Open Peer Review Journal

Urbanization bias III. Estimating the extent of bias in the Historical Climatology Network datasets

Ronan Connolly *[*1], ***Michael Connolly*** [1]

[1] *Connolly Scientific Research Group. Dublin, Ireland.*

Abstract

The extent to which two widely-used monthly temperature datasets are affected by urbanization bias was considered. These were the Global Historical Climatology Network (GHCN) and the United States Historical Climatology Network (USHCN). These datasets are currently the main data sources used to construct the various weather station-based global temperature trend estimates.

Although the global network nominally contains temperature records for a large number of rural stations, most of these records are quite short, or are missing large periods of data. Only eight of the records with data for at least 95 of the last 100 years are for completely rural stations.

In contrast, the U.S. network is a relatively rural dataset, and less than 10% of the stations are highly urbanized. However, urbanization bias is still a significant problem, which seems to have introduced an artificial warming trend into current estimates of U.S. temperature trends.

The homogenization adjustments developed by the National Climatic Data Center to reduce the extent of non-climatic biases in the networks were found to be inadequate, inappropriate and problematic for urbanization bias. As a result, the current estimates of the amount of "global warming" since the Industrial Revolution have probably been substantially overestimated.

Citation:
R. Connolly, and M. Connolly (2014). *Urbanization bias III. Estimating the extent of bias in the Historical Climatology Network datasets*, Open Peer Rev. J., 34 (*Clim. Sci.*), ver. 0.1 (non peer reviewed draft). URL: http://oprj.net/articles/climate-science/34

Version: 0.1 (non peer-reviewed)
First submitted: January 8, 2014.
This version submitted: February 3, 2014.

1 Introduction

This paper is the third in a series of three papers considering the extent to which urbanization bias has affected current estimates of global temperature trends since the late 19th century. We will refer to the first two papers as Paper I[1] and Paper II[2], respectively. In this paper, we will be considering the extent to which the two main weather station datasets currently used are affected by urbanization bias.

Urban areas tend to be warmer than the surrounding rural areas - the so-called "urban heat island effect"[3]. Although this effect genuinely alters *local* climate, it is a localised phenomenon and does not reflect the climatic trends of the non-urban surroundings. Therefore, if the area around a weather station becomes more urbanized over the years, its temperature record can be contaminated by an artificial warming trend from the growth of the urban heat island, which is unrepresentative of the actual regional temperature trends.

In other words, it introduces an *urbanization bias* into temperature trends[4]. Since many of the weather stations used for calculating global temperature trends are currently more urbanized than they were in the 19th century, it is likely that urbanization bias has introduced a substantial artificial "global warming" bias into the current estimates.

The current estimates of global temperature trends suggest that there has been a global warming trend of roughly 0.8°C/century, since the late 19th century[5]. Current climate models attribute most of this warming to "anthropogenic global warming" caused by increases in atmospheric carbon dioxide concentrations[6], and on the basis of the apparent success of

[*] Corresponding author: ronanconnolly@yahoo.ie. Website: http://globalwarmingsolved.com

these models, serious changes to economic and social policy are being proposed/implemented, e.g., see the Stern review[7]. However, if it transpires that some (or possibly all) of this apparent global warming is an artefact of urbanization bias, then this would cast doubt on the assumed robustness of the models.

Weather records are also routinely used for calibrating the various "temperature proxies" (tree rings, lake sediments, etc.) used in palaeoclimate reconstructions of temperatures of the last millennium or so[8]. Therefore, if these weather records are affected by urbanization bias, this could contaminate the proxy calibration process, reducing the reliability of the palaeoclimate reconstructions - we discuss these reconstructions in Ref. [9].

For these reasons, the urbanization bias problem seriously affects our climatic understanding of global temperatures (past, present and future). Hence, in this series of three papers, we have attempted to systematically determine the extent of the problem.

It is well-known that weather station records are often contaminated by non-climatic biases, e.g., see Mitchell, 1953 for a still-relevant summary[10]. So, the possibility that urbanization bias could be a major problem should be obvious. However, the problem seems to have been seriously underestimated by most of the groups calculating global temperature trends.

In Paper I[1] we found that there were problems with each of the papers which have claimed that the extent is only small or negligible. In Paper II[2], we assessed the data corrections applied by the only group adjusting their data for urbanization bias before constructing their global temperature trends, i.e., the National Aeronautics and Space Administration (NASA)'s Goddard Institute for Space Studies. We found that these data corrections were inadequate, and actually introduced about as many biases as they removed.

In this paper we will attempt to estimate the extent to which the weather station datasets used for constructing global temperature trends are affected by urbanization bias. We will focus on the two main datasets currently used. These are the Global Historical Climatology Network (often referred to by the acronym "GHCN")[11–13] and the United States Historical Climatology Network ("USHCN")[14–16]. Both of these datasets are compiled and maintained by the National Oceanic and Atmospheric Administration (NOAA)'s National Climatic Data Center (NCDC).

The National Climatic Data Center have developed a series of homogenization adjustments which they apply to some versions of the two datasets in an attempt to remove any non-climatic biases. Some researchers have claimed that these homogenization adjustments have removed (or at least substantially reduced) the extent of the urbanization bias problem, e.g., Menne et al., 2009[15]. Hence, a second aim of this paper will be to assess how successful (or unsuccessful) these adjustments are.

The format of the paper will be as follows. In Section 2, we will briefly describe the two datasets and their use in the current global temperature trend estimates. In Section 3, we will consider the extent to which the weather records in both datasets are likely to be affected by urbanization bias. In Section 4, we will discuss whether the National Climatic Data Center's homogenization adjustments have successfully resolved the urbanization bias problem, or not. Section 5 will then offer some concluding remarks.

2 Description of the two datasets

The Global Historical Climatology Network and the U.S. Historical Climatology Network are datasets containing monthly temperature records for a large number of weather stations. Both datasets are maintained by the National Climatic Data Center. At the time of writing, the latest versions could be downloaded from http://www.ncdc.noaa.gov/ghcnm/ and http://www.ncdc.noaa.gov/oa/climate/research/ushcn/ respectively. For the analysis in this paper, we used versions downloaded in July 2012 for the U.S. dataset and January 2013 for the global dataset[1]. The global dataset contains records for 7280 stations from around the world, and the U.S. dataset contains records for 1218 stations from the "contiguous United States", i.e., all of the U.S., except for Alaska and Hawaii.

The global dataset contains all of the station records in the U.S. dataset. However, the U.S. dataset is compiled separately by the National Climatic Data Center, and is only merged with the rest of the global dataset in one of the final steps of archiving[12, 13].

[1] For some reason, the January 2013 Version 2.0 archive for the U.S. dataset only includes data up to 2008. More recent archives (Version 2.5) are stored using a different format to the one we had originally written our analysis scripts for. Hence, we used the archive we had downloaded in July 2012.

The U.S. dataset is also generally of a higher quality and includes a *station history* archive[14]. This station history archive contains the years of any unusual changes associated with each station that had been reported by the observers - changes in the types of instruments used, station relocations, etc. In contrast, the Global Network only provides some basic *station metadata*, describing a few details about the station's current location.

For these reasons, we will consider the two datasets separately in this paper. With this in mind, we will refer to the global dataset as being those 6051 stations in the Global Historical Climatology Network which are *not* contained in the U.S. Historical Climatology Network. For the rest of the paper we will refer to this global dataset as the "Global Network" and the U.S. dataset as the "U.S. Network". When we refer to both datasets collectively, we will use the title "Historical Climatology Networks".

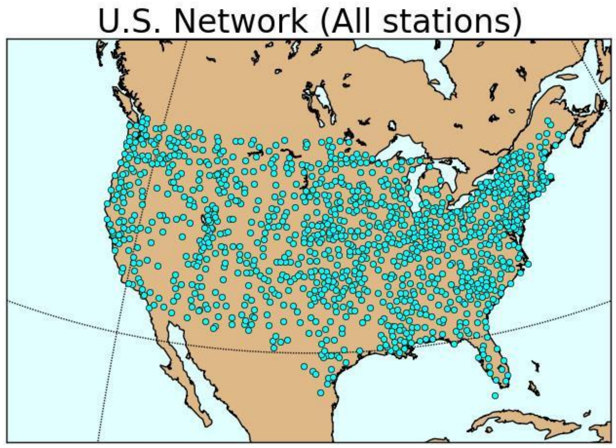

Figure 1: *Location of all the stations in the U.S. Network.*

The locations of the stations in the U.S. Network and the Global Network are shown in Figures 1 and 2. It can be seen from Figure 2 that the Global Network still contains a large number of stations (560) from the contiguous U.S. These are different stations than the ones in the U.S. Network.

Figure 3 shows the number of stations with data available for each year in the two datasets. Although the Global Network contains more than 6000 stations, the majority of these stations have relatively short records, and mostly just cover the period 1950-1990. As a result, the number of stations available drops off quite sharply outside this period.

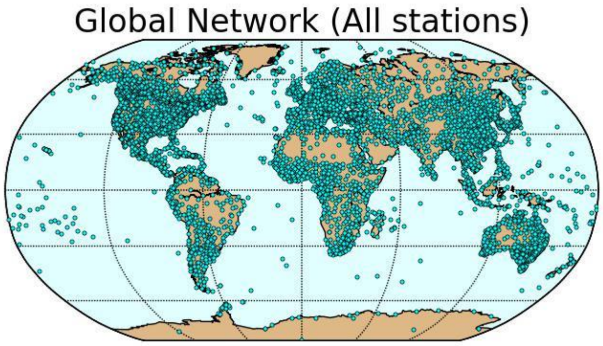

Figure 2: *Location of all the stations in the Global Network.*

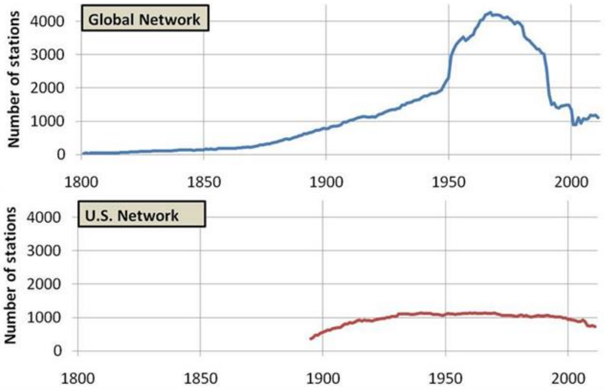

Figure 3: *Number of stations available in both networks for each year.*

The sharp post-1990 drop is particularly surprising as one might expect that there would have been an increase in station numbers and data availability in recent decades. *Some* of the reduction is due to station closures. However, apparently the post-1990 reduction is *not* predominantly a result of closing stations[13], since many of these stations are still active. Rather, it is a consequence of the fact that when the Global Network was initially compiled, a large number of monthly station archives which had been completed up to 1990 were incorporated, but these particular archives have not been updated since[13]. Rohde et al., 2013b suggest that this is mainly because the National Climatic Data Center only use monthly data for constructing the Global Network, but with the increase in global communication, a lot of stations have switched to reporting daily results and have stopped reporting the monthly averages[17].

In contrast to the Global Network, most of the station records in the U.S. Network are relatively long and cover most of the period 1895-now. A striking

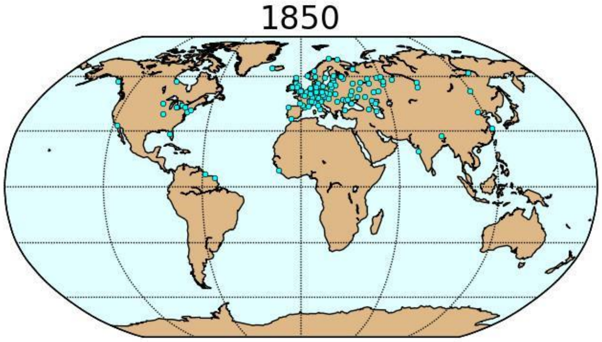

Figure 4: *Location of all the stations in the Global Network with data for 1850 A.D.*

consequence of this is that there are nearly as many stations with data for the 21st century in the U.S. Network as in the Global Network. Since the contiguous U.S. only accounts for about 4% of the global land surface, but the number of stations in both networks are comparable for long periods, this means that the U.S. Network has a much higher station density than the Global Network, particularly for the 21st century.

Nominally, the Global Network begins in 1701, but only a few of the records are available for the earlier periods, and they are mostly confined to European (and a few North American) locations, e.g., see Figure 4. It is only towards the end of the 19th century that a sample of more than a couple of hundred stations become available. For this reason, most of the weather station-based global temperature trend estimates only begin in 1880, although the Climate Research Unit begin their analysis in 1850[18] and Berkeley Earth carry out an analysis which begins in 1753[19].

Table 1 illustrates the heavy reliance on the Historical Climatology Networks by most of the current weather station-based estimates of global temperature trends. The National Climatic Data Center's estimate[23, 24] and the Rodhe et al., 2013b[17] estimate rely entirely upon them, while the Goddard Institute of Space Studies[25] and Tokyo Climate Center[20] estimates are nearly-exclusively based on them.

The Climate Research Unit's estimate is based on an in-house dataset[18], and initially it does not appear to rely too heavily on the Historical Climatology Networks. However, they still use the Historical Climatology Networks as one of their main sources. Additionally, it was constructed from similar data sources to the ones used for constructing the Historical Climatology Network. Finally, it is itself one of the main sources used by the Historical Climatology Network. As a result, the overlap between the Climate Research Unit's dataset and the Historical Climatology Networks is believed to be higher than 98%[26, 27].

The Lugina et al., 2006[22] estimate did not use the Historical Climatology Networks directly, but they used many of the same data sources used by the compilers of the Historical Climatology Networks, and so their dataset probably has many similarities to the Historical Climatology Networks[2].

Recently, following doubts about the reliability of the Historical Climatology Networks[28], the Berkeley Earth group compiled their own dataset from publicly archived weather station records. The Berkeley Earth dataset includes the Historical Climatology Networks, but also includes several other datasets, and as a result contains a much larger number of total stations ($\sim 40,000$). However, the approach of the Berkeley Earth group has been quantity over quality. They argue that quality decisions are subjective, and therefore provide as many records as they can, so that the user can make their own quality decisions. So, while their dataset nominally includes a lot of stations, many of them are only of limited value for studying long-term temperature trends, e.g., more than half of the stations have less than 30 years of data. Hence, the Historical Climatology Networks are arguably among the most useful components of the Berkeley Earth dataset.

For instance, the National Climatic Data Center compiled the U.S. Network by selecting only those records which appeared to be of a relatively high quality from NOAA's National Weather Service's larger Cooperative Observer Program (COOP) dataset[4]. Berkeley Earth did not carry out such a selection, but rather included *all* stations from both the U.S. Network and its parent dataset (the COOP), regardless of quality. In addition, much of the Berkeley Earth group's preliminary analysis has been exclusively based on the Historical Climatology Networks[17, 29].

[2] Lugina et al. did not archive their dataset, so we were unable to directly compare it to the Historical Climatology Networks, but they did publish their data sources[22], which heavily overlap with the data sources used for the Historical Climatology Networks[12].

Research Group	Stations from the Networks		Version used
Heavy reliance on the two networks:			
NOAA's National Climatic Data Center[13]	100%	7280 out of 7280	Adjusted
NASA's Goddard Institute for Space Studies	99.3%	6280 out of 6322	† Adjusted
JMA's Tokyo Climate Center[20]	99.6%	3883 out of 3900	Adjusted
Berkeley Earth - Rohde et al., 2013b[17]	100%	7280 out of 7280	Unadjusted
Similar data sources to the two networks:			
Climate Research Unit - CRUTEM3[18]	35.4%	1809 out of 5113	Adjusted
Climate Research Unit - CRUTEM4[21]	29.0%	1617 out of 5583	Adjusted
Berkeley Earth - full dataset[19]	18.7%	7280 out of 39028	Unadjusted
Lugina et al., 2006[22]	0%	0 out of 685	None

Table 1: *Use of the Historical Climatology Network datasets by the various weather station-based global temperature estimates. † The Goddard Institute for Space Studies originally used the unadjusted datasets, but in 2001, they started using the adjusted version of the U.S. Network, and since 14th December 2011, they have also being using the adjusted version of the Global Network.*

2.1 Gridding methods used in this article

When assessing the datasets, it is often helpful to examine the mean temperature trends for different subsets. In this section, we describe the averaging techniques we use for calculating these trends in this article. A complete description of these techniques is described in the supplementary information. But, briefly, we take the following approach:

1. All stations meeting the required characteristics for a particular subset are identified.

2. Each station's monthly temperature record is converted into an annual temperature record, by calculating the mean temperature of all 12 months for a calendar year. If data for one or more months is missing for a year, we do not calculate the mean for that year.

