Homogenization of Temperature Data

An Assessment

Kevin Cowtan

The author would like to thank Bärbel Winkler, Andy Skuce, Susannah Cowtan and Zeke Hausfather for comments on drafts of this report.

Contents

1. Foreword

People learn in different ways. Some learn by reading. Others by watching. Still others by doing.

I fall into the third group. So when a debate emerged in the media concerning the homogenization of historical temperature data, the thing I most wanted to do was to go and redo the science for myself, from scratch.

I already had a little experience with identifying problems in small networks of weather station records by hand, so I had some feel for the shape of the problem. However the automated homogenization of global records was a closed book to me. There is no reason to believe that an individual with limited time can reproduce a useful part of the work of dozens of full-time investigators over many years. But, if the bulk of the problem is simple and most of the work goes into dealing with exceptional cases, then it may be possible.

In this case, the basics do appear to be simple. The presence and nature of inhomogeneities in the data are easily demonstrated, and in many cases fairly easily corrected. The tests I have performed are understandable by anyone with basic scientific literacy, and are easily reproducible by citizen scientists with only modest programming skills and limited time. A working, if rudimentary, global temperature homogenization package can be implemented in only 150 lines of computer code.

All the software and data used in this report are available from the website:

`http://www-users.york.ac.uk/~kdc3/papers/robust2015/`

Conclusions

For convenience the conclusions of this report are summarised in advance. The basis for these conclusions forms the body of the report.

- Are there inhomogeneities in the data?
  Yes, there are.
- Are those inhomogeneities of a form which would be explained by sporadic changes in the measuring apparatus or protocols?
  Yes, the largest inhomogeneities are explained by sporadic changes in offset in the

temperature readings.

- Can those inhomogeneities be detected by comparing records from neighbouring stations?
  Yes, most stations have other nearby stations with substantially similar records.
- Is there sufficient redundancy in the data to allow those inhomogeneities to be corrected?
  Yes, tests using multiple benchmark datasets suggest that inhomogeneities can be corrected.
- Does the Global Historical Climatology Network (GHCN) method produce reasonable estimates of the size of the adjustments?
  Yes, both neighbouring stations and reanalysis data support the GHCN adjustments.
- Do the observations support the presence of a trend in the homogenization adjustments?
  Yes, both methods suggest that the adjustments should have a slightly skewed distribution.
- Is there evidence that trend in the adjustments could be an artifact of the methods?
  Two possible sources of bias in the method were tested and eliminated.
- If the data are correctly homogenized, how large a change will be introduced in the global temperature trend?
  The size of the required correction to the global record is much harder to determine than the direction. The simple methods described in this report cannot provide an absolute answer. The most I can say is that the GHCN correction looks plausible.

2. Background

Near-surface air temperature data from historical weather station records form one key part of our understanding of past climate, which in turn provides significant information for the evaluation of future climate change in response to human activity. Weather station temperature observations are available spanning two to three centuries, however the instrumentation, instrument locations and environments, and data collection methodologies have changed significantly over that period. As a result, historical temperature records are usually subject to a retrospective calibration or homogenization process, to enable an accurate estimate of climate change to be obtained over the period of the record.

Data inhomogeneity may arise from a number of sources. Since air temperature varies with elevation, small changes in weather station location which change the elevation of the station above sea level can lead to a spurious increase or decrease in the temperature observations (when the station elevation is lowered or raised respectively). Temperature trends spanning the station move will contain a bias which is unrelated to real-world temperature change in the vicinity of the weather station.

Similarly, changes in station equipment, such as the thermometers or the screens which are used to protect them from the influence of direct sunlight, wind or rain, can introduce a bias in the temperature observations. The introduction of electronic thermometers across the US station network over recent decades has introduced opposite biases in daytime and nighttime temperature observations, and a smaller bias in daily mean temperature observations.

The times at which temperature observations are made can also affect the results. The time-of-observation (or Tobs) bias arises when a minimum-maximum thermometer is read at a time which leads to the corresponding temperature observation not being consistently assigned to a given calendar date relative to the date of observation. For example, if a maximum temperature thermometer is reset at the hottest part of an exceptionally hot day, it will immediately return to the same high temperature, which will then be recorded for two successive days.

For these reasons, weather station temperature data are homogenized to remove artifacts arising from changes in station location, instrumentation or measurement practices. In some cases these homogenizations are performed by national weather services on the basis of local informa-

Figure 2.1: Different temperature observation equipment. Picture: Project SCREEN, Center for Climate Change, Universitat Rovira i Virgili, Spain. (CC BY 3.0)

tion about the station. The weather service may then provide raw and/or homogenized data to users. The UK Met Office record in particular makes use of data which have been homogenized by national weather services.

Other projects, notably the Global Historical Climatology Network (GHCN) used by both NOAA and NASA in their temperature records (Lawrimore et al., 2011), and the Berkeley Earth surface temperature record (Rohde et al., 2013), start from the raw records and perform their own automatic homogenization calculations using a combination of metadata concerning weather station changes, along with internal consistency tests to detect weather station changes for which no metadata are available.

Researchers investigating regional climate change do not necessarily rely on global historical temperature products, instead producing their own local climate records using carefully curated networks of station records along with manual or automatic homogenization. There are a number of software packages available for the automated or assisted homogenization of climate data including temperature data. A selection is listed in Table 2.1.

Package	**Reference**
ACMANT	Domonkos, Poza, and Efthymiadis, 2011
AnClim, ProClimDB	Stepanek, 2005
Climatol	Guijarro, 2011
GAHMDI, HOMAD	Toreti, Kuglitsch, Xoplaki, and Luterbacher, 2012
HOMER	Mestre et al., 2013
MASH	Szentimrey, 1999
RHtests	Wang, Chen, Wu, Feng, and Pu, 2010
USHCN	Menne and Williams Jr, 2009
Berkeley Earth	Rohde et al., 2013

Table 2.1: Software packages for data homogenization, modified from the 'Climatol' website (`http://www.climatol.eu/DARE`).

Both the need for, and the effectiveness of data homogenization have been well established by numerous and diverse studies. However there is a systematic trend in the adjustments which, while small, increases estimates of global mean surface temperature change over the 20th century. The trend has been questioned in the public discourse around climate science.

It should be noted that homogenization is not confined to weather station temperature data. A more substantial adjustment is made to the sea surface temperature data, which in turn makes a larger contribution to the global mean surface temperature since the majority of the Earth's surface is water. However this adjustment reduces estimates of global mean surface temperature change Figure 2.2 - primarily due to a change in measurement methods around World War 2 which led to a spurious increase in the temperature observations. This adjustment more than cancels the land station temperature adjustments. Most of the impact of the adjustments is confined to the period before 1980, and thus does not affect recent temperature trends.

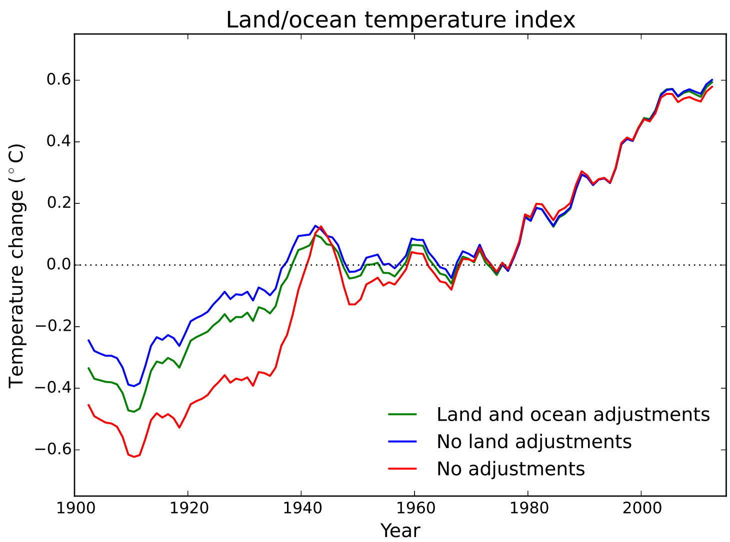


Figure 2.2: The impact of homogenization of land data alone and of both land and ocean data. Land temperature data from GHCN, sea surface temperatures from HadSST3 (Kennedy, Rayner, Smith, Parker, & Saunby, 2011).

It has been suggested that the trend in the homogenization adjustments might be evidence of a bias in the temperature homogenization process. The alternative is that the trend in the homogenization adjustments is a property of the data. How might the validity of the air temperature homogenizations be assessed?

When faced with a dispute concerning a complex technical topic, it is tempting try and draw a conclusion without engaging with that complexity. For example, an appeal could be made to philosophy of science to provide a rubric concerning how science should be done. However there are as many answers to the question of how science 'should' be done as there are philosophers of science. Similarly, the motives, preconceptions and methodology of the scientists cannot answer the question, since it is possible for scientists to reach the right answer for the wrong reasons.

If an objective answer to the homogenization question is to be determined, the complexity cannot be avoided. If the answer is to be objective, verifiable and reproducible, it must come from the data. It is of course possible that the answer may be unknowable on the basis of the current data, but that in turn should be determined from the data themselves.

An appropriate preliminary formulation of the homogenization question is therefore "Do the homogenized data better reflect the historical temperature evolution than the raw data?". The question is objective, and the answer is at least potentially knowable. Most importantly the question, if it is answerable, may be answered by the data.

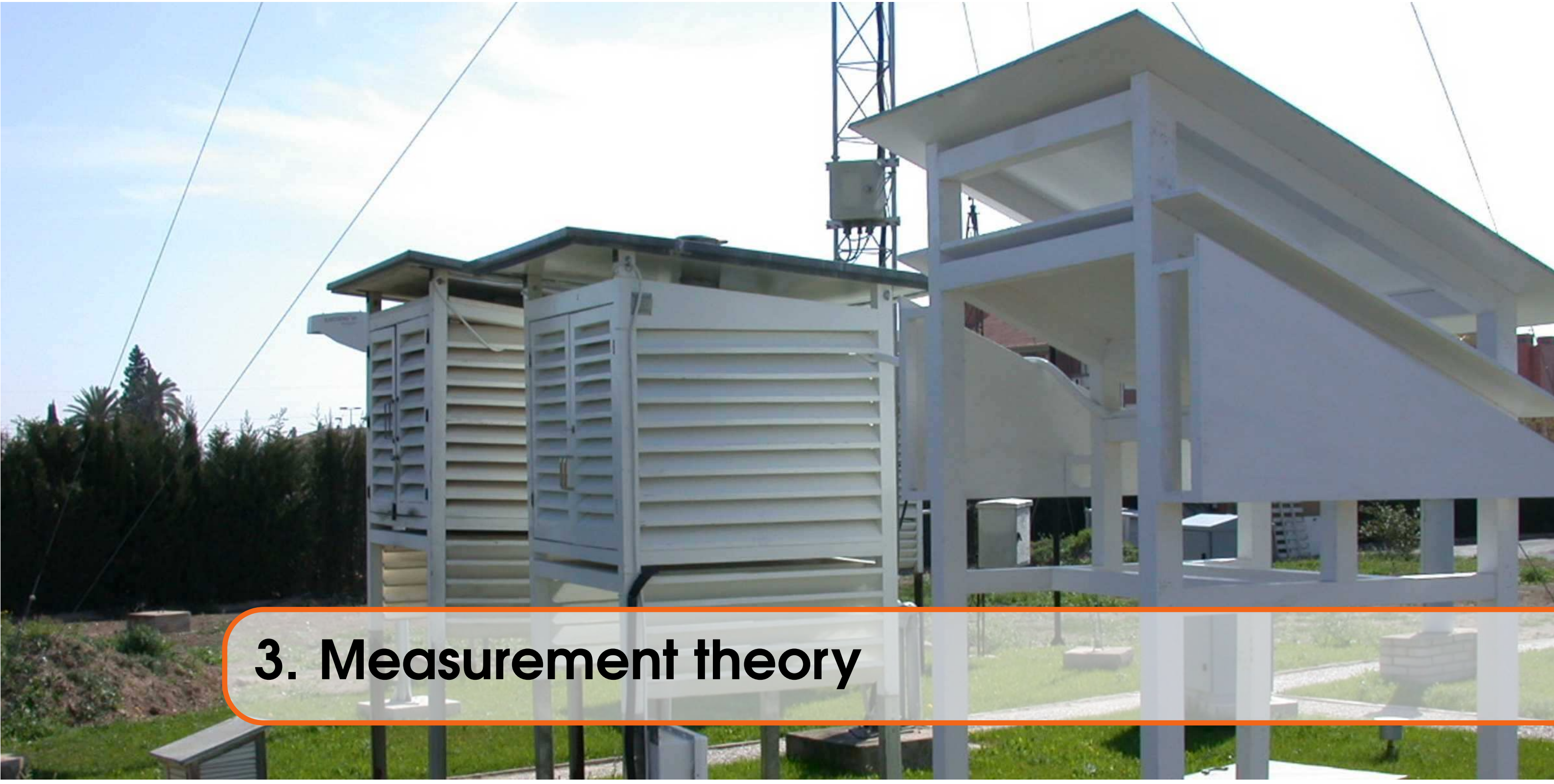
A further clarification to the question is required. For example, it is possible that homogenization might improve the correspondence of one subset of records to reality, while degrading another subset. Given that the data are in some cases both noisy and sparse, it is inevitable that some records might be degraded by homogenization. Similarly, even a faulty homogenization algorithm might improve some records by chance. How is the criterion of "better reflecting the historical temperature evolution" to be evaluated?

Of particular interest is the global mean temperature trend, which arises from a geographical mean of the local temperature trends. The temperature record may be considered to be improved if enough of the local records are improved to lead to a general improvement in the global record. The homogenization question can therefore be separated into two questions:

1. "Do *most of* the homogenized records better reflect the historical temperature evolution than the raw records?"
2. "Do *the bulk properties* of the homogenized records better reflect the historical temperature evolution than the raw records?"

The purpose of this report is threefold:

- To establish how the data may be interrogated to determine the validity of the homogenizations.
- To interrogate the data to determine to whatever extent possible the validity of the homogenizations.
- To provide a foundation on which competent citizen scientists can investigate the validity of the homogenizations for themselves.



3. Measurement theory

The underlying problem of measurement is that in most cases, observations must be carried out indirectly by means of an instrument. The observation (by which I mean the value that we record), is a property not just of the thing that we are observing. Rather it combines properties of the thing we are observing, the instrument we are using to perform the observation, and often other parts of the system which are neither the thing we are interested in nor the instrument.

