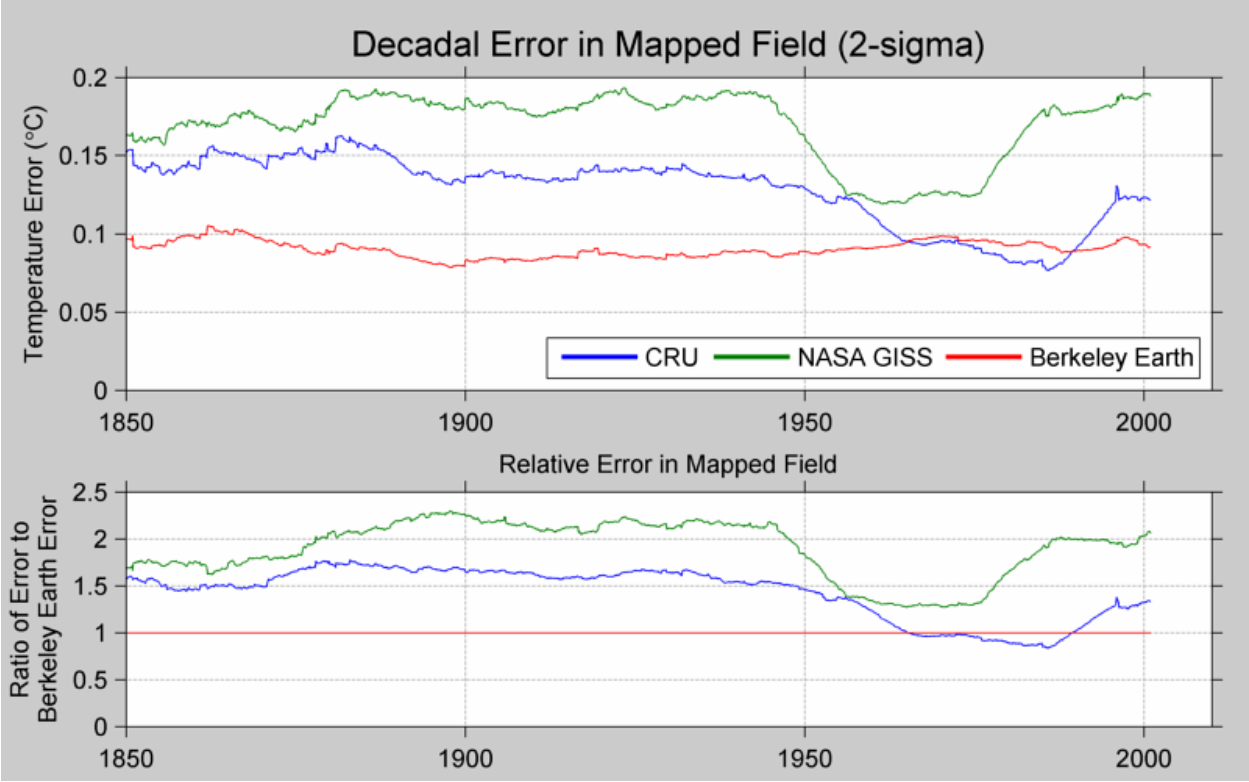

Figure 9: Same as Figure 8 but based upon the 10-year moving average at each location instead of the 12-month moving average.

In Figure 10, we show the same information as Figures 8 and 9, except presented at monthly resolution. As previously shown in Figure 6, there is a substantially seasonal component affecting the accuracy of all three reconstructions. The higher spatial variability associated with Northern Hemisphere winter leads to considerable greater errors in the reconstruction of the local field during that season. For GISS in particular, we find a more than 50% increase in local error during Northern Hemisphere winter compared with Northern Hemisphere summer. Berkeley Earth appears to show the mildest seasonality component, but still has at least a 20% increase in error in northern winter relative to northern summer.