3. The annual records for each station are then converted into "temperature anomaly records", by subtracting the mean annual temperature over the 1961-1990 period for that station from each annual value. This means that if the temperature anomaly for a given year is negative, then it was colder than the 1961-1990 average for that year, and if it is positive, then it was warmer.

4. Stations are then assigned into 5° latitude × 5° longitude boxes. The annual mean temperature anomalies for a given grid box are then calculated as the mean of all of the available anomaly records for each year, in that grid.

5. Gridded global or regional temperature anomalies are then calculated for the subset by averaging together the anomalies of all grid boxes with data for a given year, weighting by the surface area of the grid boxes.

6. Where given, confidence intervals represent twice the standard error of the mean gridded anomaly for that year. They are merely statistical in nature, and are calculated by assuming no biases in the temperature data.

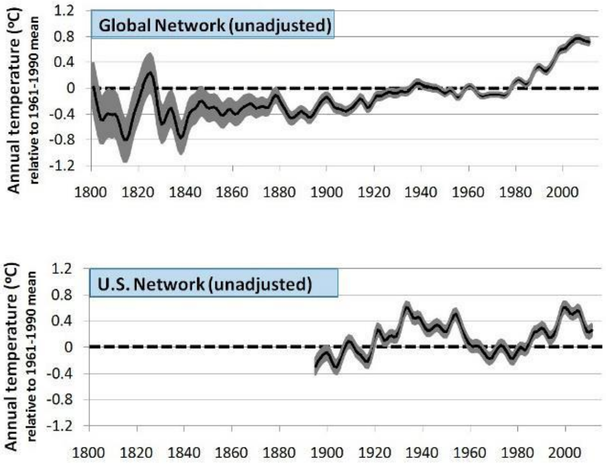

Figure 5: *Mean temperature trends (solid black lines) and confidence intervals (grey shaded regions) for the unadjusted Global Network (top) and U.S. Network (bottom), gridded into 5° × 5° boxes, relative to 1961-1990 means, with 11 point binomial smoothing applied.*

Figure 5 shows the mean gridded temperature trends of the unadjusted Global and U.S. Networks, smoothed using 11-point binomial smoothing. We can see that the Global Network covers a longer period (some stations even have data for the 18th century, which is not shown). But, the error bars and the decadal variability are much larger for the 19th century, because the estimates are only based on a small number of stations. Also, the stations with data for the earlier period are mostly confined to Europe and North America (Figure 4). This suggests the "global" temperature trends are less reliable for the 19th century, and so we will mostly confine our analysis to the period beginning in the late 19th century, when the U.S. Network also begins.

From the 1890s to the early 1940s and from the late 1970s to 2000s, both datasets show "global warming" (technically, regional warming in the case of the U.S. Network). From the 1940s to the 1970s, both datasets show a cooling trend. But, while the U.S. Network shows considerable cooling, for the Global Network there is only a slight cooling (or perhaps a "plateau").

For the U.S. Network, both the warming periods and cooling period are all of a similar magnitude, while for the Global Network, the more recent warming period is greater than the early warm period, and the magnitude of the cooling period is slight. As a result, the Global Network dataset suggests an almost continuous "global warming" since the late 19th century, while the U.S. Network merely suggests a multidecadal variation between warming and cooling periods.

However, the trends in Figure 5 are based on data that includes urban station records. Therefore, we know that at least *some* of the apparent regional warming (U.S. Network) and global warming (Global Network) trends are a result of urbanization bias. The challenge of the urbanization bias problem is in determining how much.

3 Extent of urbanization bias in unadjusted datasets

When they were compiling the Global Historical Climatology Network dataset, the National Climatic Data Center included some basic *station metadata*, i.e., data describing the station and its environment. For each station, they provided the station name, country, latitude, longitude and elevation. They also provided a number of classifications to describe the environment of the station - whether it was an airport station or not; if it was on an island, near the coast or near a lake; and what the average ecosystem of the stations' surroundings was, e.g., desert, ice, forest, etc.

Importantly for this paper, this metadata also included estimates of how urbanized the stations are. For each station they included rough estimates of the population associated with any neighbouring towns or cities in the vicinity, provided the estimated population was greater than 10,000. On the basis of this estimate they defined each station as being "rural" (population $< 10,000$), "small town" ($10,000 \leq$ population $< 50,000$) or "urban" (population $\geq 50,000$). They also considered an additional measure based on the brightness of the night-lights in the area, as determined from satellite measurements[3]. As a result, the Global Network metadata includes alternative urbanization estimates based on night brightness. As for the population estimates, there are three possibilities: rural, small town or urban[4].

In this article, we will define stations as being *"fully rural"* if they are rural according to *both* estimates, *"fully urban"* if they are urban according to *both* estimates and otherwise *"intermediate"*.

The U.S. Historical Climatology Network dataset also includes some metadata, and in some cases this metadata is more useful, e.g., station co-ordinates are more precise. However, they do not include the urbanization classifications. Fortunately, since the Global Historical Climatology Network includes the U.S. Historical Climatology Network dataset, we were still able to extract these details for the U.S. Network stations from the Global Historical Climatology Network metadata file.

Table 2 shows the total number of stations in the Global and U.S. Networks, divided into our three urban classifications, and Figures 6 and 7 show their locations. It can be seen that a large number of stations are either *fully urban* or *intermediate*. Particularly for the longer station records, many of these station records are likely to have been affected to some extent by urbanization bias.

Comparing the two networks, the Global Network has a greater percentage of *fully rural* stations than the U.S. Network ($\sim 33\%$ compared to $\sim 23\%$). As a

[3] Areas which have been highly urbanized and have access to electricity, tend to be brighter at night because of street lights, building lights, etc.

[4] In the metadata file, they use the letters A-C to indicate night brightness, with rural = "A", small town = "B" and urban = "C"

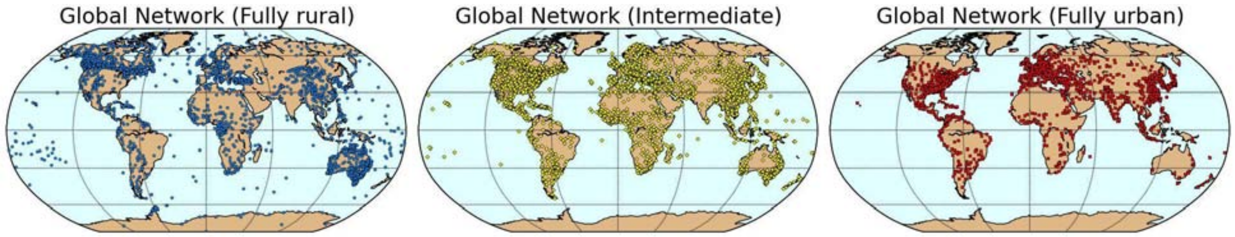
Global Network (Fully rural)
Global Network (Intermediate)
Global Network (Fully urban)

Figure 6: *Maps showing the location of the stations in the Global Network identified as* fully rural *(left),* intermediate *(middle) or* fully urban *(right)*

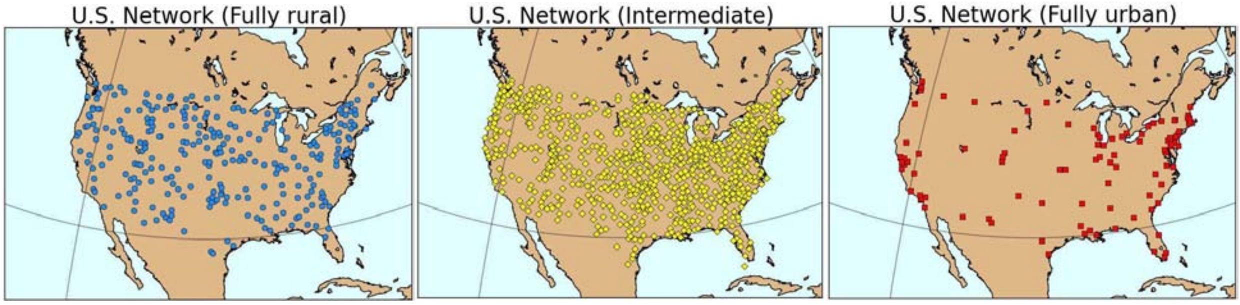
U.S. Network (Fully rural)
U.S. Network (Intermediate)
U.S. Network (Fully urban)

Figure 7: *Maps showing the location of the stations in the U.S. Network identified as* fully rural *(left),* intermediate *(middle) or* fully urban *(right)*

result, one might initially assume that urbanization bias would be less of a problem for the Global Network. However, this is wrong for at least two reasons.

First, the Global Network contains a greater percentage of *fully urban* stations (∼ 25% compared to ∼ 8%). In addition, we saw from Figure 3 that the number of stations with available data is very inconsistent over time in the Global Network. As we will discuss in Section 3.2, this inconsistency is greatest for the *fully rural* stations, and most of the *fully rural* stations only cover a relatively short period (mainly 1950-1990). However, before we discuss the Global Network, let us consider the extent of urbanization in the U.S. Network.

Subset	U.S. Network	Global Network
Fully urban	99 (8.13%)	1508 (24.92%)
Intermediate	842 (69.13%)	2556 (42.24%)
Fully rural	277 (22.74%)	1987 (32.84%)
All stations	1218 (100.00%)	6051 (100.00%)

Table 2: *Total number of stations in both networks with each level of urbanization.*

3.1 Urbanization bias in the U.S. Network

When compiling the U.S. Network, Karl et al., 1988 placed considerable effort into creating a mostly rural network[4]. Their relative success in this is apparent from Table 2 by the fact that less than 10% of the stations in the U.S. Network are *fully urban*. However, they were still unable to completely remove the urban and partially urban stations from the U.S. Network.

Since the U.S. Network has a high station density, it is possible to estimate the magnitude of its urbanization bias by separately calculating the mean gridded temperature trends for the *fully rural* stations and comparing them to the *fully urban* stations. Locations of these two subsets (as well as the *intermediate* subset, which we will not consider here) are shown in Figure 7. The mean gridded temperature trends of both subsets are shown in Figure 8. The top panel of Figure 9 shows the difference between the two subset trends (using the 11-point binomial smoothed trends for visual clarity).

The urban subset shows a noticeable warming trend relative to the rural subset of about $0.7°C$/century. While the relative warming is small

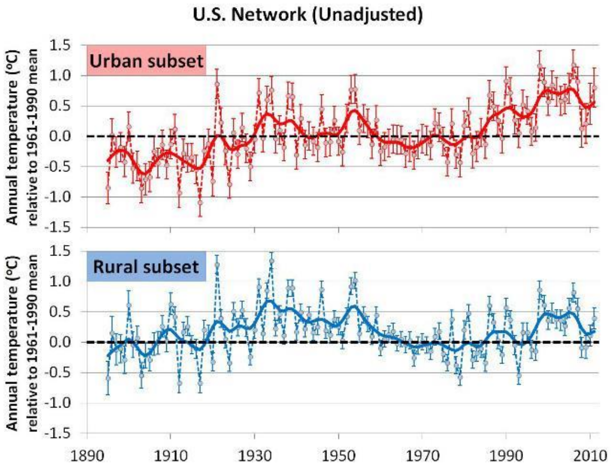

Figure 8: *Mean gridded trends of the* fully urban *and* fully rural *U.S. Network subsets of Figure 7 using the unadjusted dataset. Solid lines correspond to the 11 point binomial smoothed versions of the annual values. Confidence errors correspond to twice the standard error of the annual means.*

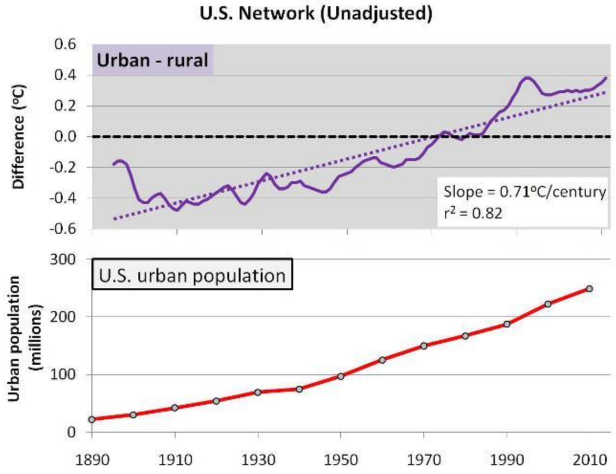

Figure 9: *The top panel shows the difference between the the* fully urban *and* fully rural *U.S. Network subsets of Figure 8 (the smoothed trends). The bottom panel shows the urban population growth for the U.S. (bottom), as determined from U.S. Census figures (Table 7 of Ref. [30], downloaded from* http://www.census.gov*).*

relative to the long-term trends of the two subsets, it is enough to make the recent 1980s-2000s warm period appear warmer than the 1920s-1940s warm period for the urban subset, the opposite to the rural subset. In other words, according to the urban subset, there has been a general warming trend for the U.S. since the start of the dataset, while according to the rural subset, the early 20th century was warmer than the recent warm period.

It is possible that some of the divergence between the subsets is due to differences between the subsets, other than urbanization. Indeed, in Section 4.2.1, we will present evidence suggesting that about $0.2°C$/century of the divergence is a consequence of different trends in observation times of the two subsets. However, from the bottom panel of Figure 9, we can see that the growing divergence between the two subsets roughly parallels the increase in U.S. urban population since the 19th century. Although population is not an exact measure of urbanization, it is a reasonable indicator of the degree of urbanization[4]. So, it is likely that much of the divergence is indeed due to urbanization bias in the urban subset.

We note that Hausfather et al., 2013 also found evidence for significant urbanization bias in the U.S. Network, although unlike us they were optimistic that most of this bias had been removed from the adjusted dataset by the homogenization algorithm which we discuss in Section 4[31]. Kalnay & Cai, 2003[32] also found evidence for urbanization bias in the U.S. Network, although their study led to some contentious debate (see our review of the debate in Paper I[1]).

3.2 Urbanization bias in the Global Network

Figures 10 and 11 show the percentage of the stations in each of the three urban categories with available data for each year, for the U.S. Network and the Global Network, respectively. Compared to the U.S. Network (Figure 10), the Global Network shows a remarkable inconsistency from year to year in the fraction of stations with different degrees of urbanization with data (Figure 11). Although nearly a third of the stations in the Global Network are *fully rural*, this fraction decreases rapidly before and after the 1950-1990 period. This decrease generally corresponds to an increase in the fraction of stations which are *fully urban*, rather than the less-urbanized *intermediate* stations.

More than half of the stations with available data for the early-to-mid-19th century are currently *fully urban*. In contrast, only about 10% of the stations with 19th century data are *fully rural*. It seems reasonable to assume that many of the stations which

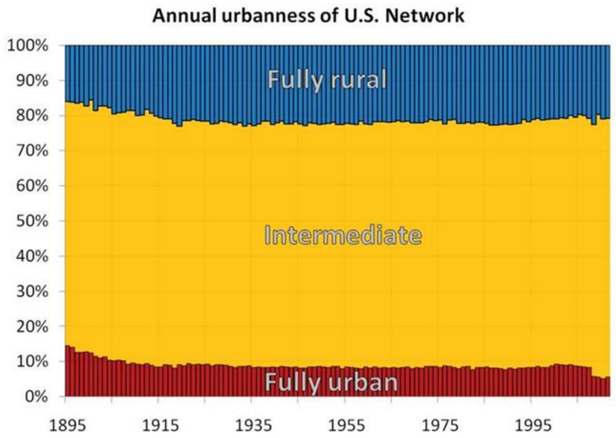

Figure 10: *Relative ratio of stations available for each year which have been identified as* fully rural, intermediate *and* fully urban, *for the U.S. Network datasets.*

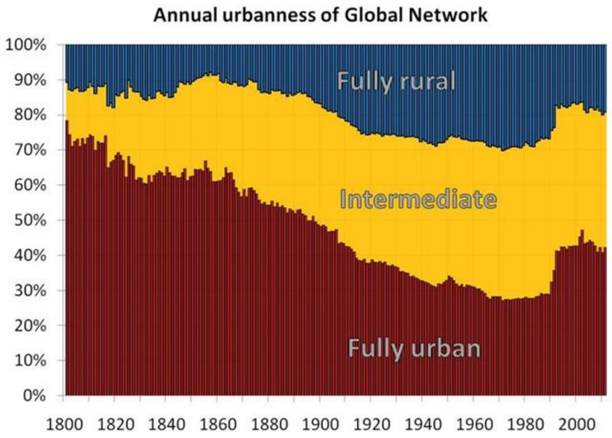

Figure 11: *Relative ratio of stations available for each year which have been identified as* fully rural, intermediate *and* fully urban, *for the Global Network. Note that the x-axis has an earlier start date than Figure 10.*

are currently identified as *fully urban* are more urbanized than they were in the 19th century. Hence, it is likely that the magnitude of the urban heat islands associated with those stations has increased since the 19th century. If so, this would have introduced an artificial warming bias to 20th century estimates of global temperature, relative to the 19th century - in the unadjusted records, at least. Records from stations identified as being *intermediate* in urbanization may also be affected by urbanization bias, suggesting that about 90% of the early-to-mid-19th century records in the Global Network are potentially biased by urbanization.