Changes to the instrument (such as a thermometer), or to the rest of the system (such as the surroundings of a weather station), will lead to changes in our observations, even if there is no change in the thing we are observing (in this case the local temperature). In order to determine a property of the thing that is being observed, we need to account for changes in the instrument and the remainder of the system.

This often leads to a distinction between raw data (the numbers produced by the instrument), and derived data (which have been adjusted to remove factors not relating to the thing being observed). The raw data remain unaltered, but the derived data are adjusted to remove both instrumental artifacts, and confounding factors not relating to the thing being observed. Those adjustments may in turn change over time as our understanding of the confounding factors evolves, or as more data become available.

Example: Measurement of X-ray diffraction patterns

X-ray diffraction patterns provide the most important source of evidence concerning the structure of biological macromolecules, and have led to the determination of more than 100,000 atomic structures deposited in the worldwide Protein Data Bank (wwPDB). These have in turn been used in areas from fundamental biology to biotechnology, the development of new medications, and industrial chemistry. Extrinsic evidence for the value of the method comes from the industrial value (measured in billions of dollars) and from the 15 or more Nobel prizes awarded for X-ray crystallography-based research.

The diffraction pattern is collected by exposing a crystal of the substance of interest to an intense X-ray beam, typically produced by a synchrotron, and measuring the resulting diffraction spots, these days using a semiconductor-based detection system.

The intensity of any diffraction spot depends on a whole range of factors including:

1. The intensity of the original X-ray beam.
2. The length of the X-ray exposure.
3. The proportion of the X-ray diffraction which is excited by the current crystal orientation (and motion).
4. The volume of the crystal which intersects the X-ray beam in the current crystal orientation.
5. The volume of non-crystalline matter (such as the crystal support and solvent, and air) absorbing the incident and diffracted beam.
6. Damage to the crystal by X-ray exposure.

All of these are subject to uncertainties. The solution of a structure by X-ray diffraction often involves the accurate determination of small differences in the intensity of diffraction spots; however these differences are often swamped by differences due to combinations of the above factors and others. In order to determine the structure, the data must be homogenized by the application of a scale factor to each diffraction image (and often to the individual regions of the image).

While attempts have been made to determine some of the scale factors by observation, in practice the most effective method is to perform an empirical scaling based on internal consistency metrics. We know that multiple observations of the same diffraction spot should produce similar values (although not the same), and so scaling is performed to optimize the internal consistency of the data. Even if some factors, such as the intensity of the beam, were measured directly, homogenization can pick up known and unknown factors which can not be explicitly corrected.

3.1 Extrinsic versus intrinsic adjustments

In some cases, corrections to raw data may be performed on the basis of extrinsic measures of confounding factors. For example, if there is a known change to the time of observation of air temperature, an adjustment can be made to reconstruct the most probable temperature at some standard time on the basis of the recorded time. If there is a change in the instrumentation, both the old and new instruments may be used in parallel to determine and correct for the impact of the change. In the case of X-ray diffraction, there have been attempts to measure the crystal shape and thus deduce the volume intersecting the X-ray beam; however these have not so far produced useful results. I will refer to these as extrinsic adjustments.

In other cases, there may be no data on which to base an intrinsic adjustment. The variations in intensity of the X-ray source during a diffraction experiment are not generally recorded along with the diffraction data. In the case of historical temperature data, there may simply be no records available concerning changes to the instruments. In this case the confounding factors could be assumed to be small or to cancel out. This assumption may be tested in the case when there is some redundancy in the data, for example many observations of the same diffraction spot, or multiple observations from nearby weather stations. If the confounding factors are not small, then a correction must be estimated using the redundancy of the data. I will refer to these as intrinsic adjustments.

Intrinsic adjustments have an important advantage over extrinsic adjustments: since they can be made without the need for additional metadata on the confounding factors, they can also adjust for confounding factors which have not been anticipated. In the X-ray case, an adjustment for beam intensity also corrects at least partially for the diffraction geometry, absorption, and other factors. In the case of temperature observations, an intrinsic adjustment for station location or instrumentation can also account for unrecorded systematic changes in time of observation.

However, intrinsic adjustments have a limitation: each degree of freedom in the adjustment function introduces an uncertainty into the result. It is therefore important to ensure that the number of degrees of freedom in the adjustment function is small compared to the level of redundancy in the data. In the extreme case the data may be adjusted to be perfectly internally consistent, at which point the results become a property of the adjustment method rather than the data.

This issue may be mitigated in several ways:

1. The adjustment function is chosen to be parsimoniously parameterized. For example, if change points are being introduced to reconcile time series, a limit can be imposed on the frequency with which change points can be introduced.
2. Tests may be performed on the explanatory power of each new adjustment parameter to determine whether the new parameter is explaining a genuine part of the signal, or merely mopping up noise.
3. The methods may be validated on synthetic data with realistic errors. With synthetic data, we know the correct answer, and so can determine whether the adjustment method is improving the data or not.

3.2 Types of adjustments

The nature of the adjustment function varies according to the system being studied, and must be determined in part on the basis of an expectation about the nature of the confounding factors affecting the observations. Adjustments may involve:

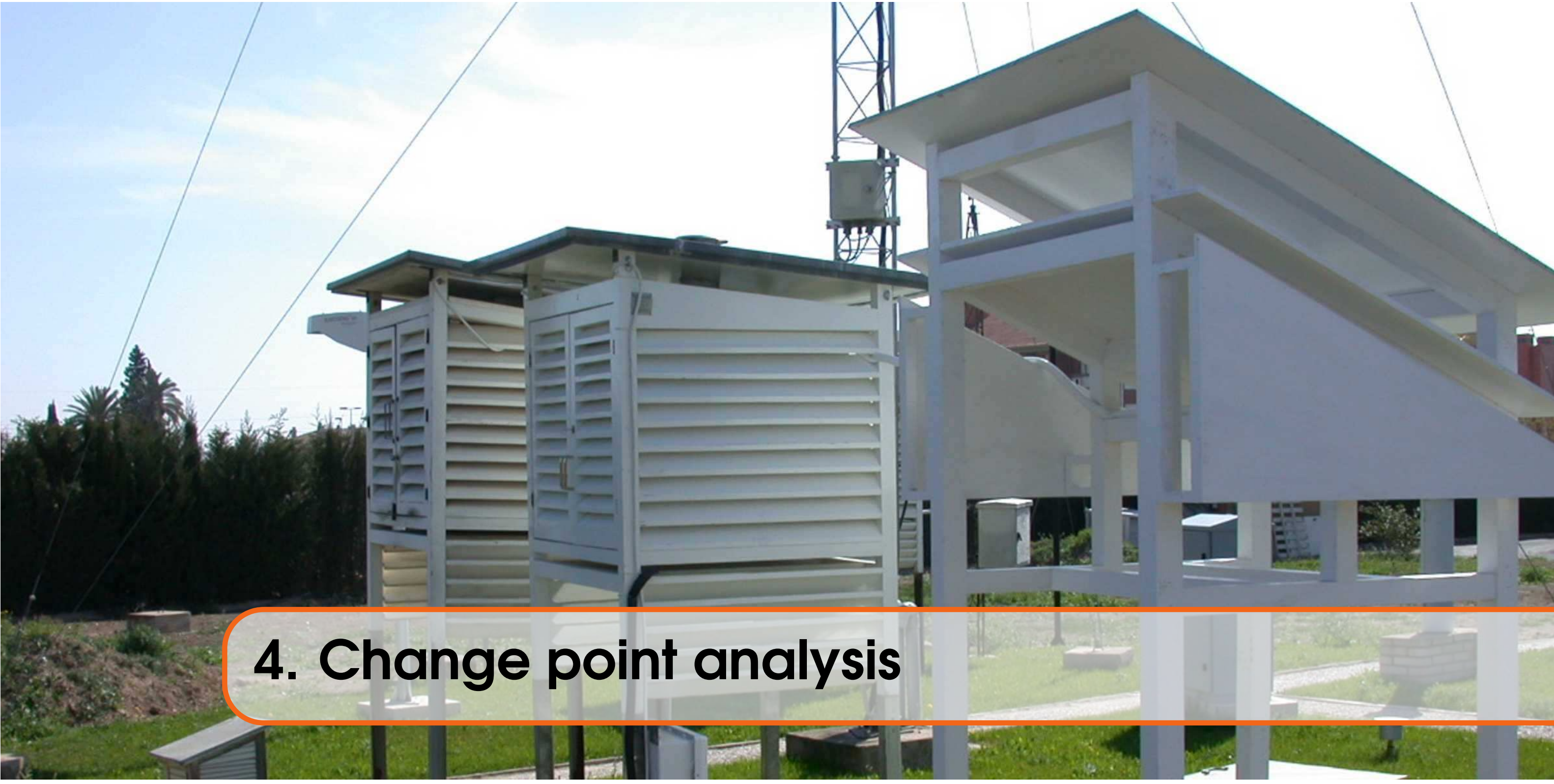
1. An offset: a constant is added to or subtracted from a set of adjustments.
2. A scale factor: a set of observations are multiplied or divided by some constant.
3. A nonlinear transformation: the observations are adjusted by some more complex function.

In the case of temperature data, thermometers must accurately capture the large variations in daily and seasonal temperature, otherwise they would immediately be identified as having a calibration problem. As a result, they do not generally suffer from major scaling inhomogeneities. However, detection of a change in climate over a long period can be influenced by a small change in the offset between the recorded observation and the true temperature. Since nearby weather station sites can have temperatures which differ from each other due to local factors, an offset inhomogeneity may go undetected. Thus homogenization by addition or subtraction of a constant to different segments of a record may be required.

By contrast, X-ray diffraction observations are based on counting statistics, and so the zero point is well known (to a first approximation and ignoring background scattering). Variations in the X-ray beam and the volume of the crystal in the beam introduce inhomogeneities of scale.

A third case is the calibration of radiocarbon dating data. In this case the inhomogeneities arise from variation over time in the amount of Carbon-14 in the atmosphere, which in turns affects its uptake by plants. In this case the homogenization is a nonlinear correction which is empirically derived from samples of known age and other sources. This is an example of an extrinsic rather than an intrinsic homogenization.

The homogenization algorithm must be appropriate for the type of errors present in the data, which in turn must be determined from the data themselves.



4. Change point analysis

Any temperature homogenization algorithm must be built around assumptions concerning the nature of the inhomogeneities in the data. For example, it is commonly (although not universally) assumed that a weather station record will consist of spans of homogeneous observations, with occasional but infrequent changes in the offset between the observation and reality. The existence of an offset is not a concern since the aim is to detect change in temperature over time; however changes in that offset introduce a spurious change in the observations which obscures the real temperature change.

The validity of a homogenization algorithm is contingent on the validity of the underlying assumptions concerning the nature of the inhomogeneities. If those assumptions are invalid, the results of the homogenization algorithm will be impacted. Therefore the assumption of spans of homogeneous data with infrequent breaks should be tested.

If the assumption is valid, then this should be apparent in the difference between temperature records from neighboring weather stations. If the stations are close enough to measure similar weather, the difference between the station readings will consist of an offset due to the differing locations and instrumentation, plus noise terms arising from any differences in weather between the sites, and various sources of noise in the individual observations. In addition, the difference temperature series will show a 'jump' at any point where either station suffers a break in homogeneity leading to a change in the temperature offset from reality for that station.

In other words, if the assumed pattern of inhomogeneities is valid, the difference temperature series will consist of a piecewise constant function, plus a noise term.

How can the difference temperature series be evaluated against this description? The detection of changes in offset in a noisy time series is well established, and is referred to as 'change point analysis'. change point analysis is a key component of temperature homogenization calculations, and so will be illustrated by a synthetic example.

Figure 4.1 shows a synthetic time series consisting of a noise signal, and possibly a piecewise constant signal. Are there any changes in the underlying signal? If so, how many changes are there, when do they occur, and how big are the changes?

Figure 4.2 shows the results of applying a simple change point analysis to the time series. The

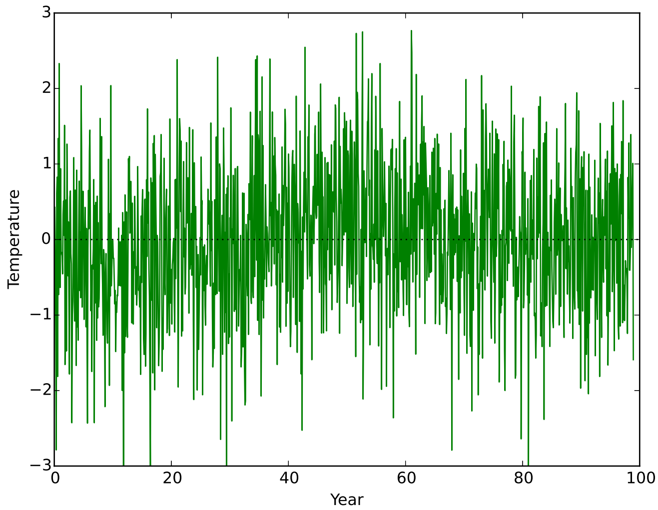


Figure 4.1: A synthetic time series consisting of a noise signal, and possibly a piecewise constant signal.

analysis identifies two change points, at around 33 and 66 years respectively. Furthermore, these change points are each identified with a confidence of at least 99.95%.

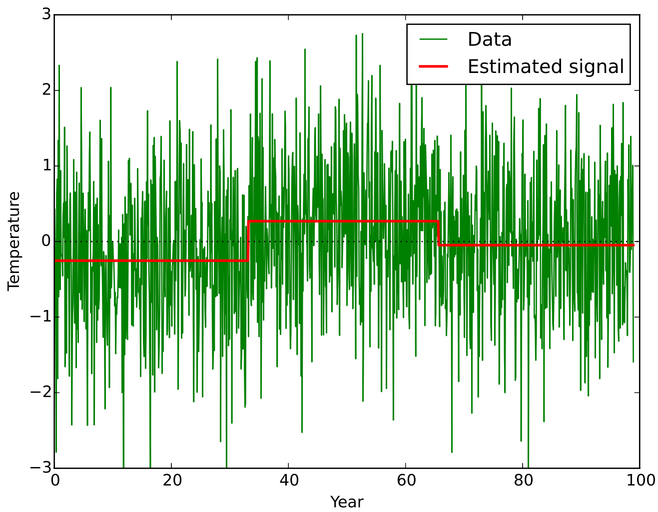


Figure 4.2: Change points detected by change point analysis in the synthetic series.