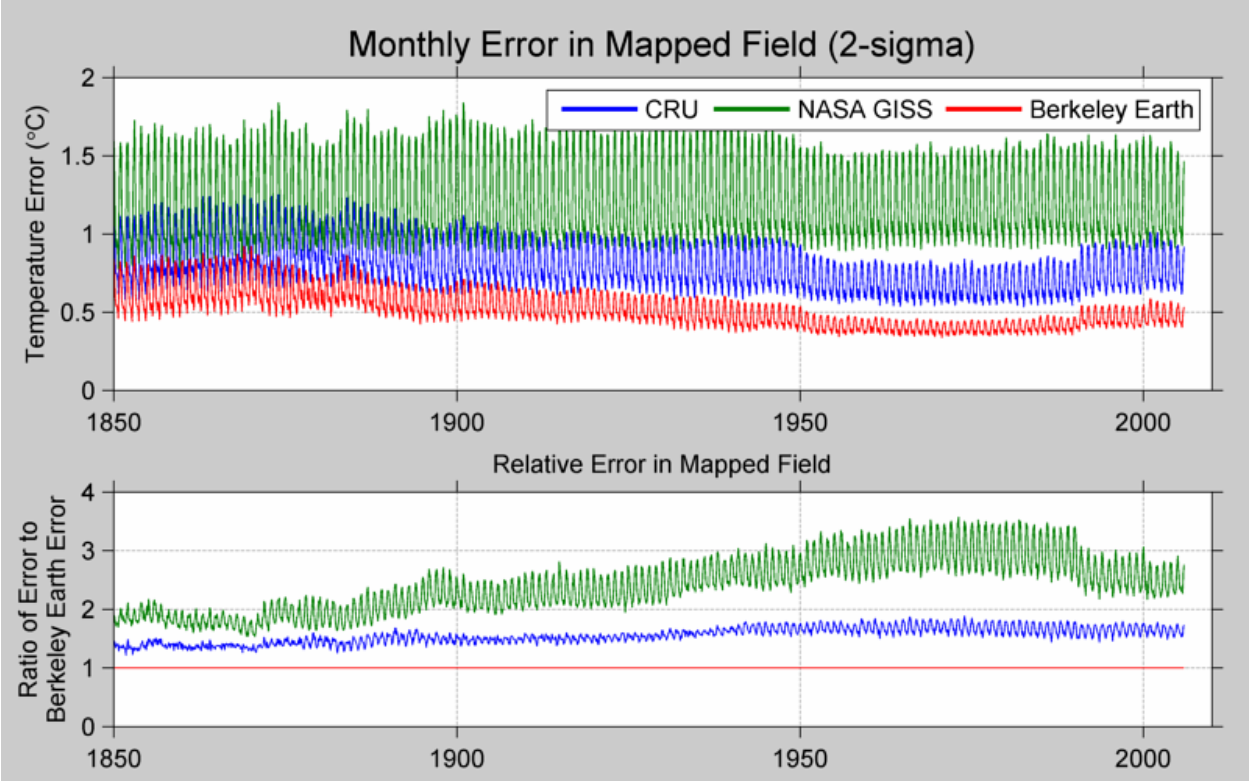

Figure 10: Same as Figures 8 and 9, but based on the typical monthly error at each occupied grid cell. As discussed, there is a significant seasonal component in how accurately the temperature anomaly field can be reproduced due to greater spatial variability in Northern Hemisphere winter.

In Figures 11, we show the equivalent annual average mapping errors as Figures 8, but consider all grid cells reported by both Berkeley Earth and GISS. By removing the constraint that CRU also report values, a substantially larger portion of the Earth can be considered (often twice as large, refer to Figure 3). As with the previous figures, we find that the Berkeley method reproduces the local structure of the simulated temperature fields more accurately than GISS. However, the difference is somewhat less severe than in Figure 8, presumably due to the inclusion areas farther from the weather stations where both methodologies will be intrinsically less accurate. The decadal and monthly figures at GISS reporting locations are similar but not shown here.

