ANALYSIS OF REVISIONS TO USHCN DATA

KEVIN KILTY

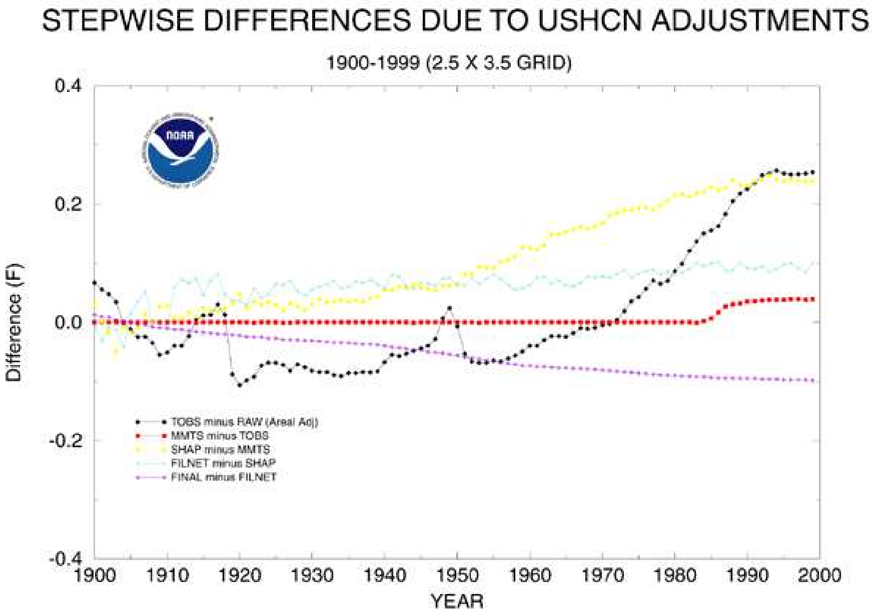

Figure 1. Stepwise summary of the effect that corrections have on the entire USHCN station data.

The corrections that USHCN stations go through in order are as follows:

(1) Identify and remove outliers.
(2) Adjust for Time of Observation Bias (TOB)
(3) Adjust for changes in equipment
(4) Homogenizing the data to account for discontinuities.
(5) Make substitutions for missing/censored data
(6) Adjust for the Urban Heat Island (UHI)

Date: December 28, 2009.

In the sections that follow I discuss questions that remain unanswered about these adjustments after reading the NCDC descriptions, and also my thoughts after reading the literature behind the adjustment schemes. The segments in italics are direct descriptions from the USHCN web site.

1. Removing Outliers

A quality control procedure is performed that uses trimmed means and standard deviations in comparison with surrounding stations to identify suspects (> 3.5 *standard deviations away from the mean) and outliers (*> 5.0 *standard deviations). Until recently these suspects and outliers were hand-verified with the original records. However, with the development at the NCDC of more sophisticated QC procedures this has been found to be unnecessary.*

There is no explanation of the degree of trimming (i.e. was it 10% or 20%?). Certainly the trimmed mean is a more robust measure of central tendency than is the raw average. The closer to 50% trimming, the more like a median is this measure of central tendency. Also, is any adjustment made to remove the annual cycle from the data before calculating statistics?

2. TOB

Next, the temperature data are adjusted for the time-of-observation bias (Karl, et al. 1986) which occurs when observing times are changed from midnight to some time earlier in the day. The TOB is the first of several adjustments. The ending time of the 24 hour climatological day varies from station to station and/or over a period of years at a given station. The TOB introduces a non climatic bias into the monthly means. The TOB software is an empirical model used to estimate the time of observation biases associated with different observation schedules and the routine computes the TOB with respect to daily readings taken at midnight. Details on the procedure are given in, "A Model to Estimate the Time of Observation Bias Associated with Monthly Mean Maximum, Minimum, and Mean Temperatures." by Karl, Williams, et al.1986, Journal of Climate and Applied Meteorology 15: 145-160.