The fractions of *fully urban* and *intermediate* stations in the Global Network decreased in the 20th century, initially suggesting that the problem is somewhat reduced for more recent decades. However, there has been a dramatic increase in urbanization over the 20th century, particularly in recent decades[33], so the potential for urbanization bias is greater. Moreover, it appears from Figure 11 that the sharp decrease in station numbers after 1990 disproportionately affected the *fully rural* stations. Hence, estimates of global temperature trends since 1990 are likely to be more affected by urbanization bias than in the preceeding decades.

Figure 12 shows the locations of all the Global Network stations with data for at least 95% of the last hundred years, that are either *fully urban* (top), *intermediate* (middle) or *fully rural* (bottom). There are a reasonable number of *fully urban* stations which satisfy that fairly modest requirement (122), and quite a few *intermediate* stations (43). But, there are only eight *fully rural* stations which do so:

1. The Pas, Manitoba, Canada. 53.97°N, 101.10°W. ID = 40371867000.
2. Angmagssalik, Greenland. 65.60°N, 37.63°W. ID = 43104360000.
3. Lord Howe Island, New Zealand. 31.53°S, 159.07°E. ID = 50194995000.
4. Sodankylä, Finland. 67.37°N, 26.65°E. ID = 61402836000.
5. Hohenpeißenberg, Germany. 47.80°N, 11.02°E. ID = 61710962000.
6. Valentia Observatory, Ireland. 51.93°N, 10.25°W. ID = 62103953000.
7. Sulina, Romania. 45.15°N, 29.67°E. ID = 63715360000.
8. Säntis, Switzerland. 47.25°N, 9.35°E. ID = 64606680000.

A Google Earth file containing these eight locations, as well as possibly more accurate locations for the stations is provided in the Supplementary Information.

The realisation that less than 1% of the 6051 stations in the Global Network are both *fully rural* **and** have data for at least 95 of the last hundred years was quite shocking to us. All of the current weather station-based global temperature trend estimates cover a period longer than a hundred years (the Rohde et al., 2013a estimate even attempts to describe the period 1753-2011[19]). But, there do not currently appear to be enough *fully rural* stations

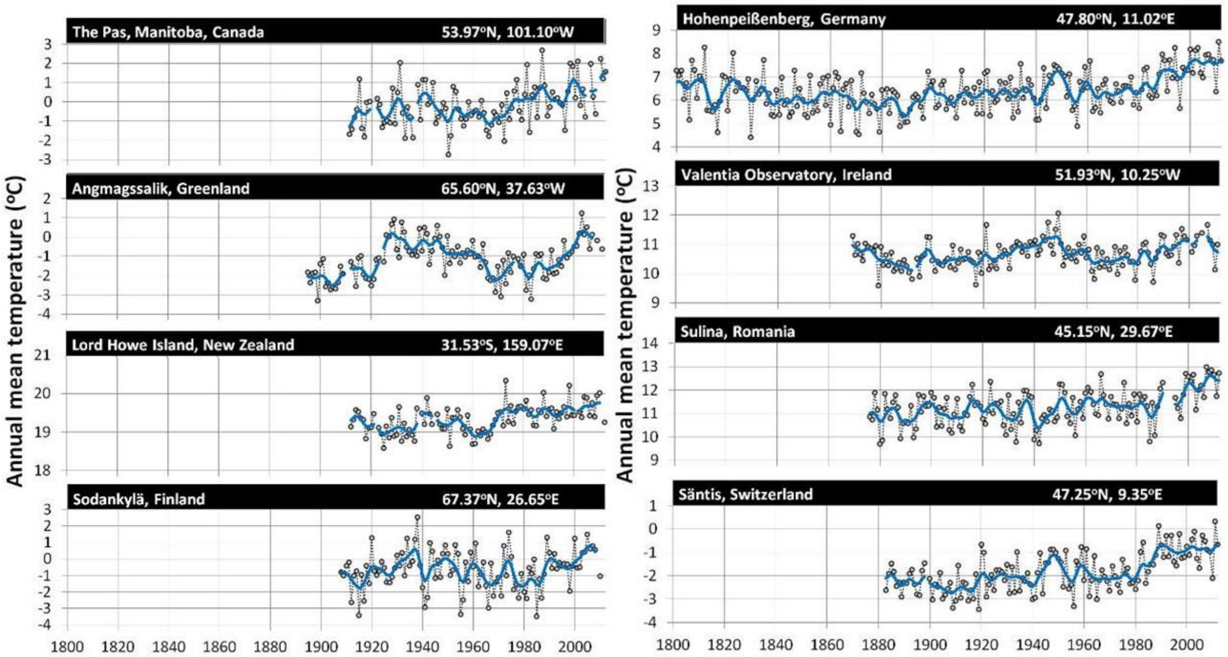

Figure 13: *Unadjusted temperature trends of the eight* fully rural *Global Network stations with data for* $> 95\%$ *of the last hundred years. Blue solid lines correspond to 11-point binomial smoothed trends.*

with enough data to do this (in the Global Network). It turns out that the global temperature trend estimates by the groups listed in Table 1 are predominantly based on a combination of:

- Stations which are **not** *fully rural*.
- *Fully rural* stations **with incomplete records**.

This means that they are strongly influenced by station records which are potentially affected by urbanization bias.

This seems a particularly insidious problem to try to overcome, and we are not sure it is possible using the current Global Network alone. One might argue that we could at least combine the 8 station records above to construct a reasonable estimate for the global temperature trends of the last 100 years. Temperature trends for these eight stations are shown in Figure 13. However, there are several serious problems with this idea.

- Between them, the eight stations only provide a very limited "global" coverage. Only one of the stations is from the southern hemisphere (Lord Howe Island) and five of the stations are from the same continent, i.e., Europe. The maximum distance between any of the European stations is only about 3,000 km (Sulina - Valentia Observatory).
- It can be seen from Figure 13 that there is a lack of consistency between the eight records. For instance, although several of the records suggest a warm period during the 1930s-1940s and another warm period during the 1990s-2000s, the relative warmth of these periods varies between stations.
- Most importantly, urbanization bias is *not* the only non-climatic bias which can affect station records, e.g., see Mitchell, 1953[10].

The last problem is worth elaborating on. There are many changes which could occur at a weather station which could introduce a non-climatic bias into the station record, e.g., changes in station location[34], observation practice[34–36], station microclimate[37, 38], instrumentation used[39, 40] or local land use[41].

For instance, although daily observations were recorded manually almost continuously at the Säntis weather station, from the time it was set up in 1882[42], in the late 1970s these manual observations were replaced with the installation of an automated

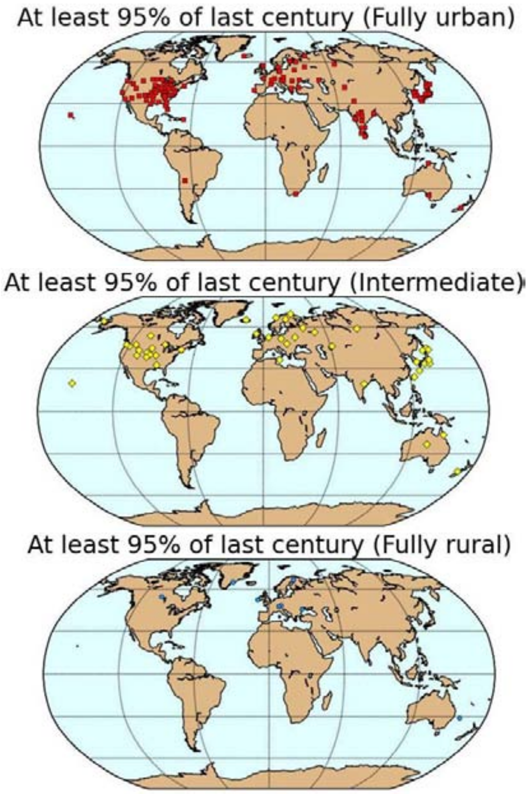
At least 95% of last century (Fully urban)
At least 95% of last century (Intermediate)
At least 95% of last century (Fully rural)

Figure 12: *Location of* fully urban *(top),* intermediate *(middle) and* fully rural *(bottom) Global Network stations with data for at least 95 of the last hundred years.*

weather station[43]. Moreover, in recent years, the location of the station has become a popular mountain resort with the construction of a large hotel and amenities (Säntis der Berg). It is plausible that some of these changes, or others introduced some non-climatic biases.

As another example, let us consider the Sulina station. By using Google Earth aerial viewing software, it appears to us that the current location for the Sulina station is probably 45.1624°N, 29.7267°E - on a concrete platform a few metres from the River Danube (not far from where the Danube enters the Black Sea). However, the metadata places the station at a location nearly 5 kilometres west of there, near the town centre, i.e., at the co-ordinates listed in the table above. It is unclear whether this is due to inaccurate metadata, or whether it represents a station move. But, a few bits of station history for the Sulina station are provided by Jones et al., 1985[44], and this suggests that the location of the "Sulina station" has varied since it was first set up, e.g., during the period 1941-1946 (i.e., around the time of World War II), measurements were made from a town 140 kilometres south of Sulina, i.e., Constanta (44.18°N, 28.67°E). The Lord Howe Island station has similarly been moved several times since it was set up in 1886[45].

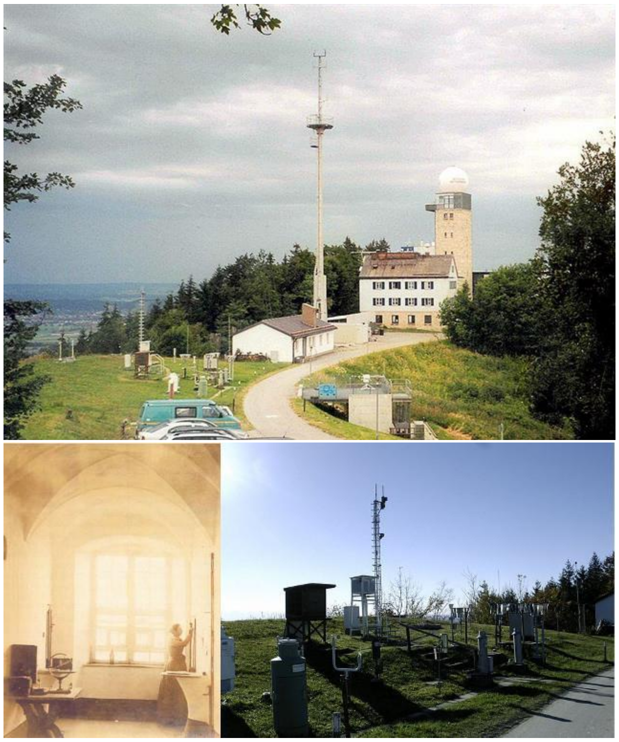

Figure 14: *Photographs of the Hohenpeißenberg Meteorological Observatory. Top photo is of the entire observatory in July 2003, by Rainer Lippert who has placed it in the public domain. Bottom two photos are of the actual thermometer station - in 1897 (left) and in July 2007 (right). The photographer for the 1897 photo is unknown, but the photo is in the public domain due to copyright expiry. The 2007 photo was by Christoph Radtke, who has placed it in the public domain. All photos downloaded from the Wikimedia Commons website (`http://commons.wikimedia.org/`).*

The Hohenpeißenberg station is also potentially affected by non-climatic biases. We found a quite-detailed review of the Hohenpeißenberg station history on the Wikipedia.de website (in German)[46]. This history recounts a number of changes which could potentially have introduced non-climatic biases into the record. For example, until the 20th century, measurements were made indoors, and in recent decades a number of buildings and new equipment have been introduced in order to improve the observatory's atmospheric monitoring system. See Figure 14. We also note that an extensive tree-felling operation was apparently carried out in 2000.

It is also difficult to completely rule out the possibility that some of the *"fully rural"* stations may have also been affected by urbanization. For instance, in 2011, it was decided to move the Valentia Observatory station from its then location a few kilometres from the small town of Cahirciveen (pop. 1,294 in 2006) to the nearby Valentia Island, over concerns that the increasing urbanization of Cahirciveen might be affecting some of the atmospheric measurements. While there was no evidence that this urbanization had affected the ground temperatures at the weather station[47], it does illustrate that urbanization is still an issue for the *"fully rural"* stations.

Even for areas which have not seen any urbanization, the station surroundings may have undergone significant *modernization* since the early 20th century (or earlier). As we discuss in Ref. [48], changes in the immediate surroundings of a weather station can introduce localised micro-climate trends with similarities to urbanization biases. For example, it is plausible that the construction of the runways and buildings at the "The Pas" weather station which is located at an airport (see Supplementary Information) could have introduced a localised warming bias. Similar potential problems may also exist for the other stations.

Of course, such problems are not unique to the stations mentioned. **Most** stations with relatively long records are affected by station changes which could potentially introduce non-climatic biases with an average frequency of at least once every twenty years[17, 34, 49]. The problem in this case is that we only have eight records, and they each suggest a different description of temperature trends since the start of the 20th century (see Figure 13). It is too small a sample to decide which (if any) of the records is the most representative of the true temperature trends.

Nonetheless, we do note from Figure 13 some common features between the records, with the possible exception of the Lord Howe Island station:

- Many of the records suggest an almost cyclical alternation between multi-decadal periods of warming and multi-decadal periods of cooling.
- Many of the records suggest that there was a relatively warm period during the early-to-mid-20th century (1930s/1940s) and also a relatively warm period in recent decades (1990s/2000s).
- In most of the records, the temperatures in recent years do not seem particularly unusual or unprecedented.

These features differ from the conventional description of global temperature trends, which suggest an almost continuous global warming since the Industrial Revolution, e.g., see the 2007 IPCC reports[5]. It seems likely that at least some of the differences between the two descriptions are due to urbanization bias in the conventional descriptions.

The Lord Howe Island station suggests that recent decades are warmer than the earlier part of the record. However, rather than suggesting a continuous warming trend, as was expected by current climate models[6], the record suggests there was a "step" increase in temperatures during the 1960s, and temperature trends were fairly "flat" before *and* after this step-change. Considering that the station has undergone a number of changes and relocations since it was first set up[45], it is plausible that much (or all) of this temperature shift may be a result of station changes, rather than an actual climatic shift. But in either case, this description also differs from the popular "global warming" description.

If we want to make more definitive statements on long-term global temperature trends, it might be helpful if more *fully rural* station records could be found that had data for at least 95 of the last 100 years. It would also be useful if more information on the station histories for the different records could be obtained.

We were able to obtain the few pieces of information mentioned above on these stations through some basic research on the internet. We suspect much more detailed information could be obtained by directly contacting the various national meteorological organizations and/or weather observers. However, the National Climatic Data Center did not collect *any* station histories for the Global Network. If this situation could be remedied, it would probably allow for more reliable assessments of the individual station records.

3.3 Urbanization bias in the Arctic

One region of the planet which is still relatively rural is the Arctic (defined here as latitudes $> 66°33'N$).