The effectiveness of the analysis may be determined in this case because the nature of the synthetic data is known. The data were indeed generated from a piecewise constant signal and a noise term, shown in Figure 4.3. The signal extracted by change point analysis is a good match for the real signal.

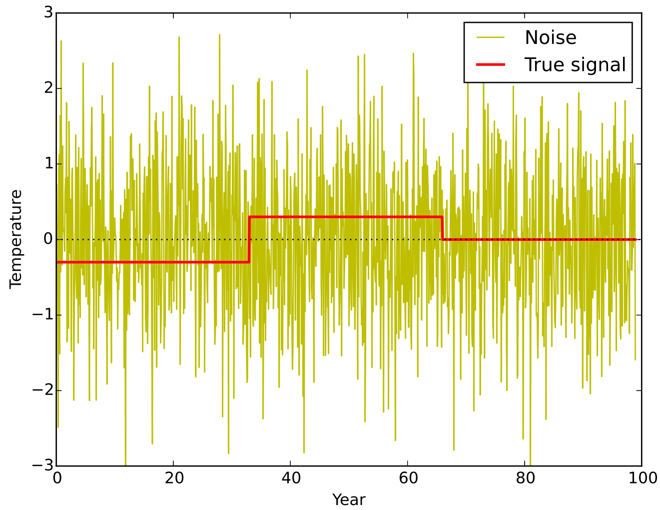


Figure 4.3: The original signal and noise components of the synthetic time series. (For real data, the signal will be the homogenization differences between two weather stations, and the noise will be other differences.)

4.1 Method

How does the change point analysis work? There are a number of methods which can be applied to the detection of change points. In this case a simple approach based on cumulative sums was used (Taylor, 2000).

The cumulative sum of a time series is another series, each element of which is the sum of all of the elements of the time series up to that point in time. The cumulative sum of a noise series is determined by random walk statistics, and wanders around zero in a range which varies in proportion to the square root of the number of terms in the series. However the cumulative sum of a constant function is equal to the constant multiplied by the number of terms in the series. As the number of terms in the series increases, the constant term rapidly overwhelms the noise contribution.

(Before calculating the cumulative sum, an offset is applied to the original time series to bring the mean to zero. This serves to further reduce the effect of the noise contribution.)

The cumulative sum for the example time series is shown in Figure 4.4. The cumulative sum shows three linear segments corresponding to the three segments of the original signal. The ends of the linear segments correspond to the change points in the signal. The change points (marked by the vertical green lines) are clear.

Can we estimate how confident we can be that a change point is present? And how do we distinguish a real change point from a random fluctuation? Both of these are achieved by a statistical approach called bootstrapping. A large number of synthetic time series are generated using random selections of values from the real time series. The cumulative sum calculation is applied to each synthetic series in turn, and plotted on the same graph, as shown by the grey lines in the top panel of Figure 4.5. This gives an indication of how likely it is that the observed change point is a result of noise. If we generate 1000 synthetic time series and the change point lies outside the range of any of them, it suggests that we can be around 99.9% confident that the change point is real. (In practice the probabilities are calculated by a slightly more sophisticated method. The confidence values may be overestimated in the presence of autocorrelation.)

This approach identifies the first change point as real with high confidence. But the second

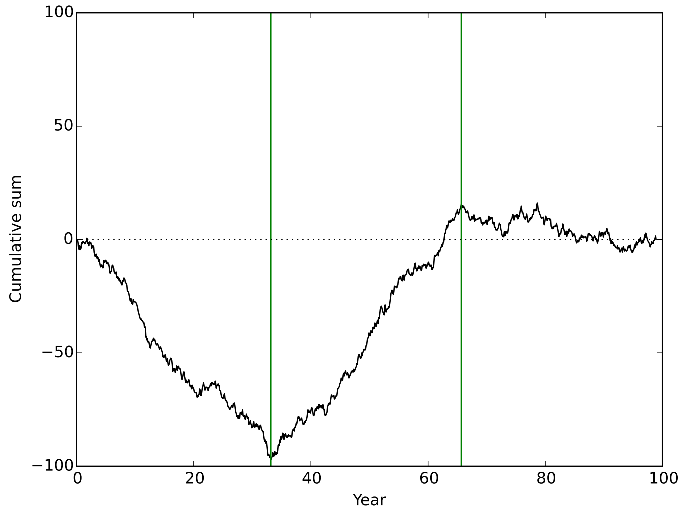


Figure 4.4: The cumulative sum of the example time series.

change point, while clearly visible, lies in the middle of the pack of synthetic series. In order to detect further change points, the original time series is split in two at the first change point, and the whole method is repeated for the two segments, shown by the lower panels of Figure 4.5. No change points are found in the left hand segment, while the second change point is clear in the right hand segment.

Cumulative sums provide a very simple method of change point detection. There are more rigorous methods which can provide better results, especially in the case of small changes in noisy data. However for the purposes of this report the cumulative sum method offers two benefits:

1. It is very simple to implement, enabling more people to reproduce the results for themselves.
2. It is easy to visualise, allowing the evidence for a particular homogenization to be clearly seen.

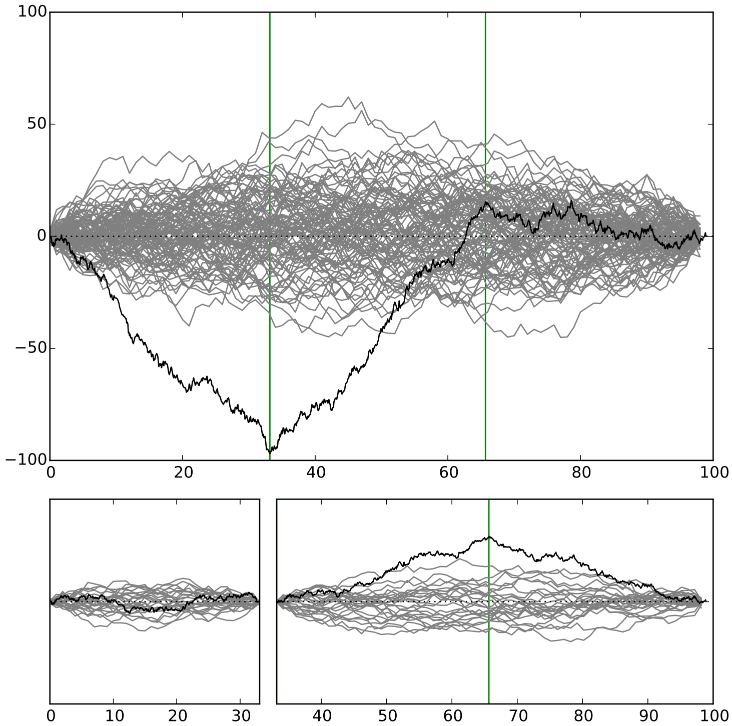
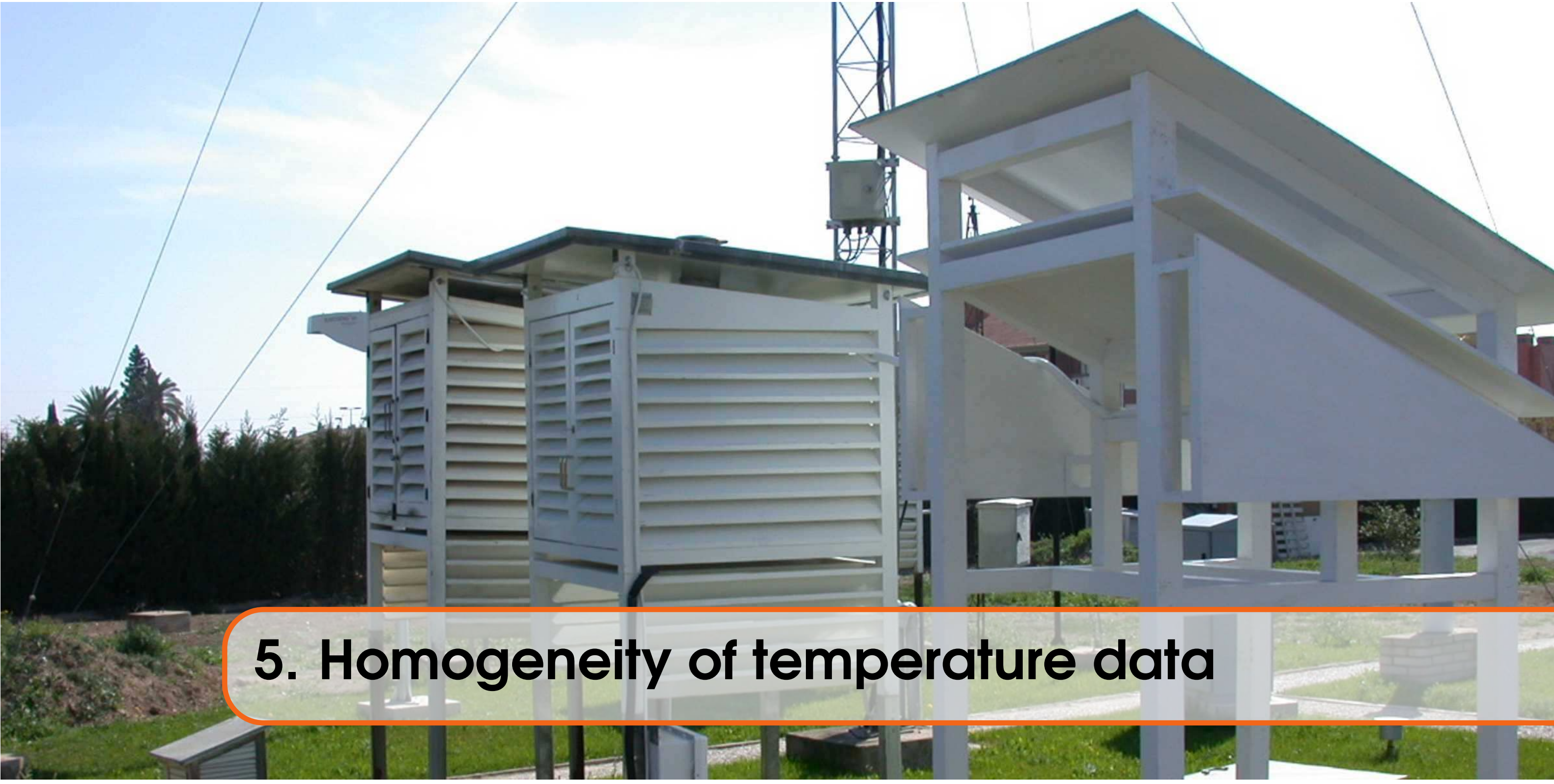


Figure 4.5: Iterative use of cumulative sums and bootstrapping to identify multiple breakpoints. The first step (top) identifies a change point at 33 years. The same process is applied to the two halves of the time series (bottom). No further change points are found in the first half of the data, however a further change point is found in the second half. (In this and subsequent figures a subset of bootstrap series are shown to limit the size of this file.)



5. Homogeneity of temperature data

If change point analysis is to be applied to the homogenization of temperature data, then it must first be established that the weather station data do indeed contain inhomogeneities which can be corrected by the application of constant offsets to segments of the data. This requires that the temperature data from nearby stations are sufficiently similar to enable the differences between the station records to be used to detect inhomogeneities in the individual records.

Two hypotheses must therefore be tested:

1. For a typical station, there exist neighbouring stations for which there are segments of temperature which show strong temperature agreement with the selected station.
2. The differences between the neighbouring stations are well explained by a constant offset with sporadic detectable changes in the value of that offset.

In order to test these hypotheses, a survey was made of the weather station records from the Global Historical Climatology Network (GHCN), using the version 3 data downloaded on 2015-02-03. GHCN provide data for 7200 stations, and both the raw and homogenized data are available. Of the 7200 stations, 6000 (including 600 US stations) are homogenized on the basis of similarity to neighbouring stations alone. The remaining 1200 US stations have time-of-observation adjustments applied to them (on the basis of station metadata) before the application of change point analysis.

Stations with at least 50 years of data were selected from the full data. For each of the resulting 2800 station records, a list of neighbouring stations was compiled comprising all stations within 500 km of the original station. The number of neighbouring stations varies according to the density of the station network. A histogram of the number of stations meeting the neighbour criterion for all of the long station records is shown in the upper panel of Figure 5.1.

Next, the existence of segments of temperature data with good agreement between the two stations was tested. For each station, the temperature series was compared against every other station in the neighbour network. The match over the whole record will be affected by inhomogeneities, so each record was divided into overlapping 7 year segments.

The records were considered to show significant agreement if a 7 year segment could be found for which the standard deviation of the difference between the stations was less than 50% of the

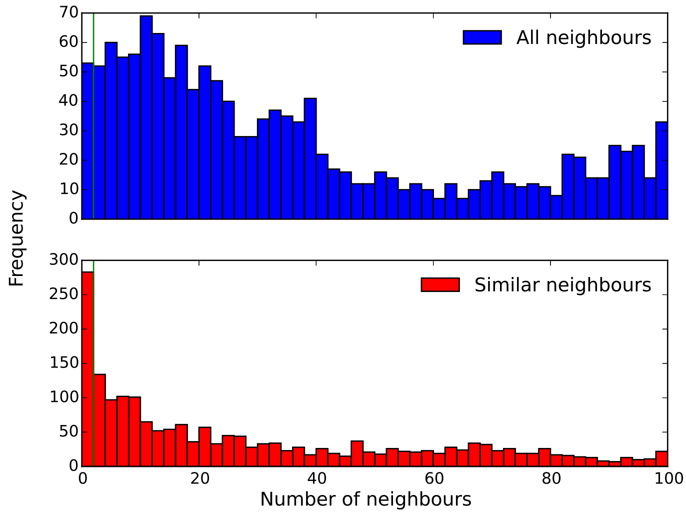


Figure 5.1: Histograms of the number of neighbouring stations for each long record station in the network. The top panel shows the number of stations within 500 km. The bottom shows the number of stations within 500 km which also meet the (strict) similarity criterion.

standard deviation of the original station over the same period. For unrelated series, the standard deviation of the difference series would be expected to be *greater* than the standard deviation of the original series by a factor of $\sqrt{2}$. The level of agreement represented by this criterion is illustrated by example segments from the real data which just achieve this criterion, shown in Figure 5.2.