The best place to begin research on this TOB adjustment is Donald G. Baker, 1975, Effect of Observation Time on Mean Temperature Estimation, J. Applied Meteorology, 14, 471-476. The presentation there is brief and very straight forward. In effect there is a mean diurnal cycle of temperature and variations of this cycle from day to day. Observations made at times other than midnight will occasionally fetch extremum temperature from the cycle on one day and enter it into maximum or minimum temperature on this and the following day. This double sampling results in a temperature bias. Cooperative stations read in the morning have a cool bias, while those read in the afternoon have a warm bias. The magnitude of the bias can't be calculated exactly as it is stochastic. It depends on the temperature of days preceding the current one. It is a function of station location, and possibly also of season. It depends on the exact course of weather, however, meaning that models built in one time period may not work well in other time periods. Baker built a model from stations in the St. Paul area and it applies to them alone. He was fortunate to have a continuous record of temperature from one station, which he could compare to nearby hourly stations. In essence he could concoct a curve of what cooperative station readings taken at any time of day would look like in comparison to true average temperature.

Karl, et al., consider a model of the diurnal variation in terms of two parameters, δ and ρ. Parameter δ represents the mean monthly variation of maximum and minimum daily temperatures (interdiurnal temperature differences). Parameter ρ represents a diurnal cycle of temperature independent of δ. If either is zero, say Karl, et al., then the TOB is zero, for reason that fetching extreme temperatures from the preceding day do not matter. They found these parameters through multilinear regression, that is interactions among parameters, by use of a model equation for Drift Corrected TOB $DCTOB = a(\delta\rho B)$; where, B is the base curve of whatever element, maximum, minimum, or mean temperature, is being corrected. This base curve is a time series averaged from the TOB corrections for a group of 79 first-order stations for each solar month.[1] Also it is constructed with reference to hour from local sunrise to mitigate issues of east-west distance in a time zone. This base curve is an initial estimate of TOB and explains about 80% of it according to Karl, et al. Meanwhile δ varies slowly across the U.S. and has a slightly different variation each month. Parameter ρ varies even more slowly across the U.S. reaching greatest magnitude during the solstice months and in the Southwest. The parameter ρ appears to be redundant as the information it carries is already encapsulated in the other two parameters, except for the month of December. The described correction explains about 90% of the TOB at a set of 28 independent test stations. Here are a couple of thoughts.

[1]Solar Month is based on the elevation of the sun, at a maximum, in 6 degree bins, with month one being for elevations less than 22.5 degrees and month 12 being elevations 82.5 degrees and above. What this does is avoid combining stations at greatly varying latitudes during the same month.

(1) Does the corrections scheme depend on the time period over which it was built (1958-1964)? Should it be tested against other time periods?
(2) The observed effect this correction has had on the USHCN data requires that a large number of new stations, or operators of existing stations, switched to afternoon observing times until about 1930, and then slowly switched to morning observing times afterward. However, there are very rapid changes in this general pattern that are difficult to explain. For example there are very rapid changes of 0.1 degree or more just after WWI and around 1950.

3. Equipment change

Temperature data at stations that have the Maximum/Minimum Temperature System (MMTS) are adjusted for the bias introduced when the liquid-in-glass thermometers were replaced with the MMTS (Quayle, et al. 1991). The TOB debiased data are input into the MMTS program and is the second adjustment. The MMTS program debiases the data obtained from stations with MMTS sensors. The NWS has replaced a majority of the liquid-in-glass thermometers in wooden Cotton-Region shelters with thermistor based maximum-minimum temperature systems (MMTS) housed in smaller plastic shelters. This adjustment removes the MMTS bias for stations so equipped with this type of sensor. The adjustment factors are most appropriate for use when time series of states or larger areas are required. Specific details on the procedures used are given in, "Effects of Recent Thermometer Changes in the Cooperative Network" by Quayle, Easterling, et al. 1991, Bulletin of the American Meteorological Society 72:1718-1724.

There is no discussion about equipment degradation nor about degradation of the station site. Apparently NOAA examined the effect of rooftop sensors in a project during the 1990s. However, one would have thought such a project to reveal all sorts of issues with the USHCN network as the surfacestations.org project has. It apparently did not. It will be somewaht difficult to separate the effect of site degradation/change with UHI effect. This might cause a very large correction.