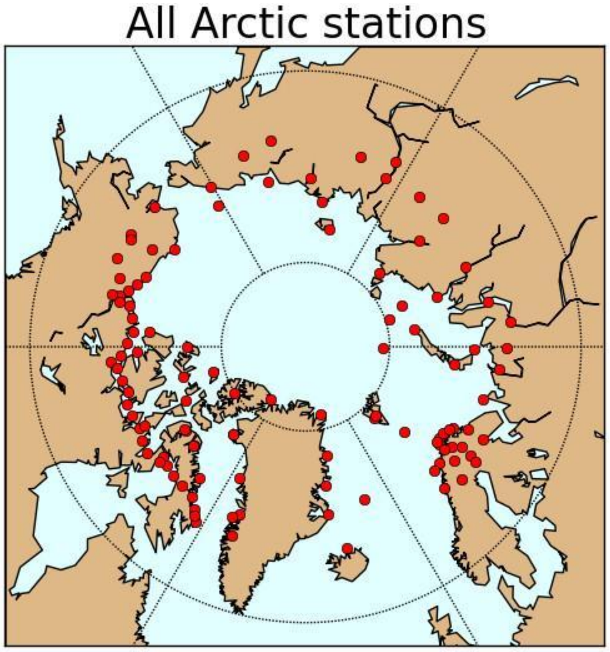

Figure 15: *Map of all Arctic stations in the Global Network.*

Therefore, we might optimistically expect that urbanization bias is unlikely to be a problem for the Global Network there. The Arctic is also a region which has attracted considerable attention in studies of climate change[50]. Hence, in this section we will consider the temperature trends of the Arctic. As can be seen from Figure 15, there are 96 Arctic stations in the Global Network. The mean gridded trends of these stations are shown in Figure 16.

The overall trends since the late 19th century appear to have comprised a warming trend (1900s-1930s), a cooling trend (1940s-1970s) and another warming trend (1980s-2000s). These general patterns are similar to the U.S. Network trends as well as the trends of some of the eight *fully rural* stations discussed in the previous section. Our estimate in Figure 16 broadly agrees with a number of other weather station-based estimates of Arctic trends, e.g., Refs. [51, 52] and references therein. However, much of the apparent agreement is probably due to the considerable overlap in the Arctic weather stations used[52].

There seems to have been an unusually large step increase at about 1920. It is hard to know whether this was due to a genuine climatic shift, a non-climatic bias or both. However, whatever the nature of the 1920 shift, there does appear to be other evidence to *qualitatively* corroborate that the three trends have at least some climatic component. For example, estimates of Arctic sea ice extent constructed from satellite readings suggest an almost continuous decrease since the satellite records began in October 1978[53], suggesting that the there was a genuine Arctic warming during the period 1980s-2000s. According to Callender, 1938's peer review comments, the reduction in Arctic ice extent and increase in air and sea temperatures from the late 19th century until the 1930s was substantial[54], suggesting that Arctic warming also occurred during the 1920s-1940s period. Finally, using several independent data sources, Kukla et al., 1977[55] found a large-scale cooling trend across the Northern Hemisphere from 1950s-1970s.

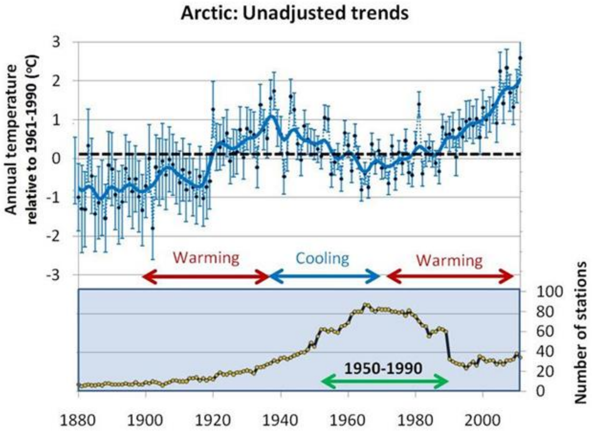

Figure 16: *Average unadjusted gridded trends of all Arctic stations in the Global Network. Solid line corresponds to 11 point binomial smoothed plot of the annual data. Confidence intervals correspond to twice the standard error of the mean.*

Unfortunately, while this appears to qualitatively confirm that the multidecadal trends of Figure 16 are broadly of the correct sign, they cannot be used to assess the exact magnitude of the trends. We saw in Sections 3.1 and 3.2 that similar warming and cooling periods occurred in other parts of the Northern Hemisphere at roughly the same times. This offers further evidence that these were genuine climatic changes. However, we also saw that there is uncertainty as to which of the two warm periods since the late 19th century are warmer (if either), and how the two cool periods compare to each other.

Figure 16 appears to suggest that the recent warm period was the warmer. But, it can be seen from the bottom panel of Figure 16 that the vast majority of the Arctic stations only have data in the 1950-1990 period, i.e., after the 1920s-1930s warm period, and only covering the beginning of the 1980s-2000s warm period. This means that very few of the stations can be used for directly comparing the early and recent warm periods.

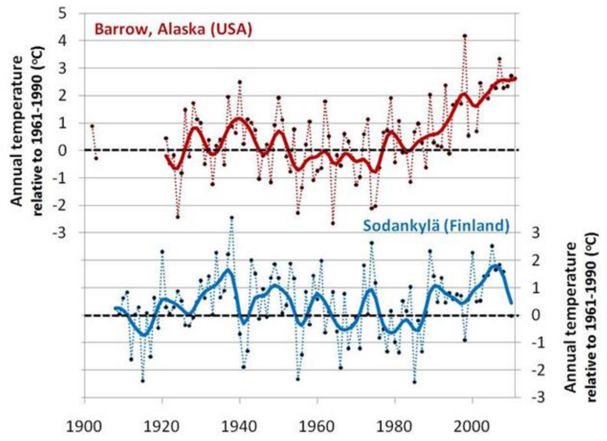

Figure 17: *Unadjusted temperature trends for the two Arctic stations, Barrow, Alaska (USA) and Sodankylä (Finland). Solid lines correspond to 11 point binomial smoothed plots of the annual data.*

Only one of the 96 Arctic stations has data for at least 95 of the last 100 years (Sodankylä), making direct comparisons of temperatures during the two warm periods and the cold period difficult. If we adopt the less restrictive requirement that stations have data for at least 75 of the last 80 years (1933-2012), this still only provides six stations :

1. **Fully rural**: Sodankylä, Finland. 67.37°N, 26.65°E. ID = 61402836000.
2. **Intermediate**: Barrow, Alaska, USA. 71.30°N, 156.78°W. ID = 42570026000.
3. **Intermediate**: Tromo, Norway. 69.50°N, 19.00°E. ID = 63401025000.
4. **Intermediate**: Vardo, Norway. 70.37°N, 31.10°E. ID = 63401098000.
5. **Intermediate**: Bodo Vi, Norway. 67.27°N, 14.37°E. ID = 63401152000.
6. **Fully urban**: Murmansk, Russian Federation. 68.97°N, 33.05°E. ID = 63822113000.

Of these, only the Sodankylä station is *fully rural*. This might be surprising since one might expect the Arctic stations to be mostly rural. However, while the Arctic is a region that is mostly unoccupied by humans, weather stations tend to be located in those areas which are near, or in, human settlements. As a result, urbanization can still be a problem for Arctic stations.

Modern settlements in harsh climatic permafrost regions often make the most of technological advances when they become available, e.g., snow ploughs, building insulation and heating, construction of regional airports. So, even towns with a modest population could have quite a large urban heat island in such regions. For instance, even though Barrow has a relatively small population (4,600 in 2000), it is known to have a substantial urban heat island[56].

Figure 17 shows the unadjusted trends of both Sodankylä and Barrow. Both stations show warming during the 1920s-1930s and 1980s-2000s and cooling during the 1940s-1970s. However, for the rural Sodankylä, *both* warm periods are comparable (and indeed the warmest year was during the earlier period), while for the urbanized Barrow, the 1980s-2000s warming was more pronounced. Since Barrow currently has a significant urban heat island, it is plausible that this urban heat island has developed over the past few decades, in which case at least some of the warming at the Arctic stations is probably due to urbanization bias. In summary, even in the Arctic, urbanization bias is a serious problem.

4 Have the adjustments removed urbanization bias from the adjusted datasets?

Most weather station records were taken with a view to reporting, understanding, and predicting weather, rather than for studying long-term climatic changes. This is why they are called "weather stations", after all. As a result, weather records are subject to a myriad of *non*-climatic changes which could easily bias analyses that are looking for climatic trends using more than a decade or so of data - see Mitchell, 1953[10] for a review.

In an attempt to correct for some of these biases, the National Climatic Data Center have developed a series of adjustments, which they apply to some versions of the U.S. and Global Networks[12–16]. Hence, the user of these datasets is offered the choice between *Unadjusted*, *Partially adjusted* or *Fully adjusted* datasets. These adjustments are commonly referred to as *"homogenizations"*, since they are designed with a view to removing "data inhomogeneities" (i.e., non-climatic biases) from the temperature records. For this reason, the terms "adjustments" and "homogenizations" are often used inter-changeably in the literature.

With the exception of the Karl et al., 1988 adjustments to Version 1 of the U.S. Network[4], none of these adjustments were designed to remove urbanization bias. However, some researchers have claimed that these adjustments *indirectly* remove any (or most) of the urbanization bias from the Historical Climatology Networks, e.g., Refs. [15, 31, 57].

If this claim were valid then the presence of urbanization bias in the *Unadjusted* datasets would not be a major concern for users of the *homogenized* datasets. As can be seen from Table 1, many of the groups using the Historical Climatology Networks currently use the homogenized versions. With this in mind, in this section we will assess the validity (or otherwise) of this claim. To do so, it is important to understand some of the theoretical basis behind the various approaches to homogenizing temperature records. Some readers will already be familiar with some of the concepts in this section.

Applying adjustments in try and correct for biases is a highly challenging topic, e.g., see Refs. [58–61]. All homogenization techniques face two main problems:

1. An adjustment technique may mistakenly treat a genuine trend as a bias, or overestimate the magnitude of an actual bias, thereby introducing artificial biases (*"false positives"*, or *"Type I errors"*)[62].

2. An adjustment technique may fail to accurately identify, or fail to completely remove, individual biases (*"false negatives"*, or *"Type II errors"*)[63].

The second problem could lead an unwary researcher into a false sense of confidence in their data, while the first problem could lead them to misinterpreting artificial homogenization adjustments as being part of genuine climatic trends. Hence, it is important to critically assess both the necessity of applying such techniques and the reliability of any homogenization techniques which are used.

In this section, we will discuss the various homogenization approaches used by the National Climatic Data Center for different versions of the Historical Climatology Network datasets. In Section 4.1, we will first summarise what adjustments have been made by the National Climatic Data Center to the different versions of the two datasets (both past and present). They have used two main types of adjustments - (1) statistically-based bulk adjustments and (2) station comparison-based adjustments. Several adjustments of the first type have been applied to the U.S. Network, and we will discuss these adjustments in Section 4.2. The second type of adjustments have been applied to both networks, and we will discuss these in Section 4.3.

4.1 Types of adjustments applied to the Historical Climatology Networks

So far, the National Climatic Data Center have compiled three versions of the Global Network and two versions of the U.S. Network. With the exception of Version 1 of the Global Network, they have provided users of the networks with a choice of datasets, each with a different set of adjustments applied. The types of adjustments applied to both networks have changed over the years, from version to version. Table 3 summarises the adjustments applied to each version.

In this paper we will refer to the raw datasets without any adjustments applied as *"Unadjusted"*. Since the National Climatic Data Center only apply one set of adjustments to the Global Network, we will simply refer to the dataset with adjustments applied as the *"Adjusted"* dataset. However, for the U.S. Network, they apply several different sets of adjustments. For this reason, they provide two adjusted datasets. In the first adjusted dataset, the only adjustments they apply are for changes in *"Time-of-Observation"* (see Section 4.2.1). We will refer to this dataset as the *"Partially adjusted"* dataset. The second adjusted dataset includes these Time-of-Observation adjustments, but also includes several others. We will refer to this dataset as the *"Fully adjusted"* dataset.

Figure 18 shows the gridded mean temperature trends for the current version (Version 3) of the Global Network using the *Unadjusted* (top) and *Adjusted* (bottom) datasets. At first glance, both trends seem similar. However, the adjustments do slightly alter the trends.

The gridded net effect of the homogenization adjustments is shown in Figure 19. It can be seen that the homogenization adjustments introduce a slight warming trend into the dataset. As a result, the recent temperatures seem "warmer" relative to the late 19th century in the *Adjusted* dataset than in the *Unadjusted* dataset.

Figure 20 shows the mean gridded temperature

Network	Adjustments	Q.C.	Step change biases				Trend biases		Infill
			TOB	MMTS	Doc.	Undoc.	UHI	Other	
Global ver. 1 (1992)[11]	Unadjusted	√							
Global ver. 2 (1997)[12]	Unadjusted	√							
Global ver. 2 (1997)[12]	Adjusted	√				√[64]			
Global ver. 3 (2011)[13]	Unadjusted	√							
Global ver. 3 (2011)[13]	Adjusted	√				√[15]			
U.S. ver. 1 (1996)[14]	Unadjusted	√							
U.S. ver. 1 (1996)[14]	Partially adjusted	√	√						
U.S. ver. 1 (1996)[14]	Fully adjusted	√	√	√	√[34]		√		√
U.S. ver. 2 (2010)[15]	Unadjusted	√							
U.S. ver. 2 (2010)[15]	Partially adjusted	√	√						
U.S. ver. 2 (2010)[15]	Fully adjusted	√	√		√[15]	√[15]			√

Table 3: *Adjustments applied to records in each of the different versions of the Historical Climatology Network datasets. The year of first release and version number are shown. Q.C. = Quality Control check; TOB = Time of OBservation adjustments[35]; MMTS = adjustments for transition to electronic Maximum-Minimum Temperature System[39]; Doc. = Documented station changes; Undoc. = Undocumented station changes; UHI = Urban Heat Island adjustments[4]; Infill = interpolation of missing periods of station records[14].*

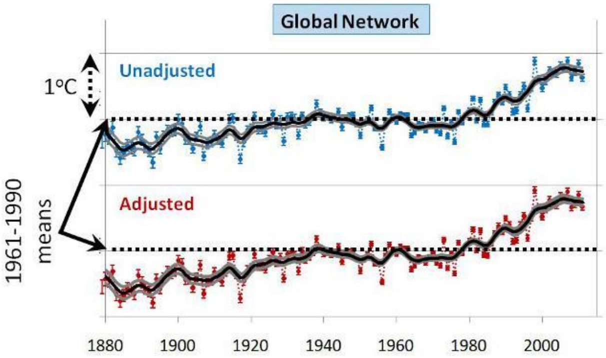

Figure 18: *Gridded mean temperature trends and confidence intervals for the Global Network before (top) and after (bottom) homogenization. Solid black (grey) lines are 11 point binomial smoothed versions of the annual means (confidence intervals).*

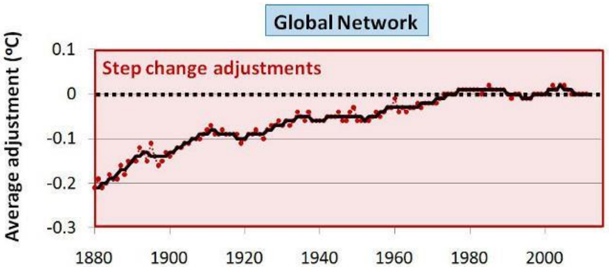

Figure 19: *The gridded mean adjustments applied to the Global Network by the National Climatic Data Center.*

trends for the current version (Version 2) of the U.S. Network using the three different datasets, i.e., *Unadjusted*, *Partially adjusted* and *Fully adjusted*. Unlike the Global Network, the effects of the U.S. Network adjustments on long-term trends are quite pronounced.

As we discussed in Section 2.1, the *Unadjusted* dataset suggests that temperatures during the 1920s/1930s were at least as warm as the 1990s/2000s. However, most of the adjustments that the National Climatic Data Center apply to the U.S. Network have the net effect of adding a warming trend. As a result, in the *Partially adjusted* dataset, the 1920s/1930s warm period seems cooler and the 1990s/2000s warm period seems warmer. The additional adjustments applied to the *Fully adjusted* dataset also do the same. This leads to the widely-quoted claim that recent U.S. temperatures are the "hottest on record"[65–67]. We can see from Figure 20 that this claim does **not** hold for the *Unadjusted* dataset, but only for the two adjusted datasets.

Figure 21 shows the annual mean gridded adjustments applied to the U.S. Network by the various homogenization algorithms. The *Time-of-Observation* adjustments introduce a warming trend of about $+0.19°C/$century, and the rest of the adjustments introduce an additional warming trend of about $+0.16°C/$century.