A histogram of the number of significantly agreeing neighbour stations for each station is shown in the lower panel of Figure 5.1. There exist at least 2 similar neighbouring stations for 90% of the stations (in practice the 50% similarity criterion is probably stronger than required). Therefore the first hypothesis appears to be satisfied for most of the network (and will be further explored by subsequent analysis).

The existence and nature of the inhomogeneities in the station records may be established by examining the differences between agreeing neighbour records. The difference series will show inhomogeneities corresponding to all the inhomogeneities in both records combined, except in the unusual case where two neighbouring stations show synchronized inhomogeneities of equal magnitude.

For each station, an agreeing neighbour was selected using a criterion which rewarded first length of overlap (to favour long difference records), and then the agreement of the best agreeing segment. The cumulative sum calculation was applied to the resulting difference temperature series. A selection of cumulative sum plots, along with the bootstrap series and detected breakpoints, are shown in Figure 5.3.

To test the second hypothesis, that the data contain sporadic change points in the temperature offset, two alternative inhomogeneity models were tested:

1. The constant offset change point model. In this case, the difference between the stations is modelled by a piecewise constant function with the change points determined from the cumulative sum plot as described.
2. A piecewise linear continuous model. In this case, the difference between the stations is modelled by a function consisting of linear segments which join at their endpoints. The model is fitted using the `UnivariateSpline` function from the python `scipy` library, with the number of segments constrained to match the number of segments for the piecewise con-

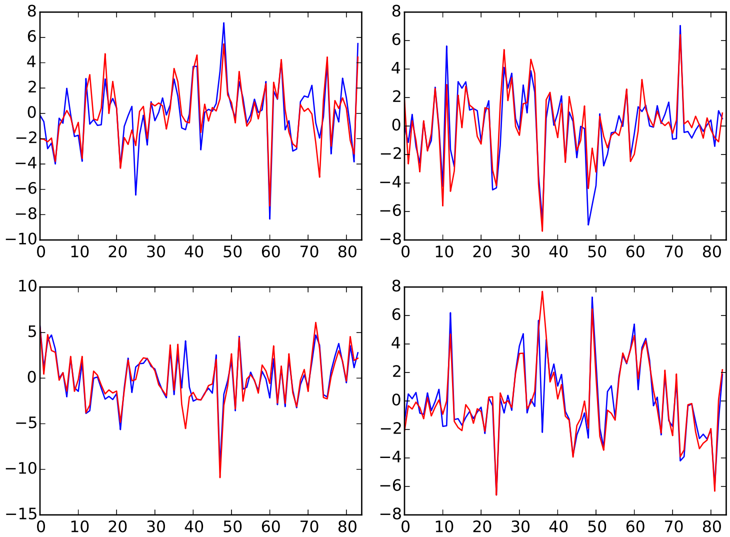


Figure 5.2: Illustration of the station similarity criterion. Four 7 year windows are shown of data which just meet the '50% reduction in standard deviation' criterion. The records show very good agreement suggesting the criterion could be relaxed somewhat, subject to benchmark results.

stant model. This however gives the piecewise linear model an advantage of one additional parameter to fit the differences.

The piecewise linear model is the simplest continuous model, i.e. the simplest model which violates the assumption of discontinuities in offset, although it should still be able to capture some of the effects of a change in offset through the use of multiple linear segments.

The two models were compared on the basis of the log-likelihood gain of the given model over applying a single constant to fit the difference temperature series. The log-likelihood gain for each model was calculated for the best neighbour difference series for each station in turn. The scores for the two models are compared on a station by station basis in Figure 5.4.

The piecewise constant model provides a better description of the differences between neighbouring stations for 99% of the stations. The mean improvement in the log likelihood gain through using the piecewise constant model over all stations is 35.7 (dimensionless). The fit of the two models are illustrated for the same selection of stations as before in Figure 5.5.

The second hypothesis, that the differences between the neighbouring stations are well explained by a constant offset with sporadic detectable changes in the value of the constant offset, is therefore supported. While there might be smaller sources of inhomogeneity which do not fit this model (Willett et al., 2014), changes in offset explain the largest part of the station inhomogeneities.

The continuous segments of the difference temperature series can also be used to estimate the effect of autocorrelation in the differences. Fitting an AR(1) autocorrelation model suggests the effective number of observation is lower than the actual number by a factor of 1.5 to 2.0 times. This reduces the 99.95% confidence criterion to a value around 98-99%.

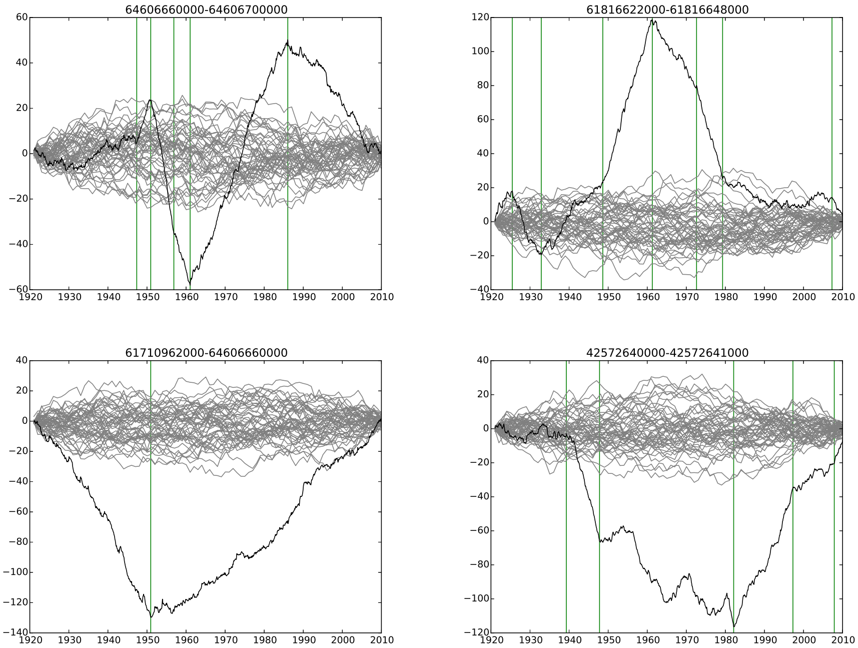


Figure 5.3: Example cumulative sum plots for the temperature difference series between neighbouring similar stations. The station numbers are given in the title of each plot. The black line is the cumulative sum, grey lines are bootstrap series, green bars are change points.

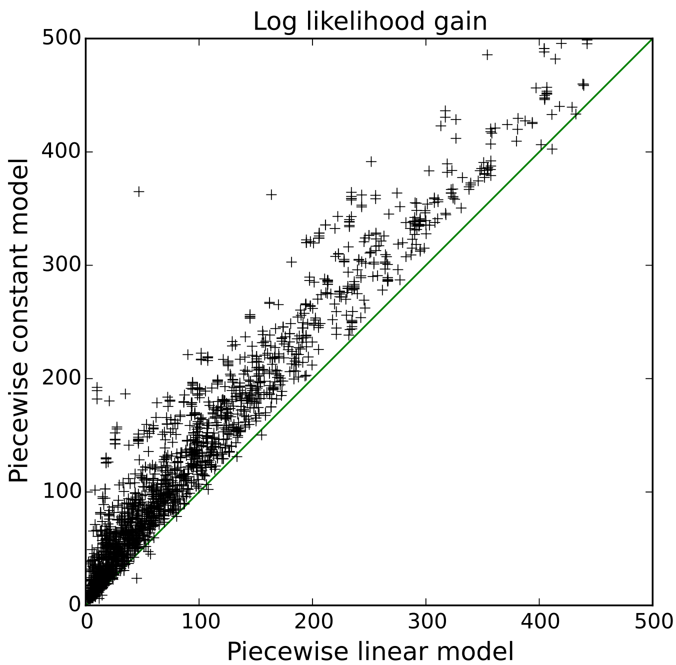


Figure 5.4: Comparison of the piecewise constant and piecewise linear continuous models of temperature inhomogeneities. The coordinates are log likelihood gains over a constant model. Points above the diagonal indicate stations for which the piecewise constant model is better than the piecewise linear model.

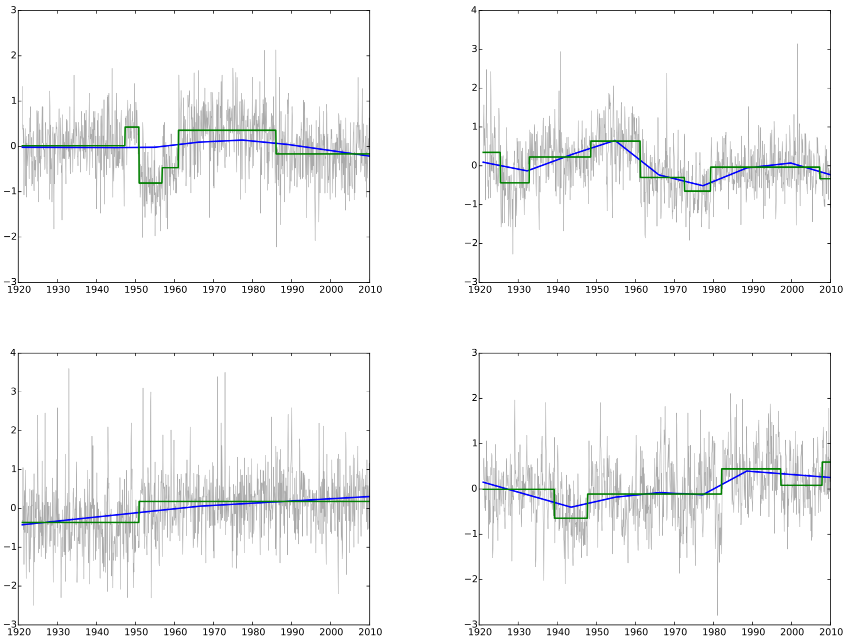
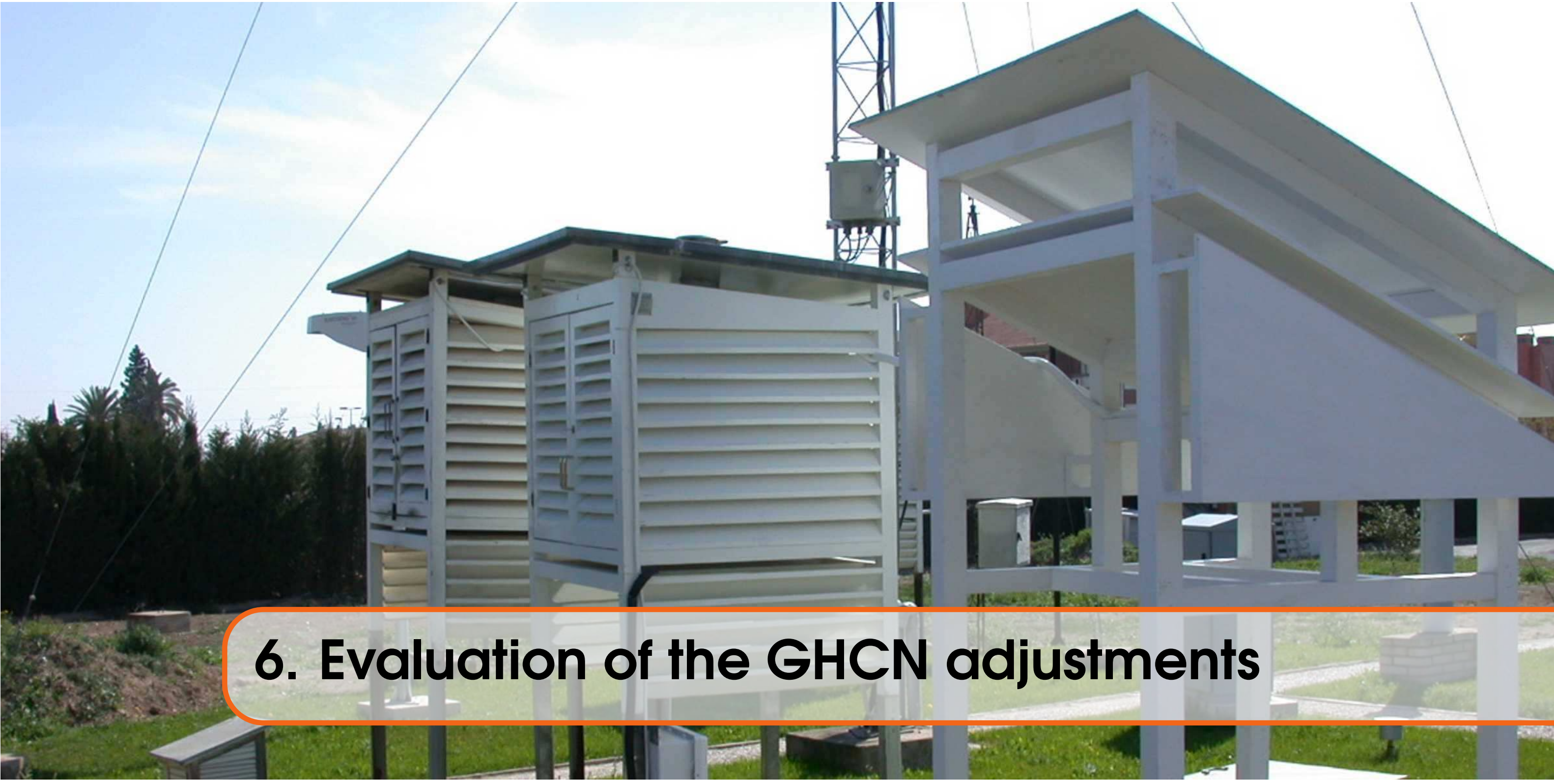


Figure 5.5: Homogeneity model fits for the four example temperature difference series. The grey line shows the temperature difference. The green line is the piecewise constant model. The blue line is the piecewise linear model.



6. Evaluation of the GHCN adjustments

Having established that the data require homogenization and that the largest source of inhomogeneity appears to be changes in temperature offset, the next step is to assess the homogenization adjustments produced by the GHCN adjustment algorithm. The GHCN adjustments are performed by a software package called PHA (standing for Pairwise Homogenization Alignment), available from the following address:

`ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/v3/software/`

An initial assessment of the PHA adjustments was performed by comparing the adjusted station records to the raw records. The data were analysed for the ninety year period 1921-2010, covering the period of greatest trend in the adjustments.