4. Homogenizing the data

The homogeneity adjustment scheme described in Karl and Williams (1987) is performed using the station history metadata file to account for time series discontinuities due to random station moves and other station changes. The debiased data from the second adjustment are *then entered into the Station History Adjustment Program or SHAP. The SHAP allows a climatological time series of temperature and precipitation adjustment for station inhomogeneities using station history information and is the third adjustment. The adjusted data retains its original scale and is not an anomaly series. The methodology uses the concepts of relative homogeneity and standard parametric (temperature) and non parametric (precipitation) statistics to adjust the data. In addition, this technique provides an estimate of the confidence interval associated with each adjustment. The SHAP program debiases the data with respect to changes other than the MMTS conversion to produced the "adjusted data". Specific details on the procedures used are given in, "An Approach to Adjusting Climatological Time Series for Discontinuous Inhomogeneities" by Karl, and Williams, Jr. 1987, Journal of Climate and Applied Meteorology 26:1744-1763.*

In theory, homogenization of temperature series ought to accomplish by itself the sum total what all of these individual adjustments seek. It should be able to make reasonable corrections for station moves, equipment changes, degradation of equipment and environment, and changes in observation schedule. However, the process is fraught with technical challenges and potential circular reasoning. As Peterson and Vose say in their overview of GHCN adjustments

> Building a completely homogeneous reference series using data with unknown inhomogeneities may be impossible, but we used several techniques to minimize any potential inhomogeneities in the reference series

Indeed, making a series of homogenized stations when none of the stations is known to be free of data quality problems in the first place is a *chicken and egg* sort of problem. The goal is to produce a database of homogeneously comparable stations.[2] The definition od such is

> A climatoligical series is relatively homogeneous with respect to a synchronous series at another place if the temperature differences (precipitation ratios) of pairs of homologous averages constitute a sereis of random numbers that satisfies the law of errors.

My first thought is why should this be the goal? Isn't it possible that even nearby stations may show divergences in long term weather patterns or climate? Thus, the urge to homogenize to excess may create consensus among stations that truly ought not exist.

[2]Conrad, V., nd l.w. Pollack, 1950, Methods in Climatology, Harvard University Press. pp. 459. Quoted in Karl, T.R., and C.N.Williams, Jr. 1987.

The path to accomplishing such a goal is to compare many stations to one another statistically and make adjustments to reduce all differences to random numbers. My suggestion would be to do so simultaneously for all stations involved, but the method of Karl and Williams is to examine candidate stations one at a time.

Here is a list of caveats presented by the authors, in italics, paired with my thoughts unitalicized.

- *The simulation studies indicated that in some circumstances, depending on network characteristics, adjustments to data based on changes before and after a potential discontinuity can actually make the data more biased than if no adjustment had been applied.* (p. 1761) This speaks for itself.
- *Station histories rarely include information on environmental changes around the station.* This may produce a slowly varying signal difficult to spot with the homogenization procedure, and then becomes part of the network to which further candidate stations are compared.
- *Stations with nonclimatic progressive changes due to urbanization may lead to inappropriate adjustments at nearby stations.* Interestingly in view of this caveat, the USHCN sequence of adjustments is to correct for urban effect *after* this data homogenization step.

According to the authors, the problem urbanization presents is mitigated because 90% of stations in the HCN have population less than 50,000. However, Karl, et al., have shown there are urbanization effects at stations with poulations as small as 2,000.

Karl and Williams suggest the best solution as one that avoids using urban stations in the adjusment method. Yet, the urban stations are those least contaminated with equipment issues, and changes in reading schedules. Arguably one is ignoring data of best quality by excluding such stations.

Frankly I also see the potential here for some circular logic that proceeds as follows: Contamination of cooperative station data with site and equipment degradation effects, and first-order station data with urban heat island effects, propagates into all station data via the homogenization process. This effect becomes amplified as more stations undergo homogenization adjustments. Circular problems like this are extremely difficult to visualize when one does not have the "big picture" of the adjustment process, and when the result is what the investigator believes to be the true in the first place.[3]

Finally, the Karl and Williams suggest having a sufficient number of "nearby" stations for comparison. They could use 20 because of the density of USHCN stations. In the GHCN, though, there are important stations used to represent large regions that have no neighboring stations at all. While Karl and Williams provide no explicit limit for how far apart neighboring stations can be, the largest separation they allude to is $150\,km$. The contiguous U.S. is covered by about 12,000 reporting stations, not counting private ones. If these were uniformly distributed they would be about $21\,km$ apart. I can't imagine any place in the world providing better coverage. Thus, how well the process works in the U.S., in fact in the northestern U.S., is not a guide to how it will work globally.