In the following sections, we will assess the reli-

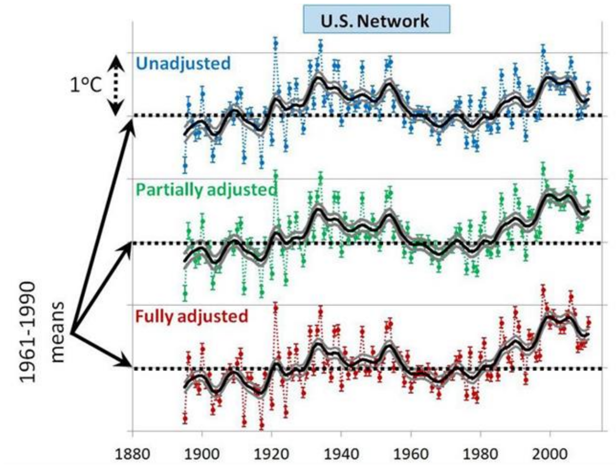

Figure 20: *Gridded mean temperature trends and confidence intervals for the U.S. Network after each stage of homogenization. Solid black (grey) lines are 11 point binomial smoothed versions of the annual means (confidence intervals).*

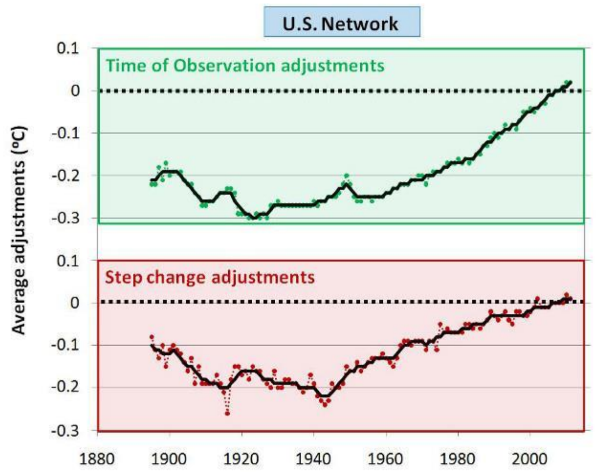

Figure 21: *The gridded mean adjustments applied to the U.S. Network by the National Climatic Data Center.*

ability (or otherwise) of the various homogenization adjustments applied to the two networks. There have been two main classes of homogenization approaches taken:

1. Statistically-averaged *bulk* adjustments
2. *Individual* station adjustments based on statistical comparisons with neighbouring stations.

The first class of adjustments are usually determined by calculating the average biases introduced by a specific bias-causing phenomenon. These average values are then subtracted from the station records which are known to be affected by that phenomenon. For example, Quayle et al., 1991 calculated the mean biases introduced to a sample of several hundred stations when their thermometers were changed from liquid-in-glass thermometers to electronic "Maximum-Minimum Temperature Systems"[39]. If it was known that a station underwent this transition at a particular time, then the Quayle et al. adjustment could be applied to the station record.

The second approach to homogenizing station records is to estimate the non-climatic biases by comparing each record to its neighbouring stations. This approach allows for station-specific adjustments, and as a result has formed the primary homogenization method used by the National Climatic Data Center for both the Historical Climatology Networks.

As these two approaches to homogenizing the temperature records are quite distinct, and both approaches have been used by the National Climatic Data Center in homogenizing the Historical Climatology Networks, we will now consider them separately. In Section 4.2, we will discuss the various statistically-based bulk adjustments that have been used in different versions of the U.S. Network, while in Section 4.3, we will discuss the station-comparison based adjustments used in different versions of both datasets.

4.2 Statistically-based bulk adjustments applied to the Historical Climatology Networks

Station history information is generally needed to apply statistically-based bulk adjustments to a temperature dataset, so that it can be decided for what portions of the records, the adjustments need to be applied. Therefore, since the National Climatic Data Center do not have a station history file for the Global Network, they do not apply any of these adjustments to the Global Network. However, at different stages, they have applied up to three sets of bulk adjustments to the U.S. Network:

1. Karl et al., 1986's adjustments for changes in *Time-of-Observation*[35]
2. Karl et al., 1988's population growth-based urbanization bias adjustments[4]

3. Quayle et al., 1991's adjustments for a known change in instrumentation which has occurred at many U.S. Network stations since the 1980s[39]

For Version 1 of the U.S. Network, they applied all three adjustments[14]. However, with Version 2, they dropped the Quayle et al., 1991 and Karl et al., 1988 adjustments[15], and so currently, the only bulk adjustment they use are Karl et al., 1986's *Time-of-Observation* adjustments.

4.2.1 Karl et al., 1986 adjustments for changes in Time-of-Observation

Most observers in the U.S. Network estimate the daily average temperature by using a maximum-minimum thermometer. These thermometers record the maximum and minimum temperatures that have occurred since the thermometer was reset. If the thermometer is reset once a day, then the mean temperature of the preceding 24 hours can be approximated as the average of the maximum and minimum. This is quite a crude approximation, but it does not require much effort on the part of the observer. As a result, it has been a very popular approach amongst weather observers, particularly before the invention of automated electronic thermometers.

Unfortunately, the mean temperatures calculated in this way are strongly influenced by the time of the day in which the observer resets the thermometer. This can be seen by considering Figure 22 and Table 4. Figure 22 shows the temperatures recorded every minute at the Ames, Iowa (U.S.) automatic weather station over an arbitrarily-chosen five-day period. Table 4 lists the different "mean temperatures" for those five days using different averaging methods. All of the methods provide different estimates relative to the mean derived from the average of all the measurements in the calendar day ("All data").

For this reason, if an observer changes the time of day during which they reset their thermometer (i.e., the *"Time-of-Observation"*), this can introduce a non-climatic shift in the record. The possibility that observation time could bias weather records has been known since at least the 19th century, e.g., Ellis, 1890[68], and early 20th century, e.g., Refs. [69, 70].

Several researchers have tried to develop methods to estimate and reduce this bias, e.g., see Refs. [35, 71, 72] and references therein. In particular, Karl et al., 1986 developed a series of *Time-of-Observation* adjustments for the contiguous U.S.[35]. The National Climatic Data Center decided to use these adjustments to correct the U.S. Network dataset for any documented *Time-of-Observation* changes in the station history file[14]. We saw the net effect of these adjustments on U.S. temperature trends in Figure 21.

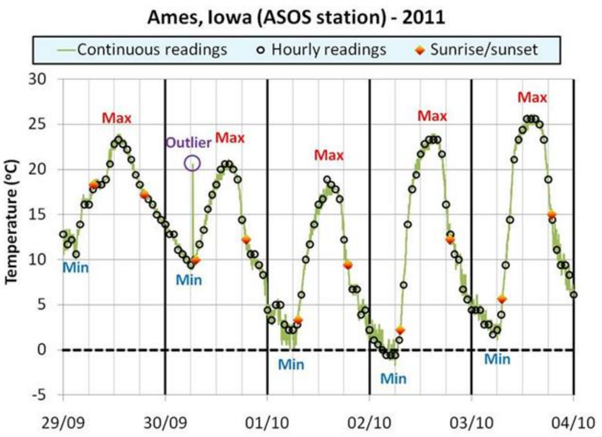

Figure 22: *1-minute interval temperature measurements over a 5-day period (29/09/2011-04/10/2011) at the Ames, Iowa (US) automatic airport weather station (located at 41.99206° N, 93.6218° W). The station is part of NOAA NWS's Automated Surface Observing System, but the data was obtained from the Iowa State University website. Measurements taken exactly on the hour (00:00, . . . , 23:00) and the times of local sunrises and sunsets are also indicated. "Outlier" corresponds to an anomalous measurement which occurred at 06:31 (Local Solar Time) on 30/09/2011. Its value of 69° F (∼20.6° C) was immediately preceded and followed by more than 30 minutes of temperatures in the range 48-50° F (∼8.9-10.0° C), and so was probably an instrumental or measurement error.*

A detailed evaluation of the *Time-of-Observation* biases is beyond the scope of this article, but as the magnitude of the National Climatic Data Center's *Time-of-Observation* adjustments is quite substantial, a brief discussion is relevant.

Figure 23 shows the relative occurrence of different reported observation times for the U.S. Network, according to the National Climatic Data Center's published station history files. At the time of writing, the National Climatic Data Center only seemed to have a public archive for the 1996 version of the U.S. Network station history files, i.e., the one archived by Easterling et al., 1996[14]. Hence, the station histories in Figure 23 only cover the period 1895-1996. Nonetheless, it is clear from Figure 23, that there

Method	**29/09**		**30/09**		**01/10**		**02/10**		**03/10**		**Mean**
	Mean	Bias	Mean	Bias	Mean	Bias	**02/10**		Mean	Bias	**bias**
All data	17.57		13.63		8.80		11.02		13.20		
Instantaneous measurements:											
24-daily	17.36	-0.21	13.85	+0.22	9.05	+0.25	10.81	-0.21	13.16	-0.04	+0.00±0.20
4-daily	17.50	-0.07	14.00	+0.37	8.73	-0.07	9.85	-1.17	12.48	-0.72	-0.33±0.54
3-daily	18.75	+1.18	13.75	+0.12	10.13	+1.33	12.50	+1.48	14.58	+1.38	+1.10±0.50
Maximum-minimum thermometer measurements:											
00:00	15.85	-1.72	16.95	+3.32	9.45	+0.65	11.10	+0.08	13.60	+0.40	+0.55±1.62
07:00	15.00	-2.57	16.40	+2.77	10.55	+1.75	11.95	+0.93	11.10	-2.10	+0.16±2.12
07:00 shift	16.95	-0.62	15.00	+1.37	10.55	+1.75	13.35	+2.33	N/A	N/A	+1.21±1.28
17:00	16.95	-0.62	15.00	+1.37	10.55	+1.75	9.45	-1.57	12.50	-0.70	+0.05±1.29
Sunset	16.95	-0.62	15.00	+1.37	9.45	+0.65	11.10	+0.08	11.10	-2.10	-0.12±1.19

Table 4: *Various estimates of the mean daily temperatures (in °C) for the data in Fig. 22, using different observation practices, and their associated biases relative to the mean of "All data" (in °C). "All data" corresponds to the simple mean of all 1-minute measurements over the 00:00-23:59 period. "24-daily" is the mean of all 24 hourly observations for each day. "3-daily" is the mean of the 07:00, 14:00 and 19:00 observations, with 19:00 receiving twice the weighting of the other observations. "4-daily" is the mean of the 00:00, 06:00, 12:00 and 18:00 observations. The observation practice for the remaining rows involves resetting the thermometer at the indicated time. For "07:00 shift", the value for the maximum temperature was taken from the following day's measurements. The column on the far right is the mean bias averaged over all five days ± twice the standard error of the mean.*

have indeed been pronounced changes in the average reported *Time-of-Observation* over this period. In particular, there has been a general decrease in the number of "evening" observers (17:00, 18:00, 19:00 and sunset), which is matched by a general increase in the number of "morning" observers (7:00 and 8:00).

As can be seen from Figure 21, the National Climatic Data Center's calculations suggest that this evening-to-morning transition has introduced a "cooling" bias to recent measurements (or alternatively a "warming" bias to earlier measurements). Balling & Idso, 2002 were sceptical of the reliability of these adjustments, since other estimates of U.S. temperature trends (e.g., satellite measurements) did not show this extra warming trend[73]. In response, Vose et al., 2003 revisited the Karl et al., 1986 adjustments[74]. However, they found the adjustments to be reliable, and also similar to an independent assessment method developed by deGaetano, 1999[75]. For this reason, the National Climatic Data Center decided to keep using the Karl et al., 1986 adjustments.

We do understand the cynicism of Balling & Idso, 2002[73] and others, e.g., Watts et al., (in preparation, 2012)[76] about these *Time-of-Observation* adjustments, since it is surprising that most of the National Climatic Data Center's homogenization attempts should coincidentally be in the same direction (i.e., to introduce "more warming" into the records). This cynicism is probably not helped by the fact that (at the time of writing), the National Climatic Data Center had not updated their publicly archived station history file in 17 years. It also took us considerable time to track down this 17-year old history file - one version is currently available at `http://cdiac.ornl.gov/ftp/ndp019/`. Perhaps if the National Climatic Data Center published a more recent station history file and made it more accessible, this would reduce some of the cynicism over their *Time-of-Observation* adjustments.

Nonetheless, our preliminary calculations (not shown) do suggest that such a transition would indeed introduce biases similar to those calculated by the National Climatic Data Center. So, we agree that, if the published station history file is accurate, *and* if the station records have not already had *Time-of-Observation* corrections applied to them[5], then the National Climatic Data Center's *Time-of-Observation* corrections are probably reasonable.

We also note that changes in observation time are a problem which is not just confined to the U.S. Net-

[5] For some station records, the archived data from which the Global Network was compiled may have already had been adjusted for changes in *Time-of-Observation*.

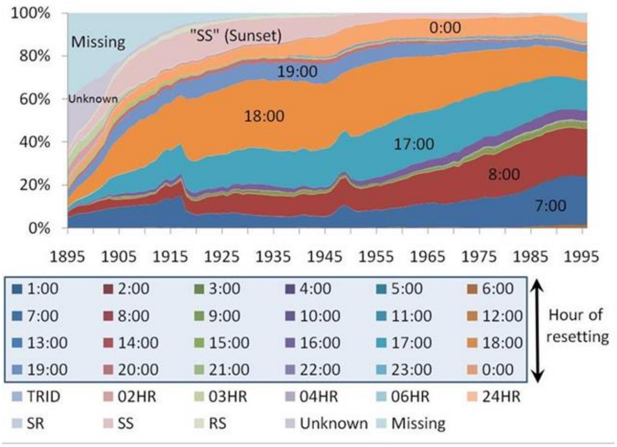

Figure 23: *Relative occurrence of different reported observation times (and averaging methods) for the U.S. Network, according to the National Climatic Data Center's station history file. The only copies we could find of this history file were those from Easterling et al., 1996[14], which only cover the period up to 1996.*

work, e.g., see our discussion of *Time-of-Observation* in Ref. [48]. No station histories were collected for the Global Network, so we do not know what *Time-of-Observation* biases exist for the Global Network. But, it seems reasonable to assume that biases of similar magnitudes exist for other parts of the world - particularly during the recent widespread shift to automated weather systems. These biases might be of either direction, i.e., they might have introduced cooling *or* warming biases to regional trends, or possibly both.

Until station histories are collected for the Global Network, it is probably not possible to reliably estimate the *Time-of-Observation* biases in the global trends. But, since they are available for the U.S. Network, it is worth seeing if the National Climatic Data Center's adjustments alter the urban-rural differences we identified in Section 3.1. Figure 24 compares the trends of the *fully rural* and *fully urban* subsets for the *Time-of-Observation* adjusted U.S. Network, i.e., the *Partially adjusted* dataset.

Comparing Figure 24 to Figure 9, we can see that the *Time-of-Observation* adjustments actually reduce the divergence between the urban and rural subsets by about 0.2°C/century. This suggests that *some* of the apparent divergence between the subsets is a result of different patterns in station observer behaviour, and not just urbanization bias. Peterson, 2003 had suggested that this was the case[77]. However, we note that even after the adjustments, there is still a substantial difference between the subsets (about 0.5°C/century), and as we saw in Figure 9, the diverging trend does seem to be similar to U.S. urban growth since the 19th century. So, unlike Peterson, 2003, we suggest that much of the divergence probably *is* due to urbanization bias.

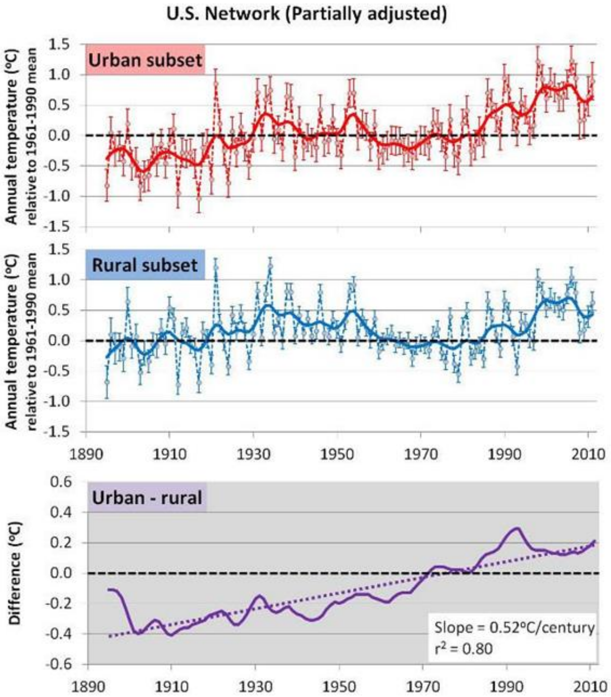

Figure 24: *Mean gridded trends of the* fully urban *and* fully rural *U.S. Network subsets of Figure 7 using the* Partially adjusted *dataset. Solid lines correspond to the 11 point binomial smoothed versions of the annual values. Confidence errors correspond to twice the standard error of the annual means. The bottom panel shows the difference between the two subsets (the smoothed versions)*

If we assume that (a) the *Time-of-Observation* adjustments improve the reliability of the dataset and (b) the rural subset is a more accurate representation of U.S. temperature trends than the urban subset, then this suggests that the recent warm period was slightly warmer than the early 20th century warm period, even if the hottest years were in the early period (1934 and 1921). However, we note in Ref. [48] that the *Time-of-Observation* adjustments actually increase the siting bias of badly-sited stations in the U.S. Network. So, it is possible that some of the warming trend introduced by the adjustments should be counteracted by adjustments to account for poor station siting.