For each station, the raw and adjusted data were compared and the dates of any adjustments to the record determined. 6048 adjustments were identified for which 3 years of data with good completeness and no further adjustments on either side of the adjustment were available. A histogram of the adjustments is shown in Figure 6.1. A gap in the centre of the distribution arises because changes in the instrument which produce almost zero effect on the observations cannot be detected by intrinsic methods. The average size of the adjustments is 0.685°C (root mean square). While adjustments may be positive or negative, the mean adjustment is very slightly positive, with a value of 0.055°C. About 54% of the adjustments are in an upward direction.

Nearby stations within 500 km were then identified with data covering the 6 year window centered on the date of the adjustment. These stations were then sorted according to their agreement with the current station for each half of the 6 year span separately, allowing for a different offset between the stations either side of the adjustment, as well as differences in the annual cycle. The 5 stations with the best match to the current station were selected and averaged to provide a local climatology for the period spanning the adjustment. The two halves of the station record (of three years each) were then fitted to the climatology to determine an estimate for the required adjustment on the basis of the neighbouring stations. This adjustment was then compared to the adjustment applied by the GHCN software.

The adjustments are compared in Figure 6.2 for 5180 adjustments where at least 3 neighbours are available. The two methods show good agreement, with large adjustments being reliably re-

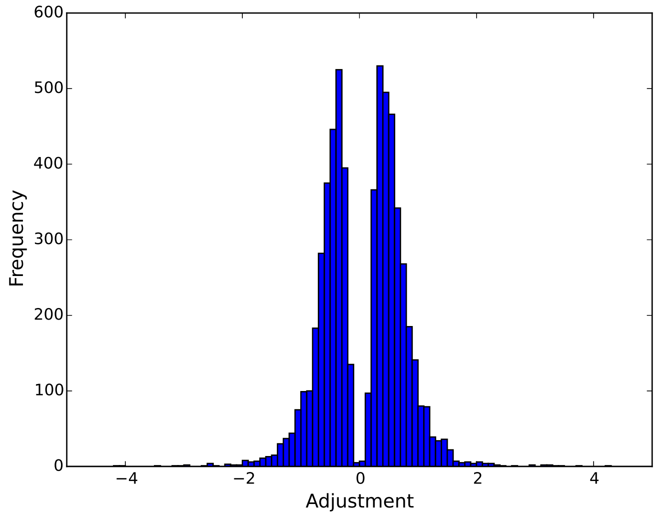


Figure 6.1: Histogram of 5180 temperature adjustments in the GHCN data.

produced. Smaller adjustments show greater uncertainty. The mean adjustment from the 6 year window method is 0.048°C, compared to 0.057 ± 0.010°C for the same adjustment in the GHCN data.

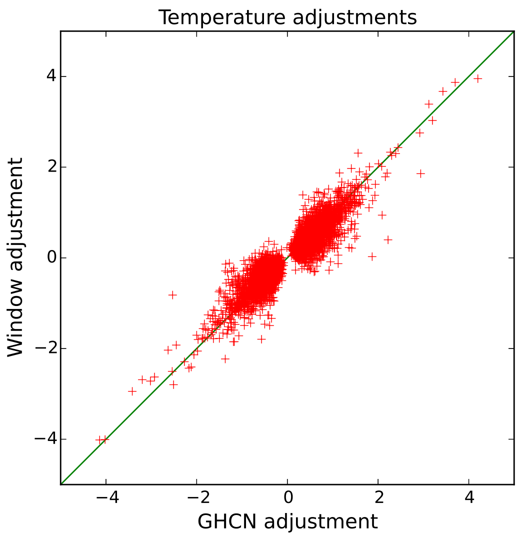


Figure 6.2: Comparison of temperature adjustments from the window method against the corresponding adjustments from GHCN.

The window method has some limitations as a test of the homogenizations. First, the short window used to establish the size of the adjustment reduces the accuracy of the estimate. However a larger window would increase the chance of the window overlapping another inhomogeneity, distorting the results. Secondly, uncertainty in the date of the inhomogeneity tends to bias the window method results. Any error in the date of the inhomogeneity will lead to some observations being placed on the wrong side of the change, which in turn will lead to an underestimation of the size adjustment. The windowed method is more prone to underestimation since it uses less data from either side of the change point. Finally, since like the GHCN method the windowed approach relies on neighbouring stations for homogenization, it is not independent and so it is unsurprising that the methods should give similar results. This result therefore serves primarily to eliminate the possibility that the size of the GHCN adjustments arise from errors in the code.

To obtain an independent estimate of the size of the adjustments, an external source of temperature data is required. Weather model reanalyses allow surface air temperature data to be inferred from a variety of sources other than weather station thermometers, including barometers, wind data, and satellites. Reanalyses have limitations, in particular in the identification of long term climate change, since the temperature reconstructions can be influenced by changes in the observing platforms, and in particular the introduction of new satellites. However the window method only requires that the temperature data be stable over a relatively short period, and so provides a novel method for utilizing reanalysis data for homogenization.

Two reanalyses were used: The Modern-Era Retrospective analysis for Research and Applications (MERRA), and the 20th (TCRv2). MERRA assimilates a variety of satellite data sources, however as a result it is only available for dates from 1979. TCRv2 assimilates data from sea surface temperature observations, and from barometer observations on land, and covers the whole period from 1900 to the present. In neither case are the weather station temperature observations used in the reanalysis, and as a result they provide an independent validation of the observations.

For each adjustment in the GHCN data, the temperature estimates for nearby grid cells in the reanalysis data were examined, again using a 6 year window centered on the adjustment. The grid cell with the best fit to the observations on the two 3 year periods either side of the adjustment (after removing the annual cycle) was selected. The required adjustment was then estimated from the difference between the raw observations and the reanalysis data in the 3 year periods either side of the adjustment.

The post 1979 adjustments are compared to the MERRA reanalysis in Figure 6.3. The estimated adjustments again show good agreement. As before, smaller adjustments show greater uncertainty. The mean adjustment according to the MERRA data should be 0.016°C, compared to 0.021°C for the GHCN data for the same set of 1313 adjustments.

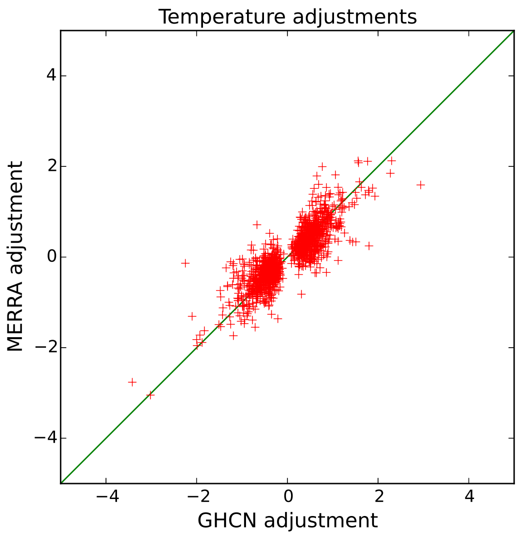


Figure 6.3: Comparison of temperature adjustments from the MERRA reanalysis against the corresponding adjustments from GHCN.

The adjustments for the whole period are compared to the TCRv2 reanalysis in Figure 6.4. The estimated adjustments are less good than for the other methods, presumably due to the limited observational bases, however there is still significant agreement. The mean adjustment according to the TCRv2 data should be 0.053°C, compared to 0.055°C for the GHCN data for the same set of 6048 adjustments.

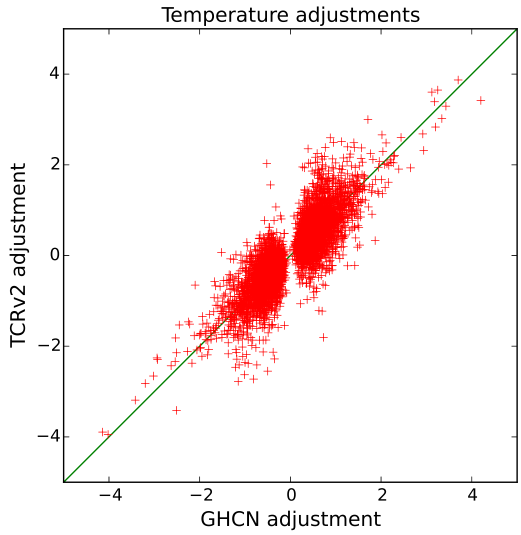


Figure 6.4: Comparison of temperature adjustments from the TCRv2 reanalysis against the corresponding adjustments from GHCN.

Three tests have been applied to the GHCN homogenization adjustments - one based on neighbouring stations as in the GHCN method, and two based on different reanalysis datasets. In each case, the alternative method provides good agreement with the GHCN adjustments. The methods applied here are based on short time windows, which are expected to underestimate the adjustment slightly. Despite this limitation, the results not only provide good agreement with the GHCN adjustments, they also confirm the existence of a trend in the adjustments: the reanalyses support the result that there should be an asymmetry in the distribution of adjustments.

This result assumes however that the GHCN adjustments are in the right places. It is possible that a systematic error in the selection of adjustment positions could lead to a bias in the adjustment directions. To test this possibility, it will be necessary to identify the adjustments from scratch. This will be the subject of the remainder of the report.

7. Fragment homogenization alignment

In order to determine whether the trend in the temperature adjustments is a property of the data, or of the algorithm, a new homogenization algorithm will be developed. If the new algorithm also shows the same behaviour, this will provide additional evidence that the trend is a property of the data rather than the algorithm.

The primary design goal for the new algorithm is that is should be as simple as possible whilst still demonstrating significant skill in the removal of inhomogeneities from benchmark data. The reasons for choosing this goal are as follows:

1. If the algorithm is sufficiently simple, it becomes possible for citizen scientists to replicate the method with only a modest level of skill and investment of time. The requirement that a result be replicable is key to the scientific method, and it is beneficial if results can be reproduced not just by rerunning the same software, but also by re-implementing it.
2. The provision of a simple software framework for the homogenization of temperature data lowers the bar of entry for investigators who wish to go beyond replication and develop their own, more advanced methods.
3. It is generally easier to identify possible causes of bias in simple algorithms than in complex ones. As complexity increases, the likelihood of unforeseen interactions also increases.

It will be shown that a complete software package for the homogenization of global temperature data can be implemented in under 150 lines of python code, which are included in this report.

The method adopted here is called 'Fragment Homogenization Alignment' (FHA). The method takes inspiration from both the 'scalpel' method used in the Berkeley Earth homogenization, and from genomic sequencing methods involving the splitting of DNA sequences into fragments. These ideas will be combined with the concept of similar neighbouring stations developed in the preceding sections of this report.

7.1 The Fragment Homogenization Alignment algorithm

The Fragment Homogenization Alignment algorithm is performed on a station-by-station basis. For each station, two steps are performed: firstly a local climatology is constructed on the basis of information from a network of stations which show significant similarity to the current station, and secondly, change point adjustments are applied to the current station to bring it into agreement with the local climatology. All calculations are performed against the raw data, so adjustments applied to one station do not influence the homogenization of subsequent stations. The two steps in the algorithm will be considered in turn.

7.2 Construction of local climatology

The local climatology for a station is constructed as follows:

1. A network of significantly similar stations in the neighbourhood of the current station is identified using the similarity criterion described previously. The size of the network is limited to the 15 best stations based on the length of overlap and similarity to the current station, in order to moderate the computation time.
2. A pairwise comparison is performed between each station in the local network and every other station (a maximum of 105 comparisons) using the cumulative sum method described previously to identify any change points in the difference series.
3. The temperature series for *both* stations are split at every change point in the difference series, to produce a set of fragments which should be free from inhomogeneities. These fragments are added to a library of fragments including every fragment from every pairwise comparison from the current network.
4. All the fragments from the library are combined into a single climatology by determining the offset which will give the best agreement between that fragment and the mean of all the fragments. The offsets are refined iteratively. (This is simpler and quicker than a least squares implementation of the same algorithm.)

The fragments determined from any individual pairwise comparison will typically be shorter than the unbroken segments in an individual record, because the pairwise comparison produces change points anywhere where there is a break in *either* record. The fragments from a single pairwise comparison do not overlap, and so are useless on their own. However different pairwise comparisons yield different change points. As long as there is at least one pair of stations in the neighbourhood neither of which contain a change point over a given period, a fragment will be available spanning that period.

It is possible (but rare) for there to be a part of the record for which there are no fragments in the library spanning that period. In this case the local climatology will be divided into two separate periods, which a break of unknown offset between them. In this case only the longest continuous part of the climatology is retained.

7.3 Determination of the station adjustments

Given that the climatology is constructed from the consensus of many neighbouring stations it is assumed to be more reliable than any individual station record. The local climatology is therefore used to correct the inhomogeneities in the current station record. The same change point calculation is applied to the difference between the current station and the local climatology, producing a list of change points. The station record is split at each of the change points, and the offsets required to match each segment to the climatology are determined. Each segment is then corrected by application of the offset, to produce an estimate of the homogenized record for that station.

If there are any parts of the current station record for which no local climatology could be constructed, those parts of the record are dropped. The homogenized record produced by this method is thus slightly less complete than the raw data.

7.4 An illustration of the FHA algorithm

The FHA algorithm will be illustrated by a simple example. Figure 7.1 shows a network of 4 neighbouring stations. The stations records each consist of a temperature signal, simulated inhomogeneity in the form of sporadic changes in temperature offset, and a noise signal specific to that station. For ease of visualisation the inhomogeneities have been made very large, while the temperature signal and noise are both very small. (More realistic data will be examined later.)

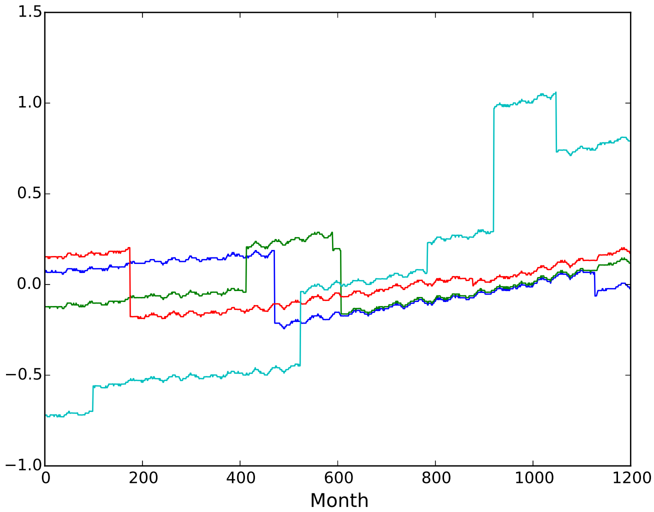


Figure 7.1: Example network of 4 stations with a small temperature signal and large inhomogeneities.