5. Missing/Censored values

Estimates for missing data are provided using a procedure similar to that used in the homogeneity adjustment scheme in step three. This fourth adjustment uses the debiased data from the third adjustment (SHAP) and fills in missing original data when needed (i.e. calculates estimated data) based on a "network" of the best correlated nearby stations. The FILNET program also completed the data adjustment process for stations that moved too often for the SHAP program to estimate the adjustments needed to debias the data.Each of the above adjustments is done is a sequential manner. The areal edits are preformed [sic] first and then the data are passed through the following programs (TOBS, MMTS, SHAP and FILNET). At the end of each program, a dataset is produced and the graphs below show the annual temperature departures for each of the adjusted values.

The process here is similar to the process for homogenizing the data. Therefore, refer to that category for comments.

6. Urban Heat Island Effect

The final adjustment is for an urban warming bias which uses the regression approach outlined in Karl, et al. (1988). The result of this adjustment is the "final" version of the data. Details on the urban warming adjustment are available in "Urbanization: Its Detection and

[3]I ran into this sort of issue while battling the recovery of past temperatures through borehole data in the 1990s. Note in this instance that Karl and Williams judge the rectitude of their adjustment process by the way it makes stations conform to their pre-existing beliefs about local climate. (bottom of p. 1758, for instance.)

Effect in the United States Climate Record" by Karl. T.R., et al., 1988, Journal of Climate 1:1099-1123.

Once again Karl, et al., look for a simple model the parameters of which they can estimate with regression. This time the model is $T_{(u-r)} = a(population)^{0.45}$. The parameter a has values different for the various seasons, different values for different sized cities, and different values for each of the weather elements (maximum, minimum, average, and range).

Karl, et al., use paired stations, one of which has high population and the other of which remained rural (< 700 population) over the time period 1901-1984. They used the census counts from 1920, 1950, and 1980 for population information.

It all seems to be a reasonable approach. However, I thought of the following:

(1) There is no discussion of correction methods for elevation difference nor latitude, even though these corrections were apparently made.
(2) I am not familiar with but a few of the stations in the study. However, I found Laramie, Wyoming paired with Chugwater, Wyoming to be very odd. They are 120km apart on opposite sides of the front range of the Rockies, and if any two stations were to ever be affected differently by a secular drift in the frequency of air mass regimes, they would be a candidate pair. I wonder how many other pairings are similarly odd and what affect this may have on correction results? Cheyenne paired against Pine Bluffs does not appear so odd, but still, because of elevation difference these two stations might have different frequency of air mass regimes.
(3) The temperature contrasts between stations are not used directly, but rather they are weighted according to what National Climatic Region they are within. Thus, the truly enormous cities are found primarily in narrow climatic zones along coasts, or around the Great Lakes and I imagine this has reduced their impact compared to a few stations located in the large regions represented by steppes and western plains. The process looks to me like it would have muted the UHI effect. I doubt the global average temperature is similarly weighted and so there is some inconsistent treatment of data going on that is difficult to see clearly. It may have very small impact of course, but it is not addressed in the report.