Nonetheless, while the claim that recent U.S. temperatures are the "hottest on record"[65–67] appears justified in the *Partially adjusted* dataset by the urban subset, it is not justified by the rural subset. This suggests that much of the apparently unusual warmth of recent U.S. temperatures is a result of urbanization bias. Claims that the recent warm period is a consequence of anthropogenic global warming from increasing atmospheric carbon dioxide[65–67] should be treated cautiously.

4.2.2 Karl et al., 1988 adjustments for urbanization bias

As we discussed in Paper I[1], quite a few studies have found evidence of urbanization bias in the various estimates of U.S. temperature trends, e.g., Kukla et al., 1986[78] or Cayan & Douglas, 1984[79]. It was partly to address this concern that Karl et al., 1988 decided to create the U.S. Network[4].

When they were constructing the network, they actively tried to select stations which were not urbanized. However, since they wanted the network to contain a high density of stations from all regions of the contiguous U.S., and to make sure the records were relatively long and complete, they were forced to include some partially urbanized stations[4]. In an attempt to remove the urbanization bias that these stations might have introduced to the network, they developed a population-based adjustment to correct for urbanization bias[4].

Since the U.S. Census Bureau have been carrying out regular decadal censuses of the U.S. for a few centuries[30], the National Climatic Data Center were able to estimate the population growth associated with each of the stations in the U.S. Network. So, for the *Fully adjusted* dataset of Version 1, they applied Karl et al., 1988's urbanization bias adjustments to all stations, using the census population figures.

As an aside, it is important to note that these urbanization bias adjustments were carried out *after* the Karl & Williams, 1987 station comparison-based adjustments which we will discuss in Section 4.3. The rationale for this was that the Karl et al., 1988 adjustments were bulk statistical adjustments, which were based entirely on local population growth. Karl et al., 1988 had cautioned that these adjustments were probably only reliable when averaged over many stations, and that individual station adjustments might be inappropriate[4]. Hence, it was thought that they might interfere with the station comparisons in the Karl & Williams, 1987 adjustments. This is a valid argument. However, it does mean that, when the Karl & Williams, 1987 algorithm was being applied to the U.S. Network, the urban stations would have had *no* urbanization bias corrections applied. A consequence of this is that the algorithm would have been strongly affected by the "urban blending" problem which we will discuss in Section 4.3.3.

Unfortunately, population growth is only an approximate indicator of urbanization. Karl et al., 1988 recognized this, but at the time, it was probably the best metric of urbanization available. Still, in the years since then, several other urbanization metrics have become available, particularly with advances in satellite technology, e.g., the GRUMP dataset[80]. This has allowed several researchers to test how reliable the Karl et al., 1988 adjustments were.

Hansen et al., 2001 proposed an alternative adjustment for the U.S. Network using night-light intensity as their urbanization metric. They found that Karl et al., 1988's population growth-based adjustment had only removed some of the urbanization bias ($\sim 0.06°C$/century) for the U.S. Network. In contrast, Hansen et al., 2001's night light-based adjustments removed more than twice that ($\sim 0.15°C$/century)[81].

Kalnay & Cai, 2003[32] suggested that the problem was even more serious, since they estimated that urbanization (and changes in land use) had introduced a warming bias of about $0.27°C$/century to U.S. temperature trends. Although, as we discuss in Paper I[1], the Kalnay & Cai, 2003 study appears to have been quite controversial, and the subject of considerable debate.

As we discussed in Section 3, both of the Historical Climatology Networks are likely to be strongly affected by urbanization bias. So, despite the inadequacies of the Karl et al., 1988 adjustments, they do seem to have been an admirable attempt to reduce the extent of the problem for the U.S. Network.

Hence, it is surprising that when Menne & Williams first began work on Version 2 of the U.S. Network, they decided they would just attempt to correct for step change biases. They discarded Karl et al.'s urbanization adjustment, and decided *not* to intentionally remove biases from urban heat islands:

> "*...we still keep some of the low frequency temperature variations that are in almost all stations, when you look at them with respect to regional trends, like urban heat is-*

lands, but that may not necessarily be a bad thing." - Claude N. Williams, Jr.; 31 January 2006[82].

However, they noticed that their step-change adjustments occasionally inadvertently removed some urban heat island bias. Initially, they regarded this as a problem[82], since they seemed to believe urbanization biases were a desirable feature. Considering the discussion above, this may at first seem surprising. The explanation is that urbanization biases are only "biases" if one is attempting to calculate regional (or global) changes in climate. The urban heat island in Phoenix, Arizona (USA), for example, is real, substantial, and affects those living and working in the area[83]. So, if you are studying the *local* climate of Phoenix, the localized urban heat island is not a "bias", but an accurate description of the local climate.

Of course, if researchers are interested in studying regional (or global) temperature trends, these localized urban heat islands would artificially bias their studies. Indeed, this was the reason why Karl et al., 1988 had introduced their urbanization adjustment to Version 1[4]. Perhaps for this reason, by the time Menne et al., 2009 was published, they had decided that the occasional inadvertent removal of some urban heat island bias through their adjustments was a feature[15]. Indeed, they suggested that this inadvertent removal accounted *"for much of the changes addressed by the Karl et al. (1988)* [urbanization bias]*correction used in* [Version 1 of the U.S. Network]*"*[15].

In Section 4.3.3, we will consider, in detail, whether the Menne & Williams, 2009 algorithm does actually account for the urbanization bias problem or not. But, we can carry out a quick check by considering Figure 25.

Figure 25 compares the urban and rural subsets of the *Fully adjusted* U.S. Network (Version 2). We can see from the bottom panel that the Menne & Williams, 2009 adjustments have slightly reduced the apparent divergence between the two subsets, compared to the *Partially adjusted* dataset (Figure 24), i.e., the divergence has decreased from $\sim 0.5°C$/century to $\sim 0.3°C$/century.

Initially, this might appear to validate the claim that the Menne & Williams, 2009 algorithm does remove (or at least reduce) urbanization bias. However, there are two major problems with the claim:

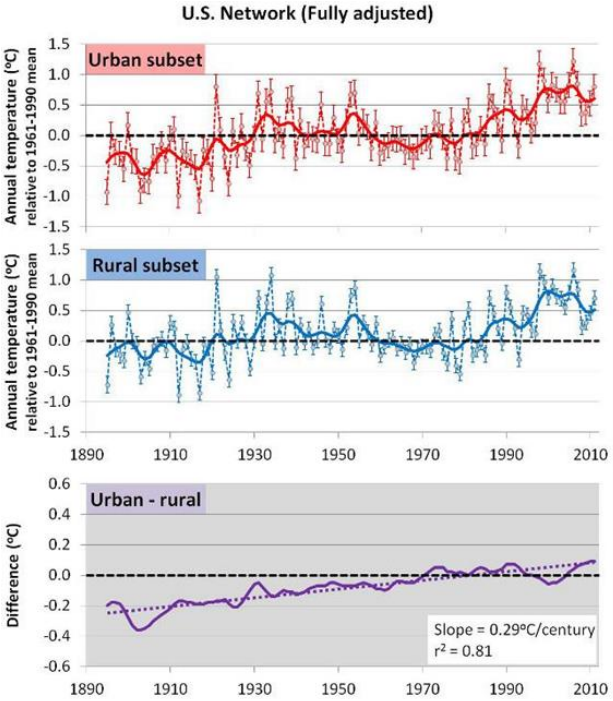

Figure 25: *Mean gridded trends of the* fully urban *and* fully rural *U.S. Network subsets of Figure 7 using the* Fully adjusted *dataset. Solid lines correspond to the 11 point binomial smoothed versions of the annual values. Confidence errors correspond to twice the standard error of the annual means. The bottom panel shows the difference between the two subsets (the smoothed versions)*

1. The divergence was only reduced by $\sim 0.2°C$/century, i.e., there is still a divergence of $\sim 0.3°C$/century.
2. As we will discuss in Section 4.3.3, the Menne & Williams, 2009 algorithm is seriously affected by the "urban blending" problem, whereby the temperature trends of rural stations in urban areas are adjusted upwards to better match those of their urban neighbours.

4.2.3 Quayle et al., 1991 adjustments for changes in instrumentation

Changes in the type of thermometer used for taking measurements at a weather station can often introduce non-climatic biases[15, 39, 40, 77, 84–87]. So, if there have been any widespread changes in instrumentation, this could significantly bias regional temperature trends.

Quayle et al., 1991 noted that, during the 1980s, a large number of stations in the U.S. replaced their Liquid-In-Glass (LIG) thermometers (housed in Cotton Region Shelters) with electronic, Maximum-Minimum Temperature Systems (MMTS)[39]. With this in mind, they estimated the mean bias associated with the transition from the liquid-in-glass thermometers to the electronic systems, by studying a sample of several hundred stations[39]. They calculated that this transition introduced a cooling bias of $-0.1°C$.

For Version 1 of the U.S. Network, the National Climatic Data Center used the Quayle et al., 1991 estimates to statistically adjust any stations with a documented change between these two thermometer systems. These adjustments had the effect of slightly increasing the values of recent (post-1980s) temperatures in the dataset (see Hansen et al., 2001[81]).

However, Pielke et al., 2002 noticed that the Quayle et al. adjustments were sometimes applied to stations which still used liquid-in-glass thermometers and argued that while Quayle et al.'s calculations might be accurate on average, instrumental biases varied from station to station[84]. Hence, both Hubbard & Lin, 2006[40] and Pielke et al., 2007[88] revisited the Quayle et al. calculations. Both studies confirmed that the bias associated with the switch from the Liquid-In-Glass thermometers to the electronic systems varied substantially from station to station. They recommended against applying a single statistical bulk adjustment to all stations.

Peterson et al., 2003 also noted that several stations used other temperature-measuring systems, e.g., hygrothermometers and hygrothermograph[77]. Station changes involving any of these systems could also introduce non-climatic biases. So, it was insufficient to *only* consider the instrumentation biases caused by the Liquid-In-Glass to Maximum-Minimum Temperature System transition.

For these reasons, when switching to Version 2, the National Climatic Data Center stopped applying specific adjustments for instrumentation change, and relied on their general step-change homogenization procedure to identify the appropriate adjustments[15]. They found that the magnitude of the larger adjustments calculated in this way were comparable to Hubbard & Linn, 2006's findings, but because their homogenization procedure is less effective at detecting small biases, the total number of stations adjusted was considerably reduced. Hence, the effective mean instrumentation adjustment is substantially less than that of Version 1[15].

In any case, the net magnitude of the Quayle et al., 1991 adjustments is relatively small. So, whether or not the adjustments are applied, this should not majorly alter the effects of urbanization bias on the U.S. Network.

4.3 Station comparison-based adjustments applied to the Historical Climatology Networks

Station comparison-based adjustments have been applied to the *Fully adjusted* versions of both networks. However, the National Climatic Data Center have used different algorithms for these adjustments at different stages. As can be seen from Table 3, they did not apply any adjustments (aside from a brief Quality Control check) to Version 1 of the Global Network. For Version 1 of the U.S. Network, they used the Karl & Williams, 1987 algorithm[34], while for Version 2 of the Global Network, they used the Easterling & Peterson, 1995 algorithm[12] (original Refs. [64, 89]). Currently, they have switched to using the same algorithm for both networks, i.e., the Menne & Williams, 2009 algorithm[49].

Although there are some important differences between these algorithms, there are a lot of similarities between them, and there are common weaknesses between all three algorithms, which we will discuss in Section 4.3.3. However, before we can assess any of the algorithms, it is important to review some of the theoretical basis behind station comparison-based homogenization algorithms.

4.3.1 Types of non-climatic biases

There are three main types of biases that are problematic for temperature records, and they each have different properties:

1. "Accounting errors", e.g., incorrectly entered temperature values or once-off errors[12, 13, 15].

2. "Step change" biases due to some *abrupt* change in the station environment or recording practices, e.g., station move[34, 90–92], change of instrumentation[15, 39, 40, 77, 84–87], new *Time-of-Observation*[35, 74, 93], the cutting down of trees in the vicinity of the station[94, 95].

3. "Trend" biases due to a *gradual* change in the station environment or the station equipment, e.g., increasing urbanization or changes in land use of surrounding area[96, 97], growth of trees in

the vicinity of the station[94, 95], gradual degradation of station equipment[94].

Of the three types of bias, the first is relatively easy to identify and correct or remove, and all of the Historical Climatology Network datasets have undergone a preliminary *Quality Control* check to remove the more obvious non-climatic outliers[11–15]. Identifying an individual monthly temperature value as being "non-climatic" or not, is, of course, a subjective process. For instance, if you decided that any measurements which are more than 3 standard deviations from the climatic mean for that station and month, then you would remove most of the unusually hot or unusually cold values. But, if the measurements were all genuinely climatic, but normally distributed, you would then be incorrectly discarding about 0.27% of genuinely climatic measurements. For this reason, the National Climatic Data Center currently use a hierarchical series of tests, in an attempt to reduce the subjectivity in this process - see Menne et al., 2009 for the U.S. Network tests[15] and Lawrimore et al., 2011 for the Global Network tests[13].

The second and third types of bias can affect quite long portions of a station's record. For this reason, they are generally more important to identify, yet simultaneously harder to identify.

Step changes will affect all readings after the change by a similar amount. Therefore, in theory, if the time of the change (usually referred to as a *"break point"* or a *"change-point"*) can be identified, the record can probably be corrected (or *"homogenized"*) by adjusting all temperatures before (or alternatively after) the break point by that amount. In practice, this is not so trivial.

Temperature variations due to natural variability in the weather are often quite substantial from month to month, and even year to year. Moreover, some step-biases may have different influences under different conditions. For example, tall trees in the vicinity of a station may shelter the station from winds in particular directions, or may shade the station from direct sunlight. If the trees are cut down, then the amount that this would influence the recorded temperatures in a given month or year may depend on the prevailing wind directions and cloud cover during that period.

So, it can be difficult to identify how much (if any) of the difference immediately before and after the break point is due to the change and not just natural variability. If there were no long-term climatic trends at the station, then this problem could be overcome by calculating the average temperatures over a longer time-scale (e.g., the five years before and after the break point). This would improve the signal-to-noise ratio, since much of the natural variability would cancel out over the longer periods. However, if there are long-term climatic trends at the station (which is in fact typically what the records are being evaluated for), this approach would remove some of the climatic trend from the record. It also assumes there are no additional break points or trend biases during the period being averaged, which is often not the case.

Trend biases have quite different properties. In particular, they do not necessarily have a single "breakpoint". For instance, if the area around a station is undergoing urbanization, the urbanization bias would be continuously increasing occur over the entire record of the station.

Some trend biases might have a specific "start" and/or "end" point, e.g., if growing trees near the station introduced a trend bias, then this bias would disappear once the trees were cut down. But, other biases might continue over the entire record.

Trend biases are often approximated by assuming they are linear, e.g., Refs. [98, 99]. This is a reasonable first approximation, and relatively simple to model. But, clearly, it may be overly crude for many actual trend biases. For instance, in the case of urbanization bias, urban growth may go through periods of rapid development or slow development depending on economic activity, local migration patterns, etc. In an attempt to introduce a more flexible model for a trend bias, Hansen et al. use a bi-linear model for their urbanization bias adjustments[81, 100]. However, as we discuss extensively in Paper II[2], there are a number of serious problems with the Hansen et al. model (both in the model itself, and in how it is implemented).