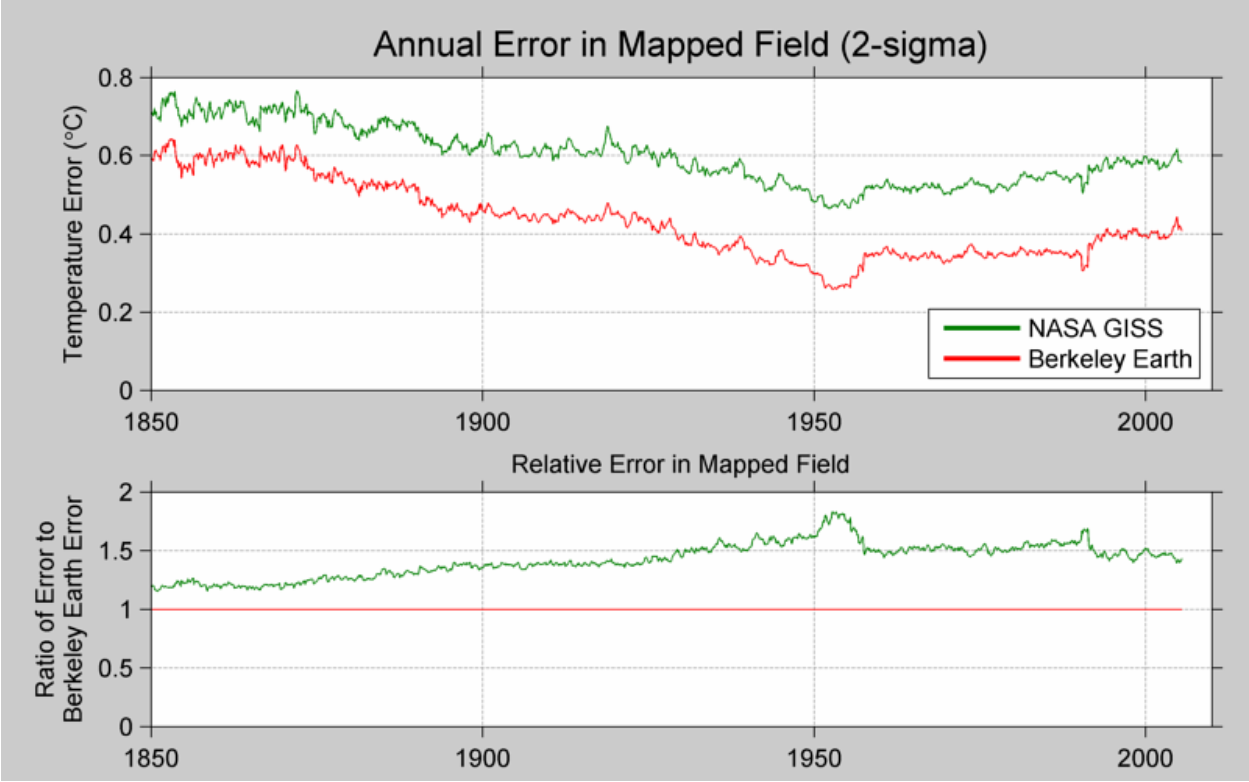

Figure 11: Same as Figure 8, except that all sites reported by GISS are considered. In many cases this means twice as much land area can be considered than when restricted to the CRU reconstruction locations, refer to Figure 3. As before, Berkeley is found to have superior accuracy in reconstructing the local field when compared to GISS

Accuracy of Measuring Trends

For a final analysis using this simulated data, we measure how accurately the various averaging techniques allow for the measurement of trends in the global land average over time. This is done over three time periods 1850 to 2005, 1900 to 2005, and 1950 to 2005. In each case we fit the reconstructed monthly averages to a line and compare the slope of the fit line to the true slope in the underlying GCM field.

The estimates of the error in measuring the trends are summarized in Table 1. As was the case with the direct estimates of the global field, we observe that Berkeley Earth is the most accurate at measuring the trend in the GCM field. Otherwise GISS is the second most accurate, while CRU typically introduces at least twice as much error in the measurement of land surface temperature trends as the Berkeley techniques.

	Error in trend reconstruction (°C / century, 2-sigma)		
Method	1850-2005	1900-2005	1950-2005
CRU	0.17	0.11	0.17
NASA GISS	0.12	0.055	0.085
Berkeley Earth	0.087	0.045	0.078

It is valuable to remind the reader that the simulated data used in these calculations were created so as to be free from any measurement errors or other biases. Hence, the errors reported here only reflect the uncertainty introduced by each averaging method in its attempt to combine sparse, spatially incomplete data. Averages involving real data will have to contend with additional sources of noise and bias, though in some cases there may also be additional stations which would help to combat noise.

We also note that even the worst performer in this test has estimated errors that are substantially smaller than the temperature trends believed to have been occurring during the twentieth century. We generally believe that all of the groups are easily capable of detecting climate change; however, we also believe that the Berkeley methodology allows one to be more precise in doing so.

Conclusions

We have provided an analysis of the effectiveness of three different averaging methodologies while using simulated data where the underlying true evolution of the field can be known exactly. We find that the Berkeley Earth method outperforms the other techniques considered in its ability to estimate the global land average, to reconstruct the details of the mapped field, and to measure the long-term trend. Hence, we believe the Berkeley Earth technique should be preferred in the reconstruction of climate fields.

In particular, the CRU methodology was found to be much less accurate at the reproduction of global averages and global trends than other methods and often exhibited 2 to 3 times as much error as the Berkeley methodology. Their techniques may be simpler to understand, and in some cases that is a virtue, however their gridding approach is appreciably limited by the incompleteness of the fields they produce. In the reproduction of global averages and trends, the GISS methodology was more similar in accuracy to the Berkeley methods, with only a 10-60% increase in error.

However, in the examination of mapped fields, GISS had the poorest performance, with local errors exceeding the errors in the Berkeley method by a factor of 2 to 3. CRU had a better performance at the reproduction of the local field, with a 50% increase in error over Berkeley being typical. We believe that the large-scale averaging intrinsic to the GISS method tends to blur out the details of the temperature anomaly field, and that it is as a consequence of this blurring that GISS is less accurate in its ability to reproduce the mapped field.

The present analysis has considered only the uncertainties introduced by the different methods of interpolating discrete temperature data. Beyond this, a full analysis of uncertainty still has to consider the effects of noise and bias as it affects the accuracy of the data. Such concerns will be explored elsewhere. However, the uncertainty associated with the reconstruction method itself imposes a fundamental limit on the possible accuracy of each reconstruction since noise and other biases can only serve to make the resulting reconstructions less accurate. Hence understanding the relative effectiveness of each averaging method is important in understanding the ultimate uncertainty associated with the reconstruction of global climate.

References

- Jones, P.D., Lister, D.H., Osborn, T.J., Harpham, C., Salmon, M. and Morice, C.P., 2012: Hemispheric and large-scale land surface air temperature variations: an extensive revision and an update to 2010. Journal of Geophysical Research, doi:10.1029/2011JD017139.
- Hansen, J., R. Ruedy, Mki. Sato, and K. Lo, 2010: Global surface temperature change. Rev. Geophys., 48, RG4004, doi:10.1029/2010RG000345.