(4) The corrections for the four elements are related to one another in the following way: $average = \frac{max+min}{2}$ and $range = max - min$; consistency would be best demonstrated if the corrections for the four elements satisfied the same relationship. In a few cases they do so well, but not in others. For instance, Table 5 in their report reports model parameters for cities over 100,000 in population, and in this the models for maximum and minimum temperatures produce the range and average quite well.
(5) If one examines the correction curves against scatter plots of data (Figures 11 in Karl, et al.) what is most striking is how puny these corrections are against the huge variation in the observed elements themselves. In the plots of range correction, in particular, urban minus rural temperatures vary from minus 9 to plus 5 $^{\circ}C$? What is going on that supposedly paired stations differ by so much? I suspect that Laramie paired with Chugwater exhibits just this sort of behavior; and, propose that stations are often not well matched to one another. Also, one can see that the corrections are influenced enormously by a few really large cities, even though I just argued that this data was discounted too heavily by bringing in climatic zones weighting. This influence is a result of weight being provided by population.
(6) The 95% confidence intervals for the parameter (a) is always quite large, indicating that statistically acceptable corrections could be very different in value–twice as much down to half as much.
(7) The table of maximum temperature differences between urban and rural areas according to Oke's model (Table 8) shows some very large values. Certainly these are instantaneous maxima, and apply to calm, nocturnal conditions, but still they are more like the sorts of differences we have come to expect by looking at measurements in urban heat islands. They make me even more skeptical about the value of this correction model.
(8) Karl et al., themselves, list the limitations of this method. The first is that it will underestimate UHI for small stations even though the underestimate is probably small. They do not recommend using this process to predict the impact of UHI at any particular station, even though this seems to me to be exactly what the USHCN adjustments are doing. Instead, they say, these corrections should be applied only if a number of stations are available to represent regional averages. I am not sure

> what this means. They caution against applying these corrections globally because cities even in Western Europe are not like cities in North America. They also speak of the really large UHI effects of places dominated by stable air in winter (Think Antarctica or places in the Arctic.)

7. Conclusions

After reading these documents many times I feel I can draw several conclusions:

1) There have been many changes to cooperative stations, equipment, observation scedules, and so forth that I was not aware of. These require adjustments which are sometimes large in comparison to the precision required for some uses of this data. This is not a good position to be in unless the corrections are known to be very precise. I doubt anyone can argue this is the case here.

2) Perhaps the adjustment schemes depend upon too great a degree on models built from regression. This situation would be improved if one could figure adjustments on the basis of physical principles that could be applied per station.

3) Considering that TOB adjustments produce a signal very much like greehouse warming itself, what can we do to minimize this confounding? My suggestion would be to limit collection of raw data to first-order stations that require no TOB corrections. There will still be issues of changes in how stations have collected data, and changes in instrumentation, but these seem less problematic in my mind.

4) For making global temperature estimates, perhaps we need to weight data in the manner Karl, et al., did for making UHI adjustments. Collect stations into climatic zones, average, then weight by the area of the climatic zone. We would be able to work with less data of higher raw quality, and get away from the arbitrary looking assignment of single, inappropriately sited stations to huge land regions (like Antarctica).[4]

---

[4] I had actually thought about doing this years ago, because it had worked well for Chapman and Pollack when they were using a very small data set to compute heat flow over the African continent. In this case they used tectonic/physiographic provinces. But the idea is the same.