4.3.2 Theoretical basis of step-bias adjustments

Surprisingly, although many non-climatic biases (especially urbanization bias) are better described as a trend bias, there seems to have been very little research into homogenization methods for removing trend biases. Instead, most of the station-comparison homogenization methods used on climate records seem to focus on removing step biases. The only exceptions we could find were Alexandersson & Moberg's homogenization techniques that were developed for analysing long-term Scandinavian temperature records[98]; Vincent et al.'s homogenization of Canadian temperature records[99, 101, 102]; and the Goddard Institute of Space Studies urbanization bias adjustments[25, 81, 100] which we discuss in Paper II[2]. It is possible that we have overlooked some other approach, but it can be seen from Table 3, that *none* of the station-comparison adjustments in any of the versions of the National Climatic Data Center's Historical Climatology Networks are for trend biases.

On step-bias homogenization techniques there is currently a considerable body of work, e.g., see reviews by Peterson et al., 1998[103]; World Meteorological Organization (WMO), 2003[58]; Ducré-Robitaille et al., 2003[59]; deGaetano, 2006[60]; Reeves et al., 2007[61]; Costa & Soares, 2009[104]; Domonkos, 2011[105] or Venema et al., 2012[106]. This work has been very useful to the climate community, and should be condoned, with continuing research encouraged. However, we strongly recommend that the community should *also* be simultaneously working on homogenization techniques for trend biases, particularly in light of the urbanization bias problem which we are discussing in this current series of papers.

Nonetheless, it has been frequently claimed that the step-bias homogenization techniques that the National Climatic Data Center have used on the Historical Climatology Networks are somehow able to remove the trend biases caused by urbanization, e.g., Refs. [15, 31, 57]. This is a surprising claim, which does not appear to have attracted much scrutiny until now, aside from Pielke et al., 2007[88]. Hence, we will now evaluate the reliability of this claim. To do this, let us first consider the theoretical basis behind the step-bias homogenization algorithms.

A common approach to homogenizing temperature records is to make use of *difference series*. All three of the step-bias homogenization algorithms so far used by the National Climatic Data Center do this. Difference series are simply the series constructed by subtracting the temperature values for the neighbour records from the temperature record of the station being homogenized (the *"target series"*). However, since there is usually more than one neighbour, there are several ways to do this. Two methods which have been particularly popular are,

1. Reference series construction
2. Pair-wise station comparisons

In the first method, the temperature records of several of the *target station*'s neighbours are averaged together into a single series, known as a *"reference series"*. The difference series is then made by subtracting this new reference series from the target series. The reference series is assumed to represent the climatic trends of the region, and so any differences between the target series and the reference series are probably non-climatic.

In the second method, multiple difference series are constructed, with a separate series for each of the neighbours. If there are any unusual differences between the target record and a neighbour, then it is assumed that a non-climatic bias occurred in either (a) the target or (b) the neighbour. The year of that difference is flagged. If the same year is flagged in comparisons with several other neighbours, then it is assumed that the target record is at fault, and an adjustment is applied.

Menne & Williams, 2009 found that the pair-wise method is more accurate than the reference series method if the number of neighbours used is at least seven[49].

Once the difference series has been constructed (single or multiple), it can then be analysed using statistical tests. In terms of the actual statistical test for a step bias, two general approaches have been popular,

1. Application of statistical tests to either side of a potential breakpoint
2. Testing of a particular *regression model* for all potential breakpoints, and assessing which model gives the closest fit to the difference series.

The first approach is to go through each potential breakpoint in the record and apply some statistical test to the portion of the record up to that year and the same test to the rest of the record. If the tests give very different results, then it is assumed that the year corresponds to a break-point, and an adjustment is applied to the period of the record up to that year (or alternatively, to the period of the record from that year on).

For example, in the *"Standard Normal Homogeneity Test"*[107], a statistic is determined for each datapoint by calculating the mean value of the part of the record up to that point and comparing it to the mean value of the rest of the record. When all datapoints have been tested, the statistics for each datapoint are ranked. If any of the test statistics have a value greater than a pre-defined threshold value, then a step-bias is assumed to have occurred at the datapoint corresponding to the highest value of the test statistic. The magnitude of the bias is then estimated as the difference between the two means.

In the second approach, a crude statistical model, with only a few parameters, is proposed. For instance, in the *"Two-Phase Regression"* model, the difference series is assumed to comprise a constant value for the period of the record up to a potential break-point, and then a different, constant value for the rest of the record. Each potential breakpoint is tested, and the success of the fit is calculated. Sometimes, the process is repeated with several different regression models, e.g., the Vincent, 1998 model[99]. The success of the *"null hypothesis"* fit, which assumes there is no bias in the record, is also calculated.

The relative success of each of the fits is then compared. If one of the models gives a much better fit than the null hypothesis fit, then whichever of the models does the best is assumed to be accurate. If this model suggests there has been one or more step-biases, then the relevant portions of the record are adjusted to account for those biases.

As we mentioned earlier, one of the main challenges in homogenizing temperature records is in making sure you correctly identify all of the biases (minimising false negatives) and avoid mistaking genuine trends as being biases (minimising false positives). One way to improve confidence that the breakpoints you have identified are actual biases would be if the breakpoints coincided with documented station changes which are likely to have caused a non-climatic bias, e.g., a station relocation, or a change in instrumentation.

The National Climatic Data Center have a station history file for the U.S. Network which includes the times of any *reported* station changes associated with each station. For this reason, for Version 1 of the U.S. Network, the step-bias adjustment algorithm (Karl & Williams, 1987[34]) used this file for deciding all the potential breakpoints. The Karl & Williams, 1987 algorithm *only* treated those years associated with a station change as potential breakpoints.

The Karl & Williams, 1987 algorithm offered the advantage that the only years which were tested were those which were associated with a known station change. It was hoped that this would reduce the number of false positives.

However, when the National Climatic Data Center were compiling the Global Network, they did not bother collecting any station histories. So, this algorithm would not work with the Global Network. For this reason, they used the Easterling & Peterson, 1995 algorithm[64] for homogenizing the Global Network[12]. In this algorithm, *all* years were treated as potential breakpoints.

One problem with the Karl & Williams, 1987 algorithm is that station observers do not always report changes which could introduce biases, e.g., they may not notice them, or realise their significance[108]. So, by only checking years that had a documented station change, the algorithm was potentially overlooking a large number of *undocumented* station changes. As a result, their algorithm was likely to be indirectly introducing false negatives.

In addition, not all station changes introduce a bias. But, since the Karl & Williams, 1987 algorithm was limited to only checking the documented years, any undocumented biases (of either the trend or the step type) which occurred outside of the tested years could skew the statistical tests. Some of the undocumented biases could be mistakenly attributed to the year of the documented change, even if the documented change had not itself caused any bias. This would increase the number of false positives.

With this in mind, for Version 2 of the U.S. Network, the National Climatic Data Center switched to using the Menne & Williams, 2009 algorithm[49]. Like the Easterling & Peterson, 1995 algorithm, it treated all years as potential breakpoints, i.e., it was designed for detecting undocumented step biases. But, it could still be used for testing the documented station changes. So, they run the algorithm twice - once only using the documented station changes as potential breakpoints ("test for documented biases"), and once treating all years as potential breakpoints ("test for undocumented biases")[15].

Since the breakpoints associated with actual documented station changes are probably more likely to be genuine than the ones estimated by purely statistical means[62, 109], a lower statistical threshold is required by the National Climatic Data Center to justify identifying a breakpoint as genuine, if it coincides with a documented station change[15].

For Version 3 of the Global Network, they also switched to using the Menne & Williams, 2009 algorithm. But, obviously, since they do not have a station history file for the Global Network, they only carry out the test for undocumented biases[13].

4.3.3 Assessment of step-bias homogenization approaches used by the National Climatic Data Center

In this section, we will assess whether or not the step-bias homogenization algorithms used by the National

Climatic Data Center on the Historical Climatology Networks have successfully removed the urbanization biases from the *Fully adjusted* datasets, as has been claimed by some, e.g., Refs. [15, 49, 57, 77].

The three algorithms used at various stages for the Historical Climatology Networks are quite distinct, and have some different properties. However, many of their characteristics, problems, and limitations, are similar. Hence, in this section, we will discuss all three simultaneously, although we will predominantly focus our attention on the Menne & Williams, 2009 algorithm, since this is the one currently used for homogenizing both the U.S. and the Global Network.

For brevity, we will refer to the three algorithms by the author initials and the year of publication:

1. **EP95**: Easterling & Peterson, 1995[64]. (See also Peterson & Easterling, 1994[89]). Used for Version 2 of Global Network[12].
2. **KW87**: Karl & Williams, 1987[34]. Used for Version 1 of U.S. Network[14].
3. **MW09**: Menne & Williams, 2009[49]. Used for Version 2 of U.S. Network[15] and Version 3 of Global Network[13].

Most of the technical differences between the three algorithms are important for a rigorous algorithm comparison, but do not significantly alter the main conclusions we will draw in this section, which are as follows:

1. Their methods for allocating station "neighbours" introduce undesirable selection biases.

2. Their algorithms are less effective when the station records are short and/or contain a large number of data gaps. This is a particularly serious problem for the Global Network.

3. Their algorithms treat trend biases as step biases.

4. Their algorithms lead to "urban blending", which means that urbanization bias will only (at best) be partially removed from urban stations, and will actually be introduced to many rural stations.

1. Problems with methods for allocating neighbours

The EP95 algorithm uses a reference series (constructed from 5 neighbours), while the other two algorithms use pair-wise neighbour comparisons (KW87 uses 20 neighbours, and MW09 uses 40 neighbours). However, in all cases, the neighbours are (a) partly chosen to minimise the distances of the neighbours from the target station, and (b) partly on the basis of maximising the correlation between the target record and the neighbour records.

Initially, these might seem like good ideas. If you want to identify non-climatic biases, it is important that the neighbouring stations are from the same climatic region. The further away a neighbour is from the target, the more likely it is to be in a different climatic region. So, it makes sense that we would prefer neighbours to be as close to the target station as possible. Hence, to get a higher density of potential neighbours, for Version 2 of the U.S. Network, Menne et al., 2009 decided to use the larger COOP Network dataset, rather than just using neighbours from the U.S. Network.

All of the station records in the U.S. Network were originally taken from the COOP dataset, and the COOP dataset has more than five times as many stations. So, this would substantially reduce the average distances of the nearest neighbours from the target station. However, it was an unwise decision for several reasons:

- The U.S. Network stations were carefully selected from the COOP dataset on the basis that they were of a relatively high quality[4]. So, the non-U.S. Network stations remaining in the COOP dataset are generally of a lower quality.

- Specifically, the average COOP record is of a much shorter length, and contains more data gaps than the U.S. Network records.

- The National Climatic Data Center do not currently provide easy access to the COOP dataset, making it harder for other researchers to assess the U.S. Network homogenization adjustments.

Also, station records from the same climatic region tend to be highly correlated. So, it might initially seem desirable to select the neighbours with the highest correlations to the target record. However, there are other, non-desirable reasons why two records would be highly-correlated.

Although, the National Climatic Data Center do not currently appear to provide public access to the COOP dataset, we had previously downloaded a 2011 version of the datasets from a temporary folder on the National Climatic Data Center's public ftp website (we downloaded them from ftp://ftp.ncdc.noaa.gov/pub/data/williams/). So, using this 2011 dataset, we wrote a script to calculate

Network	U.S.	U.S.	Global
Stations	1218	1218	6055
Record	93±14 yrs	93±14 yrs	44±33 yrs
Mean neighbour properties:			
Network:	COOP	U.S.	Global
Distance	107±24km	286±78km	683±598km
Overlap	29±6 yrs	78±12 yrs	23±15 yrs
r^2	0.85 ± 0.07	0.78 ± 0.09	0.79 ± 0.12
Tested	87%	100%	96%

Table 5: *Analysis of the nearest neighbours for the U.S. and Global Networks identified by Menne & Williams, 2009[49]'s sorting algorithm. Where appropriate, the shown values correspond to the mean ± the standard deviation (σ).*

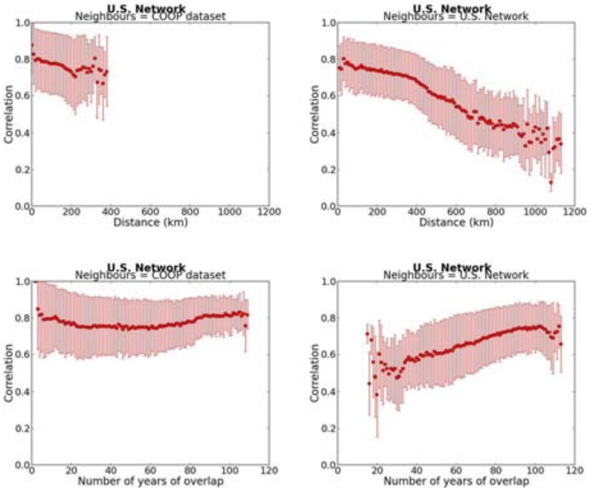

Figure 26: *Mean correlations of stations from the U.S. Network with their nearest neighbours with increasing distance (top) or number of years of overlap (bottom) as calculated using Menne & Williams, 2009's sorting algorithm. Confidence intervals correspond to two standard deviations. Left: U.S. Network stations with COOP neighbours - the current approach for the U.S. Network[15]. Right: U.S. Network stations with U.S. Network neighbours.*

the 40 COOP neighbours for each U.S. Network station, using our recreation of the MW09 algorithm. Additionally, we calculated the 40 neighbours they would have if the U.S. Network dataset was used for the neighbours instead. We also wrote a separate script to calculate the 40 neighbours for each of the Global Network stations[6]. The main results of these calculations are summarised in Table 5 and Figures 26 and 27.

In Figure 26, we plot the mean correlation for the U.S. Network between the neighbours and the target station, depending on (a) the distance between the neighbour and target (top panels), or (b) the length of overlap between the records (bottom panels). Figure 27 shows the equivalent plots for the Global Network.

We can see that generally the correlation decreases with increasing distance. This is as predicted - the further away two stations are, the less likely they are to be in the same climatic region. Indeed, this result has already been reported a few times, e.g., Refs. [17, 110, 111].

We also find that for stations with a fairly long period of overlap between their records (e.g., >50 or 60 years), the correlation tends to increase as the overlap increases. However, for stations with very short overlaps (e.g., <20 or 30 years), the opposite occurs, i.e., the correlation increases as the number of years of overlap decreases. This is a purely statistical artefact; when the overlap between the two records is too short there are not enough data-points to reveal how uncorrelated the climatic trends actually are. In other words, when the overlap period is short, a high correlation is *not necessarily* an indication that the records are showing the same climatic trends.

So, if the length of overlap for some of the records is quite small (e.g., less than 30 or 40 years), then selecting neighbours on the basis of maximising correlation can be counter-productive. Indeed, it may actually preferentially select for neighbours with shorter overlap periods.

From Table 5, it can be seen that the average overlap period for the Global Network stations is only 23 years. So, short overlap periods are a serious problem for the Global Network. The average overlap period for the U.S. Network would be relatively long if the neighbours were only selected from the U.S. Network (78 years). However, the National Climatic Data Center use the COOP dataset for their neighbours instead, and so the average overlap period is only 29 years.

There is another problem with selecting neighbours on the basis of having a highly correlated station record. If the target station contains non-climatic biases, then this should *reduce* the correlation between its record and any neighbours whose records are relatively unbiased. Selecting neighbours on the basis

[6] Both scripts are included in the Supplementary Information.

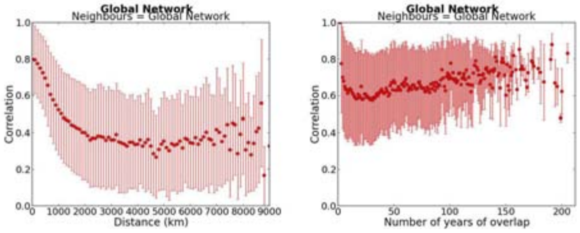

Figure 27: *Mean correlations of stations from the Global Network with their nearest neighbours with increasing distance (left) or number of years of overlap (right) as calculated using Menne & Williams, 2009's sorting algorithm. Confidence intervals correspond to two standard deviations.*

of high correlations increases the likelihood of selecting neighbours whose records are affected by similar non-climatic biases to the target station. A biased station record might be better correlated to similarly biased neighbours than to its unbiased neighbours.

2. Problems when records have data gaps or are too short

A problem that all three algorithms share with most step bias homogenization algorithms is that they require a minimum of several years uninterrupted data on either side of a potential breakpoint. This is because the magnitude of the bias is generally estimated by comparing the mean temperature of a period of several years before the breakpoint to the mean temperature of the same length period after the breakpoint. For the KW87 and EP95 algorithms, a period of at least 5 years on either side of a breakpoint, although the MW09 algorithm only requires 2 years data on either side[7].