The first two stations selected, shown in Figure 7.2a, each have 2 change points (although one is quite small). The difference series therefore shows 4 change points (Figure 7.2b). The change points are used to split each of the two stations into 5 fragments, which are added to the fragment library (Figure 7.2c).

The process is repeated for every pair of stations in the network. This gives a final fragment library of around 50 fragments (Figure 7.3).

Offsets are then determined to optimally merge the fragments into a single temperature series, essentially by joining overlapping fragments so that the overlapped regions agree (Figure 7.4) which is the local climatology series. Note that the original temperature series has been correctly recovered, apart from some spurious spikes at change points due to off-by-one errors in identification of the month of the change: this is commonly addressed by refining the change points after the cumulative sum step, however for real data the noise is such that the change points are unlikely to be exact even with refinement.

Finally, the change points for the difference between the first station and the climatology are determined. The fragments of the station record are fitted to the climatology to determine the temperature offset for each fragment.

7.5 Limitations

The primary design criterion for the FHA algorithm was simplicity, with the aim of making the method easily reproducible. As a consequence the method has a number of limitations. The cumulative sum method, while easy to understand, is less effective than more recent likelihood-based methods. In addition, the method is dependent on several parameters.

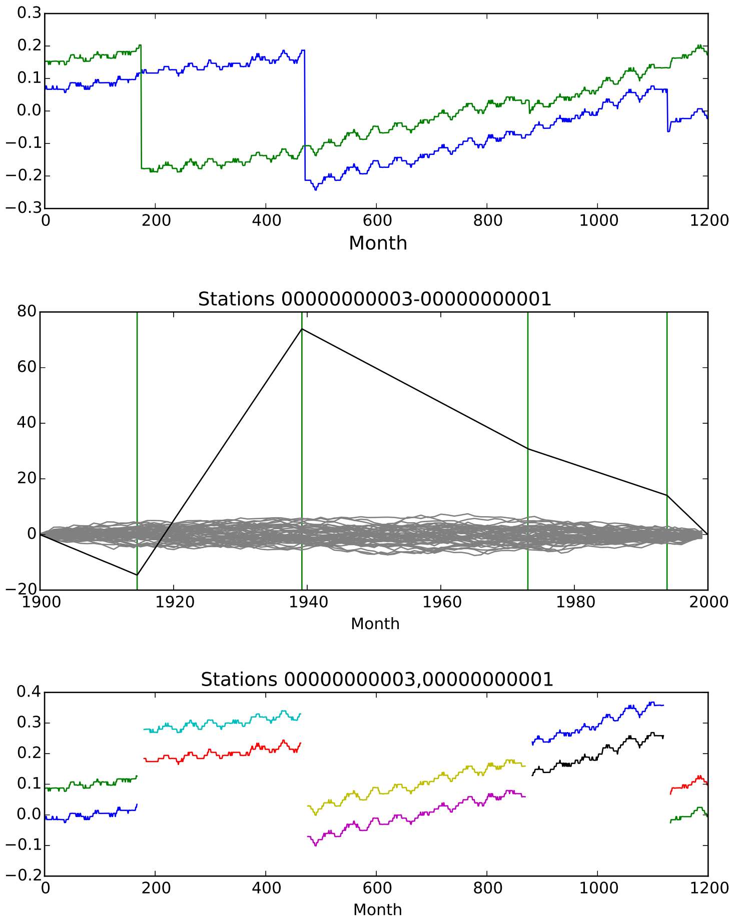


Figure 7.2: Compiling the FHA fragment library. (a) Two stations from the network are compared. (b) Change points in the difference temperature series are identified. (c) Both stations are split into fragments at each change point, and the fragments are added to the library.

The most significant parameters are:

- The threshold for introducing a break (set at 3.5σ, corresponding to a 99.95% confidence level or 98-99% after including autocorrelation).
- The station similarity threshold (set at 50% of the standard deviation for the most similar 7 year window).
- The number of stations which will contribute to the fragment library for homogenizing a given station (set at 15, to limit the computational overhead).

These parameters were set on an ad-hoc basis from inspection of the data and intermediate results. The similarity threshold is particularly important: if this is set too low then too few similar stations will be found to construct the fragment library. If it is set too high then the resulting difference temperature series will be too noisy to allow the detection of change points. Likely areas for improvement include the use of a better change point method, and station-dependent criteria for selecting a local station network.

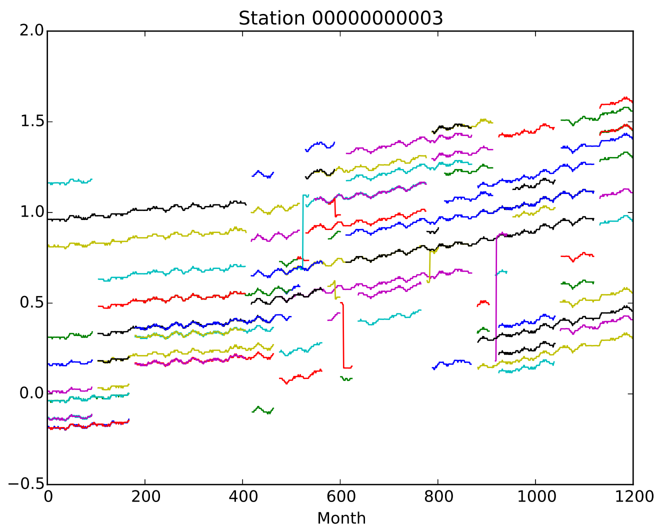


Figure 7.3: Complete fragment library for the network of 4 stations.

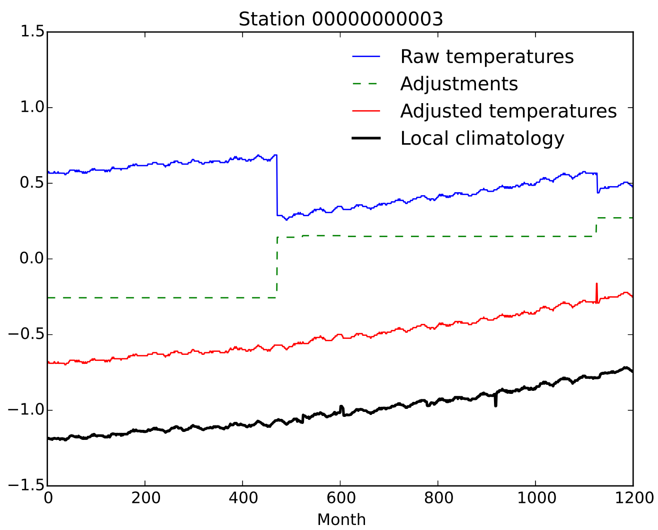


Figure 7.4: Correcting a station to match the local climatology determined from the fragment library.

8. Benchmarking FHA

Benchmarking, in which an algorithm is tested on real or synthetic data for which the correct answer is known, forms a critical tool for the evaluation of any data processing algorithm. A true temperature series is generated by some means, and then homogenization errors are added to the data. The corrupted data are presented to the homogenization algorithm, which is then evaluated on its ability to reconstruct the true signal. Two major benchmarking studies have produced data sets against which the FHA algorithm will be evaluated:

1. Venema et al. (2012) tested a range of homgenization algorithms on small networks of weather stations in Austria, France and Spain. Test data were generated with the same statistical properties as the original station data, and random homogenization errors (both instantaneous and trend-like changes) were add to the data, both on a random basis and in clusters.
2. Williams, Menne, and Thorne (2012) tested the PHA algorithm used by the GHCN dataset against 8 synthetic datasets representing the contiguous United States. The 'true' temperature signals were derived from climate model outputs for the period 1900-1999. The size and frequency of the inhomogeneities varied between the 8 datasets, as did the presence of metadata concerning possible change points (e.g. due to station moves). Some of the datasets included inhomogeneities with a non-zero mean, thus introducing a trend in the adjustments to the data.

The Venema et al. (2012) benchmarks involve small and comparatively sparse networks of stations, while the Williams, Menne, and Thorne (2012) benchmarks involve densely spaced stations with a lot of problems.

The FHA algorithm was first tested against the Venema et al. (2012) data. The algorithm was able to construct a local climatology and thus determine homogenization adjustments for 77% of the data (or 85% if the similarity criterion is relaxed). The FHA adjusted data were compared to the true data, as were the corresponding data from the subset of methods in Venema et al. (2012) which were also tested against all of the stations. The algorithms were scored on the basis of the correlation coefficient between the homogenized and true data for each station, and an overall score determined from the mean of the correlation over all stations in all of the networks. (Correlation is not an ideal score for this purpose - Venema et al. suggest better metrics - but it is simple to evaluate on the grounds that the bounds are zero and one.)

The correlations for the raw data and different homogenized datasets are given in Table 8.1. The FHA algorithm produced a clear improvement over the raw data. The results were also comparable to the other methods tested, producing a slightly higher correlation than the Climatol and AnClim algorithms, but a lower correlation than MASH and ACMANT.

Dataset	**Correlation to true temperatures**
Raw data	0.933
USHCN 52x	0.974
USHCN main	0.971
USHCN cx8	0.971
Climatol	0.938
MASH main	0.983
ACMANT	0.982
AnClim main	0.943
Climatol2.1a	0.974
Climatol2.1b	0.968
ACMANT late	0.990
FHA (this study)	0.978

Table 8.1: Correlation coefficients between homogenized data using different algorithms and the true data for the Venema et al. (2012) benchmark data.

The FHA algorithm was then tested against the 8 different datasets from Williams, Menne, and Thorne (2012). These datasets differ in the source of the temperature data and the size and frequency of the inhomogeneities; their properties are summarised in Table 8.2. Only station records for which more than 50 years of data were present were homogenized, although all of the stations were available for use in neighbour networks. For one dataset ('world 5') the similarity criterion had to be relaxed due to larger than expected inter-station differences.

Dataset	**Description**
World 1	Temperatures from MIROC 3.2. Breaks, some clustered, some with metadata, some with a negative trend.
World 2	Temperatures from CSIRO 3.5. Breaks as world 1.
World 3	Temperatures from HadGEM1. Breaks as world 1.
World 4	Temperatures from CCSM 3.0. Breaks as world 1.
World 5	Temperatures from GFDL CM2. Big breaks with good metadata, no trend in the breaks.
World 6	Temperatures from NCAR PCM. Many small breaks, some with metadata, some with a trend.
World 7	Temperatures from MIROC 3.2 hires. Breaks, some with metadata, some false breaks in the metadata.
World 8	Temperatures from MIROC 3.2 hires. No breaks (i.e. perfect data).

Table 8.2: Descriptions of the different datasets in the Williams, Menne, and Thorne (2012) benchmark data. From Zeke Hausfather.

The correlations for the raw and FHA adjusted data, based on only those data for which an adjustment was performed, are given in Table 8.3. The FHA results are almost perfect for every dataset with the exception of ‘World 6’. This dataset is especially challenging because the inhomogeneities are both small and frequent, making them hard to detect, however even in this case the adjusted data are closer to the truth than the raw data.

Benchmark	Raw correlation	FHA correlation
World 1	0.893	0.993
World 2	0.907	0.992
World 3	0.924	0.992
World 4	0.929	0.993
World 5	0.912	0.995
World 6	0.934	0.984
World 7	0.957	0.994
World 8	1.000	1.000

Table 8.3: Correlation coefficients of the raw and FHA homogenized data with the true data for the different benchmark datasets from Williams, Menne, and Thorne (2012).

Of particular interest is the ability of the homogenization algorithm to detect a trend in the inhomogeneities, given that such a trend exists in the adjustments to the real world data. Area weighted ‘global’ temperature series were calculated for the raw, adjusted and true data using only those data for which adjustments were determined. The trend on the whole period 1900-1999 was determined for each dataset. The results are shown in Figure 8.1.

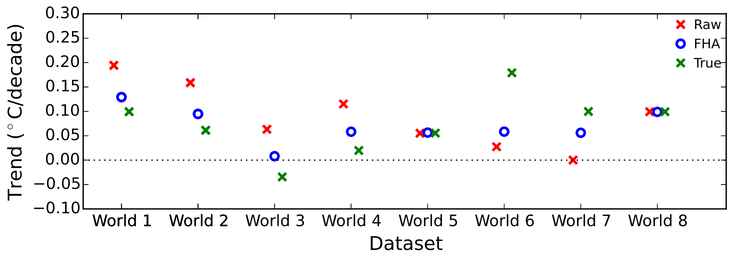


Figure 8.1: Temperature trends for the period 1900-1999 for the 8 benchmark datasets (‘worlds’). Red crosses indicate the trend in the raw data. Blue circles indicate the trend in the FHA homogenized data. Green crosses indicate the true trends.

While the FHA algorithm always corrects the trend in the direction of the true data, it only recovers part of the trend in the adjustments, with the portion recovered ranging from 69% for world 1 to 20% in the case of the challenging world 6. By contrast, the PHA algorithm of Menne and Williams Jr (2009) recovers almost all of the trend in the adjustments in the same tests, or all of the trend if metadata are also used.

It is interesting that FHA can almost perfectly reconstruct the true data in every case except world 6 as measured by correlation, and yet it typically recovers only part of the trend in the adjustments. This highlights the fact that the trend in the adjustments is tiny compared to the individual adjustments. Furthermore, the trend can only arise from the long term cumulative effect of multiple adjustments which are determined from shorter term changes in the records - in other words it is a kind of ‘dead reckoning’ problem.

Why does the FHA algorithm underestimate the change in trend? Two factors are suggested for further investigation. Firstly, the cumulative sum algorithm is insensitive to small changes compared to more advanced (e.g. likelihood based) methods. Secondly, the fragment method produces a large number of unnecessarily short fragments: it is possible that this lack of longer fragments weakens the detection of longer term trends in the adjustments. Both of these possibilities are testable.

9. Application of FHA to real world data

The FHA algorithm was next applied to the real world GHCN data, and the results compared against the results of the Menne and Williams Jr (2009) PHA algorithm, using the same metrics as in the benchmark. All data, including those with time-of-observation adjustments, were used in the test. As with previous tests the time period was limited to 1921-2010, no use was made of metadata, and homogenizations were calculated only for stations with long records.

The global mean trends for the period 1921-2010 using the raw data and the homogenized data from each method are compared in Table 9.1. The FHA adjustment to the trend is 41% of that determined by PHA. This is consistent with the benchmark result that FHA only detects 20-70% of the true trend in the inhomogeneity of the data. The FHA result may also be affected by working with a shorter run of data, since it is harder to detect inhomogeneities at the ends of the data.