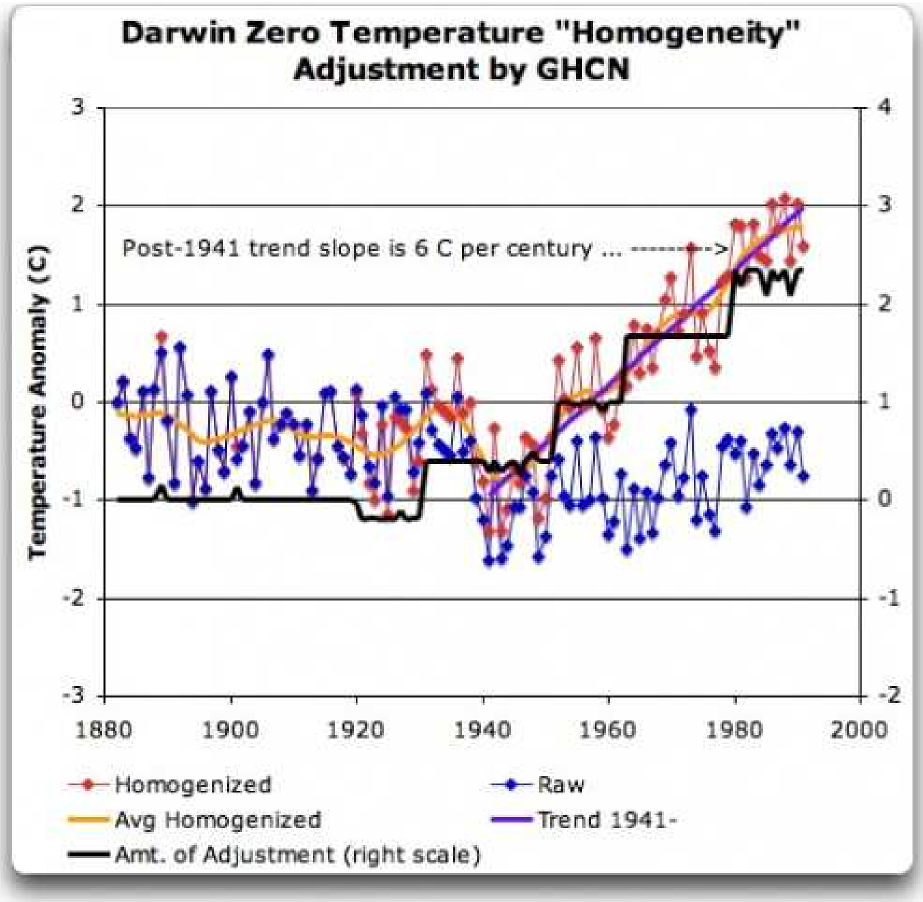

FIGURE 2. Darwin temperature adjustments (after Willis Eschenbach). Note the step-like adjustments, which appear to have come from station discontinuities that the adjustment process itself detected.

5) In the case of Darwin Zero, which Willis Echsenbach describes in detail on WUWT, the homogenization process appears to be adjusting data for a series of discontinuities (the step-like changes) and then makes progressive adjustments based on comparison with other stations. Do the discontinuities actually exist? Can one point to the events that resulted in the discontinuity? Was this process let loose on its own, using inappropriate and inadequate coparisons, and is one example of homogenization increasing the bias rather than reducing it, as Karl and Williams admit sometimes occurs?

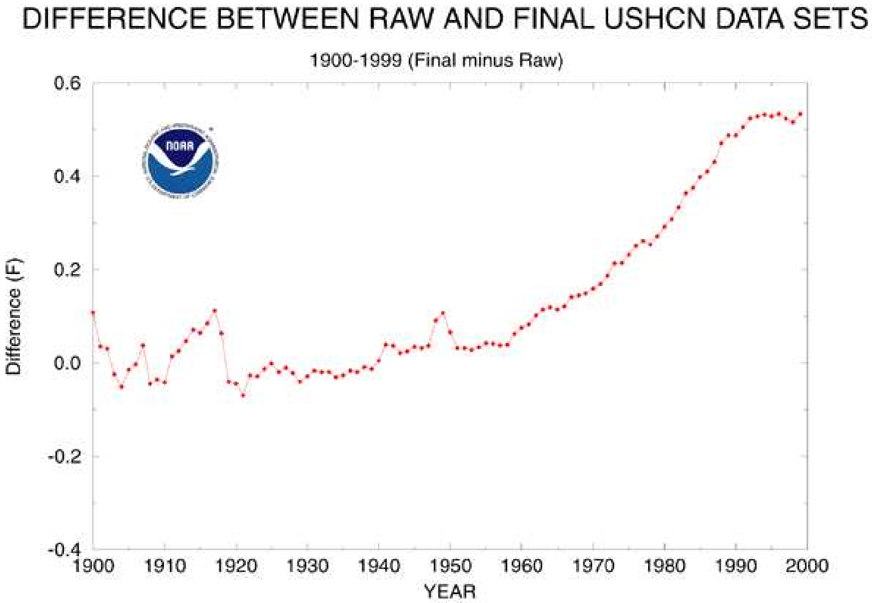

FIGURE 3. Overall summary of the effect that corrections have on the entire USHCN station data.

6) I find it surprising that every correction except UHI enhances the alleged greenhouse effect in the time series. I do not understand why the homogenization process should do this unless it is picking up unaddressed TOB and UHI. What I feel might be happening is that UHI, which is still in the data at this point, as the correction is apparently done out of order, is corrupting the homogenization effort. Note that the homogenization correction is almost a mirror image of the UHI.