Data	**Trend (°C/decade**
Raw	0.091
FHA homogenized	0.102
PHA homogenized	0.117

Table 9.1: Global mean surface temperature trends over the period 1921-2010 for the raw data, and for homogenized data from the FHA and PHA methods.

Global temperature series were calculated for the raw and homogenized data using a method similar to Jones et al. (2012), and used to calculate the impact of the adjustments on global land temperature estimates. The results are shown in Figure 9.1. The adjustments from the FHA algorithm are broadly similar in shape but underestimated in magnitude, compared to PHA as would be expected from the benchmark results.

Trend difference maps for the two homogenization methods are shown in Figure 9.2. The geographical distribution of trend adjustments shows similarity between the two methods. FHA supports the upwards adjustment of the southeastern US data, although by a smaller amount, presumably due to lack of metadata and insensitivity to cumulative small changes. FHA supports the upward adjustments to the Paraguay record, but rejects the upward adjustments to the Icelandic

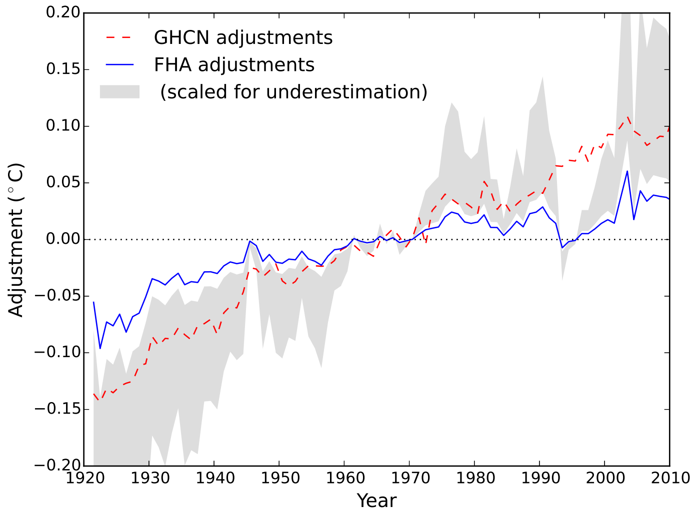


Figure 9.1: Impact of the temperature adjustments on global mean surface temperatures using adjustments determined from the PHA and FHA algorithms. The shaded region shows the effect of scaling the FHA adjustments to correct for the underestimation of trend using the range of scale factors determined from the Williams, Menne, and Thorne (2012) benchmark.

record. FHA introduces many localised adjustments in central Asia which are not supported by PHA: this is currently unexplained, however the absence of any outlier rejection in the FHA calculation may play a role. Outside of Asia local patterns of adjustments are generally similar between the two methods but with small differences in the regional trend.

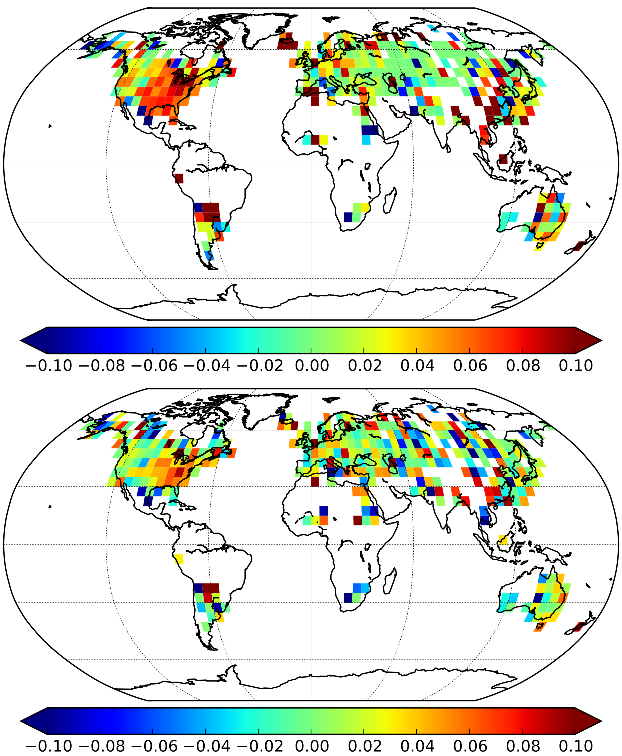


Figure 9.2: Geographical distribution of the trends in the adjustments for the PHA adjustments (top panel) and the FHA adjustments (bottom panel).

10. Testing for bias

The FHA algorithm is symmetrical in its response to increasing or decreasing temperatures, and so by definition any trend in the adjustment should be a property of the data rather than the method (subject to small variations arising from the random bootstrap step). While the simplicity of the algorithm also reduces the likelihood of bias, even simple algorithms can be subject to unforeseen behaviours. It is therefore prudent to perform explicit tests for any bias in the trend in the adjustments, particularly in the case of the much more complex PHA algorithm of Menne and Williams Jr (2009).

Two hypothetical biases which would create a spurious trend in the adjustments will be tested:

1. That the method systematically favours adjustments in one direction over the other.
2. That the method systematically favours adjustments which amplify the regional trend.

In each case, the hypothesis will be tested by application of the PHA homogenization algorithm to two sets of temperature data and comparing the results. The temperature data will be either the raw observed temperature data, or a modification of the raw temperature data which will lead to a predictable change in the results for a given hypothesis. If that change is not observed, then the hypothesis is falsified. Each hypothesis will be considered in turn.

10.1 Hypothesis 1: That the trend in the adjustments arises from the method systematically favouring adjustments in one direction over the other.

The implication of this hypothesis is that the homogenization algorithm will produce adjustments with a positive trend for any input data. The hypothesis can be tested very simply by inverting the temperature record. If the hypothesis is correct, the algorithm will produce adjustments with a positive trend for the inverted (i.e. cooling) record as well. If however the trend in the adjustments is a property of the data, then inverting the data will also invert the adjustments.

The individual station records were inverted by determining the mean temperature for that station, and inverting the station data about that mean. Temperature records were calculated from both the original and inverted station records, both before and after homogenization. US stations were excluded to guarantee that no time-of-observation adjustments were present. The results are shown in Figure 10.1.

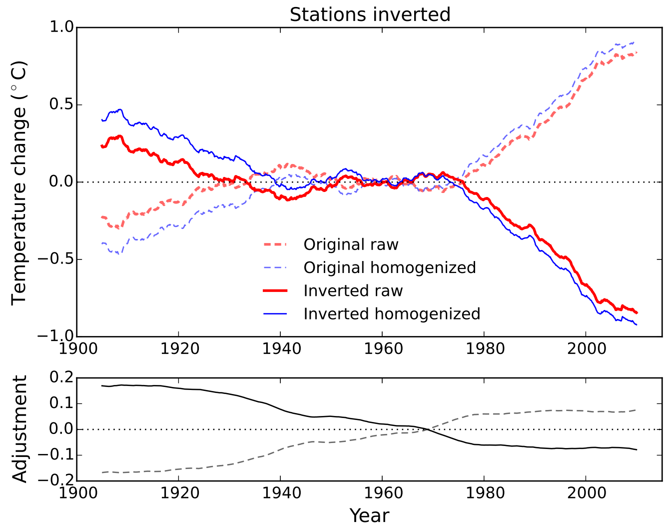


Figure 10.1: Comparison of the impact of homogenization on both the original and inverted station records.

Inverting the original station records also inverts the direction of the homogenization adjustments. There is no evidence that the PHA method favours adjustments in one direction over the other on a scale which is significant compared to the global adjustments. The hypothesis is therefore falsified.

10.2 Hypothesis 2: That the trend in the adjustments arises from the method systematically favouring adjustments which amplify the regional trend.

The implication of this hypothesis is that the homogenization algorithm will produce adjustments which tend to be in the same direction as the regional trend in the input data. The hypothesis can be tested by subtracting the regional temperature trend from every station record, in order to produce a record which has no regional trend. Alternatively, twice the regional trend can be subtracted from each record to produce a record in which the regional and global trends are inverted, while leaving the local month on month variations largely intact.

The second approach is adopted here. The regional trend is determined using a per-gridcell lowess smooth of a gridded spatially smoothed temperature record (Hansen, Ruedy, Sato, & Lo, 2010), with a smoothing parameter of $1/3$. The smoothed regional temperature record for the gridcell is scaled by a factor of -2 and added to each station record in that grid cell. Temperature records were calculated from both the original and inverted station records, both before and after homogenization. The results are shown in Figure 10.2.

Inverting the trends in the station records, while maintaining the month-on-month variation, does not affect the direction of the homogenization adjustments. Some differences in the adjustments are expected as geographically variable changes to the station records push individual stations over or under adjustment thresholds. However there is no evidence that the PHA method favours adjustments which reinforce the regional trend on a scale which is significant compared to the global adjustments. The hypothesis is therefore falsified.

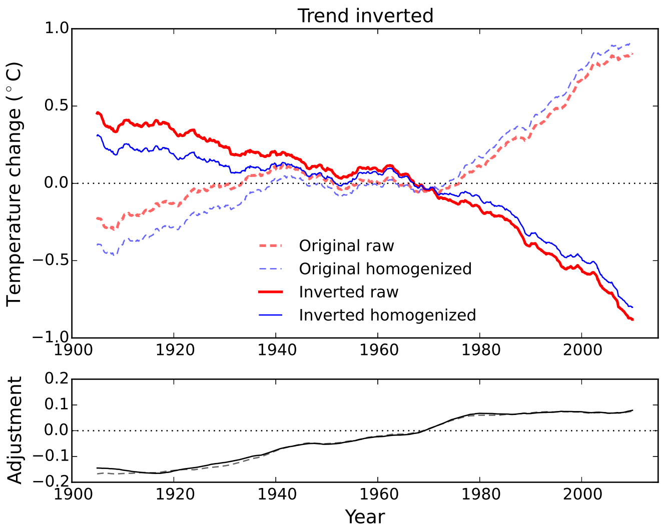


Figure 10.2: Comparison of the impact of homogenization on both the original records, and records in which the smoothed regional trend has been inverted.

11. Conclusion

Arthur C. Clarke famously noted that "any sufficiently advanced technology is indistinguishable from magic". In many cases the same law applies to science: the specifics of most sub-disciplines are sufficiently opaque to an outsider (whether a lay person or a scientist from another field) that they may as well be magic.

As a result, we often rely on 'fast-thinking' strategies to evaluate the validity of others' work (Kahneman, 2011). Criteria vary in their objectivity, from how often the results have been reproduced, to whether the results are threatening to the reader's worldview. The use of fast thinking strategies is inevitable, since the alternative is to reproduce every result for ourselves. However that does not mean that such strategies give the right answer.

A few years back I saw a number of examples on blogs and web forums of claims that James Hansen sat in his office each month trying to decide what number GISTEMP would report for that month. I don't see those claims very often any more (Hansen's retirement notwithstanding). I suspect that the work of citizen scientists like Nick Barnes, Nick Stokes, Caerbannog and others who have produced their own versions of the temperature record has played a role in this. They have demystified the temperature record, transforming it from something handed down by experts, into something which anyone can produce for themselves.

One of my goals for this work is to do the same for homogenization. The problem is not much more complex than the temperature record calculation, and easily within the reach of many citizen scientists and bloggers. The steps I have taken are rudimentary, and the statistical methods are merely (and in some cases barely) adequate for the job. But having demonstrated that it is possible, I hope that others will go further. The first goal will be to produce a method which recovers more of the trend in the benchmark data. Applying such a method to the observations will tell us more.

To get to this point I've explored a number of questions, and obtained (in some cases tentative) answers to them:

- Are there inhomogeneities in the data?
  Yes, there are.
- Are those inhomogeneities of a form which would be explained by sporadic changes in the

measuring apparatus or protocols?
Yes, the largest inhomogeneities are explained by sporadic changes in offset in the temperature readings.
- Can those inhomogeneities be detected by comparing records from neighbouring stations?
Yes, most stations have other nearby stations with substantially similar records.
- Is there sufficient redundancy in the data to allow those inhomogeneities to be corrected?
Yes, tests using multiple benchmark datasets suggest that inhomogeneities can be corrected.
- Does the Global Historical Climatology Network (GHCN) method produce reasonable estimates of the size of the adjustments?
Yes, both neighbouring stations and reanalysis data support the GHCN adjustments.
- Do the observations support the presence of a trend in the homogenization adjustments?
Yes, both methods suggest that the adjustments should have a slightly skewed distribution.
- Is there evidence that trend in the adjustments could be an artifact of the methods?
Two possible sources of bias in the method were tested and eliminated.
- If the data are correctly homogenized, how large a change will be introduced in the global temperature trend?
The size of the required correction to the global record is much harder to determine than the direction. The simple methods described in this report cannot provide an absolute answer. The most I can say is that the GHCN correction looks plausible.

However the answers are less important than the demonstration that the questions can be meaningfully investigated by sufficiently motivated citizen scientists. I hope that another law of Arthur C. Clarke is relevant:

> Every revolutionary idea - in science, politics, art, or whatever - seems to evoke three stages of reaction. They may be summed up by the phrases:
> 1. “It’s completely impossible - don’t waste my time”;
> 2. “It’s possible, but it’s not worth doing”;
> 3. “I said it was a good idea all along.”

Bibliography

Domonkos, P, R Poza, and D Efthymiadis (2011). “Newest developments of ACMANT”. In: *Advances in Science and Research* 6.1, pages 7–11 (cited on page 8).

Guijarro, José A (2011). “User’s guide to CLIMATOL”. In: *An R contributed package for homogenization of climatological series, Report, State Meteorological Agency, Balearic Islands Office, Spain* (cited on page 8).

Hansen, James et al. (2010). “Global surface temperature change”. In: *Reviews of Geophysics* 48.4 (cited on page 46).

Jones, PD et al. (2012). “Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010”. In: *Journal of Geophysical Research: Atmospheres (1984–2012)* 117.D5 (cited on page 41).

Kahneman, Daniel (2011). *Thinking, fast and slow*. Macmillan (cited on page 49).

Kennedy, JJ et al. (2011). “Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization”. In: *Journal of Geophysical Research: Atmospheres (1984–2012)* 116.D14 (cited on page 9).

Lawrimore, Jay H et al. (2011). “An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3”. In: *Journal of Geophysical Research: Atmospheres (1984–2012)* 116.D19 (cited on page 8).

Menne, Matthew J and Claude N Williams Jr (2009). “Homogenization of temperature series via pairwise comparisons”. In: *Journal of Climate* 22.7, pages 1700–1717 (cited on pages 8, 39, 41, 45).

Mestre, Olivier et al. (2013). “HOMER: a homogenization software–methods and applications”. In: *Időjárás-Quarterly Journal of the Hungarian Meteorological Service* 117.1, pages 47–67 (cited on page 8).

Rohde, Robert et al. (2013). “Berkeley earth temperature averaging process”. In: *Geoinfor. Geostat.: An Overview* 1.2, pages 1–13 (cited on page 8).

Stepanek, P (2005). “AnClim-software for time series analysis”. In: *Dept. of Geography, Fac. of Natural Sciences, Masaryk University, Brno* (cited on page 8).

Szentimrey, T (1999). “Multiple analysis of series for homogenization (MASH)”. In: *Proc. Second Seminar for Homogenization of Surface Climatological Data*, pages 27–46 (cited on page 8).

Taylor, Wayne A (2000). *Change-point analysis: a powerful new tool for detecting changes*. `http://www.variation.com/cpa/tech/changepoint.html` (cited on page 17).

Toreti, Andrea et al. (2012). “A novel approach for the detection of inhomogeneities affecting climate time series”. In: *Journal of Applied Meteorology and Climatology* 51.2, pages 317–326 (cited on page 8).

Venema, Victor KC et al. (2012). “Benchmarking homogenization algorithms for monthly data”. In: *Climate of the Past* 8.1, pages 89–115 (cited on pages 37, 38).

Wang, Xiaolan L et al. (2010). “New techniques for the detection and adjustment of shifts in daily precipitation data series”. In: *Journal of Applied Meteorology and Climatology* 49.12, pages 2416–2436 (cited on page 8).

Willett, K et al. (2014). “A framework for benchmarking of homogenisation algorithm performance on the global scale”. In: *Geoscientific Instrumentation, Methods and Data Systems* 3.2, pages 187–200 (cited on page 23).

Williams, Claude N, Matthew J Menne, and Peter W Thorne (2012). “Benchmarking the performance of pairwise homogenization of surface temperatures in the United States”. In: *Journal of Geophysical Research: Atmospheres (1984–2012)* 117.D5 (cited on pages 37–39, 42).

A. The FHA code

The FHA algorithm is implemented in a python program consisting of 150 lines of code, not counting comments and blank lines.

The change point analysis calculation is performed by the function `changepoint()`:

```
1  # Change point detection using Taylor (2000)
2  def changepoints( dorig, nsig=3.5, nbuf=2 ):
3    if dorig.shape[0] < 3*nbuf: return []
4    dnorm = dorig - numpy.mean(dorig)
5    cusum = numpy.cumsum(dnorm)
6    dboot = numpy.random.choice(dorig, size=(dorig.shape[0],500))
7    dboot -= numpy.mean(dboot, axis=0)
8    cusumb = numpy.cumsum(dboot, axis=0)
9    custd = numpy.std(cusumb, axis=1)
10   ratio = numpy.absolute(cusum)/numpy.maximum(custd,1.0e-30)
11   i = numpy.argmax(ratio[nbuf:-nbuf])+nbuf
12   if ratio[i] <= nsig: return []
13   return ( changepoints(dorig[:i],nsig,nbuf)+
14            [i]+
15            [x+i for x in changepoints(dorig[i:],nsig,nbuf)] )
```

Line 3 checks whether the data are long enough for change point detection. If not, it returns an empty list of changes.

Lines 4-5 calculate the cumulative sum for the input series.

Line 6 generates 500 bootstrap series.

Lines 7-8 calculate cumulative sums for the bootstrap series.

Line 9 calculates the standard deviation of all the bootstrap series for a given time step. This is an optimization: rather than calculating 10,000 series to obtain the confidence level of a change, a smaller number are determined and used to determine a standard deviation under the assumption of normality (which is good for random walk statistics of more than a few steps).

Lines 10-11 identify the time step for which the cumulative sum is most significant compared to the bootstrap distributions.

Line 12 causes an empty list of change points to be returned if no values exceed the sigma threshold.

Line 13 implements a recursive search for further change points. A list is constructed from any change points found before the current change point, the current change point, and any change points found after the current change point. This list is the return value of the function.

A helper function, `changemissing()`, is used for change point detection in data containing missing values. It also removes any annual cycle in the timeseries. Finally it converts the indices of the change points in the complete data back to indices of change points in the original data.

```
# calculate changepoints on data with breaks
def changemissing( dorig, nsig=3.5, nbuf=2 ):
  dnorm = anomaly( dorig )
  mask = ~numpy.isnan(dnorm)
  diff = dnorm[mask]
  chg = changepoints(diff,nsig,nbuf)
  index = numpy.arange(dnorm.shape[0])[mask]
  return [index[i] for i in chg]
```

The third key function merges a fragment library down to a single temperature series by determining optimal offsets to align the fragments.

```
# align and merge a set of fragments
def mergeall( data, cmin=1 ):
  # select only fragments which overlap the longest fragment
  flags = ~numpy.isnan(data)
  imax = numpy.argmax(numpy.sum(flags,axis=1))
  flagc = flags[imax,:]
  mask = numpy.zeros([flags.shape[0]],numpy.bool)
  for c in range(10):
    for i in range(flags.shape[0]):
      if ~mask[i]:
        if numpy.sum(numpy.logical_and(flags[i,:],flagc)) >= cmin:
          mask[i] = True
          flagc = numpy.logical_or(flagc,flags[i,:])
  # now find best offsets for merging
  dsel = data[mask,:]
  for c in range(100):
    davg = numpy.nanmean(dsel,axis=0)
    dmax = 0.0
    for i in range(dsel.shape[0]):
      diff = numpy.nanmean(davg-dsel[i,:])
      dsel[i,:] += diff
      dmax = max( dmax, abs(diff) )
    if dmax < 1.0e-4: break
  # return the result
  return davg
```

This function is divided into two sections. The first section determines whether there are fragments with sufficient overlap to define a single, connected climatology from all of the fragments. Sometimes there is a break in the library with no fragments spanning it. In this case all fragments on the smaller side of the break are discarded. The second section determines the optimum offsets to align the fragments. This is achieved by averaging the fragments, and determining the offset to fit each fragment to the resulting mean. The process is iterated to convergence. A direct least squares solution is also possible, but slower.

The full FHA code is as follows:

```
import sys, math, numpy, scipy.stats

# calculate great circle distance
def distance( la1, ln1, la2, ln2 ):
  a = 0.99999999*( math.sin(math.radians(la1))*math.sin(math.radians(la2)) +
                   math.cos(math.radians(la1))*math.cos(math.radians(la2))*math.cos(math.radians(ln1-ln2)) )
  return 6371.0*math.acos( a )


# convert to anomalies removing seasonal cycle
def anomaly( t ):
  a = numpy.empty_like(t)
  for m in range(12): a[m::12] = t[m::12] - numpy.nanmean(t[m::12])
  return a


# align and merge a set of fragments
def mergeall( data, cmin=1 ):
  # select only fragments which overlap the longest fragment
  flags = ~numpy.isnan(data)
  imax = numpy.argmax(numpy.sum(flags,axis=1))
  flagc = flags[imax,:]
  mask = numpy.zeros([flags.shape[0]],numpy.bool)
  for c in range(10):
    for i in range(flags.shape[0]):
      if ~mask[i]:
        if numpy.sum(numpy.logical_and(flags[i,:],flagc)) >= cmin:
          mask[i] = True
          flagc = numpy.logical_or(flagc,flags[i,:])
  # now find best offsets for merging
  dsel = data[mask,:]
  for c in range(100):
    davg = numpy.nanmean(dsel,axis=0)
    dmax = 0.0
    for i in range(dsel.shape[0]):
      diff = numpy.nanmean(davg-dsel[i,:])
      dsel[i,:] += diff
      dmax = max( dmax, abs(diff) )
    if dmax < 1.0e-4: break
  # return the result
  return davg


# Change point detection using Taylor (2000)
def changepoints( dorig, nsig=3.5, nbuf=2 ):
  if dorig.shape[0] < 3*nbuf: return []
  dnorm = dorig - numpy.mean(dorig)
  cusum = numpy.cumsum(dnorm)
  dboot = numpy.random.choice(dorig,size=(dorig.shape[0],500))
  dboot -= numpy.mean(dboot,axis=0)
  cusumb = numpy.cumsum(dboot,axis=0)
  custd = numpy.std(cusumb,axis=1)
  ratio = numpy.absolute(cusum)/numpy.maximum(custd,1.0e-30)
  i = numpy.argmax(ratio[nbuf:-nbuf])+nbuf
  if ratio[i] <= nsig: return []
  return changepoints(dorig[:i],nsig,nbuf)+[i]+[x+i for x in changepoints(dorig[i:],nsig,nbuf)]


# calculate changepoints on data with breaks
def changemissing( dorig, nsig=3.5, nbuf=2 ):
  dnorm = anomaly( dorig )
  mask = ~numpy.isnan(dnorm)
  diff = dnorm[mask]
  chg = changepoints(diff,nsig,nbuf)
  index = numpy.arange(dnorm.shape[0])[mask]
  return [index[i] for i in chg]



# MAIN PROGRAM
numpy.random.seed(1)
years = range(1921,2011)
datef = []
for y in years:
  for m in range(1,13):
    datef.append(y+m/12.0-1/24.0)

# read ghcn inventory
stations1 = {}
tnull = numpy.empty([len(datef)])
tnull.fill( numpy.nan )
for line in open(sys.argv[1]):
  stations1[line[0:11]] = { "lati":float(line[12:20]), "lngi":float(line[21:30]), "data":tnull.copy() }
# read and store the temperature data
for line in open(sys.argv[2]):
  id = line[0:11]
  if id in stations1:
    year = int( line[11:15] )
    if year in years:
      for m in range(12):
        temp = int( line[19+8*m:24+8*m] )
        flag = line[24+8*m:37+8*m]
        if temp != -9999 and flag[0:2] == "  ":
          stations1[id]["data"][12*(year-years[0])+m] = 0.01*temp
```

```python

# output file
of = open( sys.argv[1].rsplit('.',2)[0]+"_adj.dat", "w" )

# loop over stations
nwin = 7*12
for key1 in stations1:
  # check for enough data
  t1 = stations1[key1]["data"]
  a1 = anomaly(t1)
  if sum(numpy.isnan(t1)) > t1.shape[0]*2//5: continue

  # find neighbours
  la1,lo1 = stations1[key1]["lati"], stations1[key1]["lngi"]
  neighbours = []
  for key2 in stations1:
    la2,lo2 = stations1[key2]["lati"], stations1[key2]["lngi"]
    d = distance( la1, lo1, la2, lo2 )
    neighbours.append((d,key2))
  neighbours = [k for d,k in sorted(neighbours)[:75]]

  # find the most correlated neighbours
  keys,rmss,miss = [],[],[]
  for key2 in neighbours:
    # fetch temperatures
    t2 = stations1[key2]["data"]
    a2 = anomaly(t2)
    # calc min of moving rmsd
    nrmsd,nrmsdmin = nwin, nwin*2//3
    smin = 1.0e20
    for i in range(0,a1.shape[0]-nrmsd,12):
      nnan = numpy.sum(numpy.isnan(a1[i:i+nrmsd]-a2[i:i+nrmsd]))
      if nnan < nrmsd-nrmsdmin:
        s = numpy.nanstd(a1[i:i+nrmsd]-a2[i:i+nrmsd]) / numpy.nanstd(a1[i:i+nrmsd])
        if s < smin: smin = s
    #print key1, key2, smin, tmin
    if smin < 0.5:
      keys.append(key2)
      rmss.append(smin)
      miss.append(numpy.sum(numpy.isnan(a1-a2)))
  s1,s2 = scipy.stats.rankdata(rmss), scipy.stats.rankdata(miss)
  scores = [(1*s1[i]+2*s2[i],keys[i]) for i in range(len(keys))]

  # assemble up to 15 temperature series (speed)
  scores.sort()
  nkeep = min( len(scores), 15 )
  ts = numpy.empty( [nkeep,len(datef)] )
  for i in range(ts.shape[0]):
    ts[i,:] = stations1[scores[i][1]]["data"]
  print key1, len(neighbours), len(scores), nkeep

  # prune months with insufficient data
  counts = numpy.sum(~numpy.isnan(ts),axis=0)
  ts[:,counts<4] = numpy.nan

  # loop over all pairs of stations
  fnull = numpy.empty([len(datef)])
  fnull.fill( numpy.nan )
  frags = []
  npad = 6
  for i in range(0,nkeep-1):
    for j in range(i+1,nkeep):
      diff = ts[i,:]-ts[j,:]
      if numpy.sum(~numpy.isnan(diff)) > 180:
        chg = [-npad]+changemissing(diff,3.5,6)+[len(diff)+npad]
        for c in range(1,len(chg)):
          f = fnull.copy()
          f[chg[c-1]+npad:chg[c]-npad] = ts[i,chg[c-1]+npad:chg[c]-npad]
          f[numpy.isnan(diff)] = numpy.nan
          frags.append(f)
          f = fnull.copy()
          f[chg[c-1]+npad:chg[c]-npad] = ts[j,chg[c-1]+npad:chg[c]-npad]
          f[numpy.isnan(diff)] = numpy.nan
          frags.append(f)
  frags = numpy.array(frags)

  if len(frags) == 0: continue

  # combine
  local = mergeall(frags,36)

  # calculate changepoints
  diff = local - t1
  chg = changemissing(diff,3.5,6)
  for i in chg: print datef[i]

  # calculate shifts
  adj = numpy.empty_like(t1)
  chg = [0] + chg + [diff.shape[0]]
  for i in range(1,len(chg)):
    adj[chg[i-1]:chg[i]] = numpy.nanmean(diff[chg[i-1]:chg[i]])
  adj[numpy.isnan(diff)] = numpy.nan
  adj -= numpy.nanmean(adj)

  # apply adjustment
```

```
    t2 = t1+adj

    # write
    t2[numpy.isnan(t2)] = -99.99
    s = ""
    for i in range(0,len(datef),12):
       c = numpy.sum(t2[i:i+12]>-99.00)
       if c > 0:
          s += "%11s%4dTAVG"%(key1,int(datef[i]))
          for j in range(12): s += "%5d   "%(int(round(100*t2[i+j])))
          s += "\n"
    of.write(s)

# done
of.close()
```