As a result, these algorithms are known to be less reliable near the starts and ends of a record, as well as on either side of a gap in the record[59, 60, 112, 113]. Since the three algorithms are based on analysing difference series (Section 4.3.2), this problem applies to both the target record *and* the overlapping parts of the neighbour records.

From Table 5, the mean length of the records in the Global Network is only 44 years. So, this is a serious problem for the Global Network. As we discussed in Section 3.2, in the Global Network, the *fully rural* records tend to have less data. This means that (a) the *fully rural* records are harder to homogenize and (b) the problem is accentuated when using the *fully rural* records as neighbours.

The mean length of the records in the U.S. Network is considerably longer (93 years), and as we discussed in Section 3, the U.S. Network seems to be less heavily urbanized. However, as we mentioned above, the National Climatic Data Center do not currently use the U.S. Network stations for homogenizing their records. Instead, they use the COOP stations as neighbours. The mean number of overlapping years between these neighbours and the U.S. Network stations are only 29 years (Table 5). In other words, by using the much shorter COOP records for their neighbours, they reduce the reliability of the adjustments.

3. Problems with treating trend biases as step biases

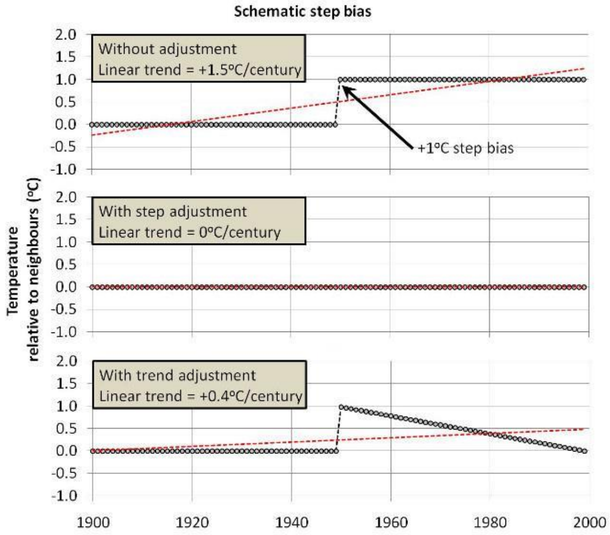

Figure 28: *Schematic illustration of the different effects of applying a step-change or a trend-change adjustment to a step bias.*

One of the biggest challenges in homogenizing climate records is in dealing with multiple biases. The KW87 (a Student "t" test) and EP95 ("two phase regression") algorithms are designed for dealing with multiple *step* biases. However, neither of those algorithms consider the presence of *trend* biases. As we discussed in Section 4.3.1, these two types of bias have quite different statistical properties.

The MW09 algorithm (a modified version of the

[7] The MW09 algorithm uses monthly data and so 2 years represents 24 data-points, as opposed to the 5 yearly data-points evaluated by the other two algorithms.

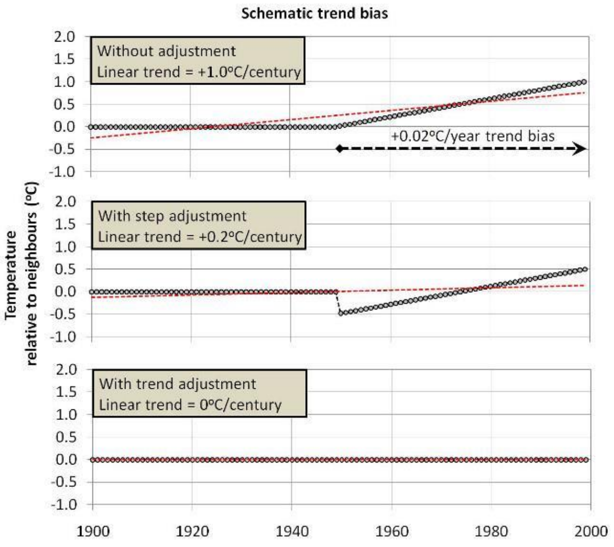

Figure 29: *Schematic illustration of the different effects of applying a step-change or a trend-change adjustment to a trend bias.*

Standard Normal Homogeneity Test) uses a more sophisticated approach in that it distinguishes between trend and step biases during the *identification* process. In doing so, Menne & Williams, 2009 were considering some of the advances in data homogenization proposed by Lund, Reeves, Wang and others[61, 62, 109, 114]. However, once the biases had been identified, the MW09 algorithm treated all biases as "step biases" during the actual adjustment process.

Treating a trend bias as a step bias (or vice versa) is a serious flaw in most of the current homogenization approaches. The problem is schematically illustrated by Figures 28 and 29. If a step bias is treated as a trend bias, then this will introduce an artificial trend into the climatic record, and keep a step bias (see Figure 28). Similarly, if a trend bias is treated as a step bias, then this will introduce an artificial step into the climatic record, and keep a trend bias (see Figure 29).

Menne & Williams, 2009[49] and Menne et al., 2009[15] have claimed that it is acceptable to treat a trend bias as a step bias, on the basis that it partially reduces the long-term trend of the series. Essentially they seem to be agreeing that any trend biases should be reduced, but suggest that a step-bias adjustment is sufficient to do so.

We strongly disagree with this claim. In our opinion, the middle panel of Figure 29 is *not* an acceptable homogenization. It is true that the net *linear trend* of the biases has been reduced, relative to the unadjusted top panel. However, the latter portion of the record *still has a trend bias of the same magnitude as before*. The only difference is that now the record has *two* biases instead of one.

Williams et al., 2012[115] carried out a series of tests of the MW09 algorithm using synthetic data with introduced errors. The algorithm did very well in identifying and removing most of the introduced errors. However, all of the introduced errors were step biases. So, the tests did not reveal how the algorithm fared in the presence of trend biases.

Venema et al., 2012[106] also tested the MW9 algorithm (and several others) using synthetic data. Again, the MW09 algorithm fared well in detecting and removing step biases. Unlike the Williams et al., 2012 studies, they introduced trend biases as well as step biases. However, they did not actually assess how successful the algorithms were at dealing with trend biases. Instead, they limited their analysis to evaluating the introduced step biases. Hence, Venema et al., 2012 did not reveal how the algorithm fared in the presence of trend biases, either.

In contrast, the results of those studies which have actually assessed how effective step adjustment methods are at treating trend biases have been negative[60, 88]. DeGaetano, 2006 found that when step bias adjustments were applied to a trend bias, only about half of the trend was included in the adjustment[60]. Pielke et al., 2007 found that, if a step bias occurred during a trend bias, then the magnitude of the step bias was overestimated when the two biases were of the same sign, but underestimated when the two biases were of opposite sign[88].

We recommend that future approaches to homogenizing temperature records should make more effort to try and correct for trend biases using trend adjustments and step biases using step adjustments. Ironically, this recommendation was actually made by Menne & Williams, 2009, although they chose to disregard it as being beyond the scope of their paper:

> *"Ultimately, a better solution would be to remove trend inhomogeneities via trend adjustments and step inhomogeneities via step adjustments."* - Menne & Williams, 2009[49]

4. The urban blending problem

The "urban blending" problem is a serious concern if you are using station comparison-based algorithms to homogenize temperature records and some of the stations are affected by urbanization bias. As we saw in Section 3, both of the Historical Climatology Networks are affected by urbanization bias, particularly the Global Network (Section 3.2). However, aside from a brief discussion in Hausfather et al., 2013[31], the National Climatic Data Center do not appear to have considered implications of the urban blending problem.

The problem arises from the assumption that the mean temperature trends of a station's neighbours, on average, *"...accurately reflected the climate of the region so that any significant departures from climatology could be directly associated with discontinuities in the station data"* (Peterson & Vose, 1997[12]). While this assumption might hold if (a) non-climatic biases were relatively rare, or (b) the biases were confined to occasional step changes that could be pinpointed to a few easily identified break-points in each record, it does *not* hold if a large number of stations are affected by *trend* biases like urbanization bias.

Consider the case of a target station in an urbanized area. If the area is currently urbanized, then it is likely that many of the stations in the area have been affected by urbanization bias. Some stations might be heavily affected, and other stations might be unaffected. This creates three scenarios for the station record:

1. The record is, on average, *more* affected by urbanization bias than its (urban) neighbours, e.g., a down-town station with a long record.
2. The record is affected by urbanization bias to about the same extent as its neighbours.
3. The record is, on average, *less* affected by urbanization bias than its neighbours, e.g., a rural station on the outskirts of the area.

In the first case, the homogenization algorithm would reduce the amount of urbanization bias in the record (assuming the algorithm overcomes the other problems discussed above). However, the amount of bias will *only* be reduced to match those of its neighbours.

In the second case, the homogenization algorithm would *not* reduce the amount of urbanization bias, because its neighbours are, on average, similarly affected.

In the third case, the homogenization algorithm would actually *introduce* a warming bias, so that the previously unbiased station record can better match the trends of its biased neighbours.

In all three cases, the homogenized trends of the station and its neighbours will be much closer to each other than they were before homogenization. In other words, the trends of all stations will be more "homogeneous", i.e., of a uniform nature. But, "homogeneous" does *not* mean less biased. The urbanization biases present in the unadjusted data will merely have been evenly distributed (or "blended") amongst the different stations.

The extent of this problem will depend on how heavily urbanized the neighbours are relative to the stations being homogenized. Let us first consider the U.S. Network.

Since the National Climatic Data Center use the COOP dataset as the source for their neighbours, we cannot use the Global Historical Climatology Network metadata to identify the degree of urbanization of the COOP stations. However, we can estimate their urbanization by assigning the station coordinates to the gridded urbanization estimates from the GRUMP dataset (see Ref. [80] for details on the GRUMP dataset).

Subset	**Neighbours:**		
	Urban	Rural	Water
Fully urban	24.4 ± 7.5	15.3 ± 7.6	0.3 ± 0.6
Intermediate	18.0 ± 7.6	21.8 ± 7.7	0.2 ± 0.5
Fully rural	15.0 ± 7.7	24.9 ± 7.7	0.1 ± 0.4
All stations	17.8 ± 8.0	22.0 ± 8.0	0.2 ± 0.5

Table 6: *Average number of urban and rural COOP neighbours used for homogenizing the U.S. Network stations with the Menne & Williams, 2009[49] algorithm. "Water neighbours" are in grid-boxes GRUMP identifies as being mostly water, and so are not identified as urban or rural. The ranges correspond to the standard deviation, which incorrectly assumes a Gaussian distribution, and hence should be treated cautiously.*

Using the lists of neighbours identified for each station that we calculated earlier, we calculated the average number of urban and rural COOP neighbours used for homogenizing the U.S. Network stations. Table 6 summarises the results. On average, about 44.5% of the 40 neighbours are identified as urban using the GRUMP metric. If we assume that a substantial fraction of the neighbours identified as urban using the GRUMP metric are at least partially affected by urbanization bias, then this leaves a strong possibility of urban blending.

If we look at the subset of the *fully urban* U.S. Network stations, then the average number of urban neighbours increases to 24.4 out of 40 (61%). This is not surprising, since urban stations are often clustered together, e.g., in metropolitan areas. However, it means that *fully urban* stations will tend to be homogenized by mostly urban neighbours, thereby leading to the first two scenarios of the urban blending problem.

The urban blending problem should be somewhat reduced for the *fully rural* subset, since an average of about 25 out of 40 neighbours are rural (62.5%). However, that still means that about 15 of the neighbours used for homogenizing *fully rural* stations are urban. In other words, the third scenario of the urban blending problem is a serious concern.

By just considering average statistics, this glosses over the fact that some areas may be heavily urbanized, but other areas may have almost no urbanization. This can be seen in Figure 7, where the *fully urban* stations tend to be clustered around metropolitan areas.

This adds a nuance to the urban blending problem. In very rural areas, very few of the neighbours (if any) would be urban. In these areas, urban blending should not be an issue. However, in very urban areas, very few of the neighbours (if any) would be rural. In these areas, urban blending could be a major problem.

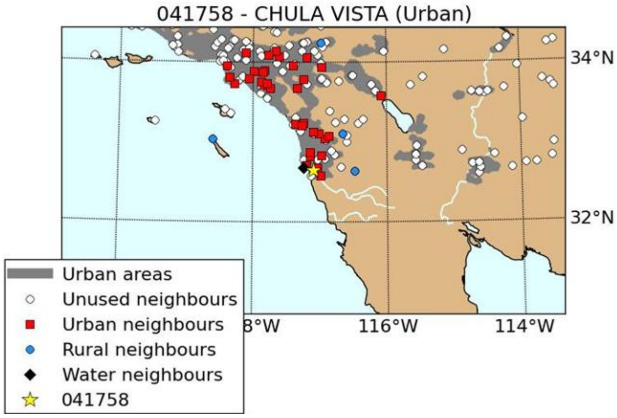

Figure 30: *Location of the* fully urban, *Chula Vista, California station (COOP ID=041758) and its neighbouring COOP stations. Urban areas and urban COOP neighbours are determined using the GRUMP dataset. "Water neighbours" are in grid-boxes GRUMP identifies as being mostly water, and so are not identified as urban or rural, although in this case the single station identified as such appears to be urban.*

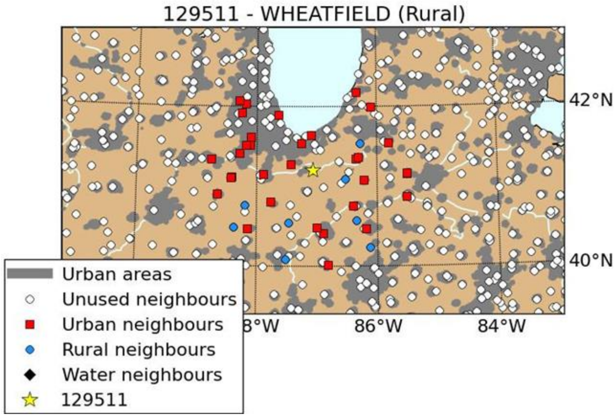

Figure 31: *Location of the* fully rural, *Wheatfield, Indiana station (COOP ID=129511) and its neighbouring COOP stations. Urban areas and urban COOP neighbours are determined using the GRUMP dataset.*

We can illustrate this by simply considering Figures 30 and 31. Both of these figures show stations in highly urbanized areas. Figure 30 shows the COOP neighbours used for homogenizing the *fully urban* Chula Vista station. Only 4 of the 40 neighbours (10%) are identified as rural by the GRUMP metric. Figure 31 shows the COOP neighbours used for homogenizing the *fully rural* Wheatfield station. Although the Wheatfield station is itself *fully rural*, 32 of its 40 COOP neighbours (80%) are urban. So, it is likely that it could be affected by urban blending.

As part of their study of urbanization bias in the U.S. Network, Hausfather et al., 2013 briefly considered the urban blending problem (see their Section 4.4)[31]. They found that the MW09 algorithm reduced the apparent divergences between urban and rural subsets, as discussed above. But, they recognised the possibility that this could be due to urban blending.

To test this, they divided up the COOP neighbour stations into urban and rural subsets, using the Impermeable Surface Area (ISA) associated with the station as an indicator of urbanization. This urbanization metric identified about 29% of U.S. stations as being "urban" and about 71% as "rural"[8]. They then applied the MW09 homogenization algorithm three times - once using all COOP stations (urban and rural) as neighbours; once using only the urban COOP stations as neighbours; and once using only the rural COOP stations as neighbours.

[8] We estimate this from their Table 1, in which they identified 857 of the U.S. Network stations as rural and 357 as urban, using the ISA metric.

Hausfather et al. found that using the urban-only stations led to more warming, as would be expected from urban blending. However, the rural-only adjustments were very similar to the standard adjustments using all stations. On this basis, they concluded that the COOP stations were *'...sufficiently "rural" to...'* avoid urban blending. This is an invalid conclusion. Their ISA urbanization metric identified more than 70% of the U.S. stations as "rural", whereas we saw from Table 2 that only 22.74% of the stations in the U.S. Network are *fully rural* by our metrics. This means that many of their "rural" neighbours may well have been affected by urbanization bias.

We do not think it is particularly surprising that their rural-only homogenization gave similar results to their all-COOP homogenization, since their "rural" subset comprised the majority of stations in the COOP. Indeed, we note that their all-COOP homogenization led to about 30% more adjustments than their rural-only subset. They suggested that this was because the rural-only subset was less dense (and therefore less reliable). However, an alternative explanation is that it caused slightly less urban blending.

Hausfather et al., 2013 claimed that urban blending was not a major problem for the U.S. Network, because they believed the COOP dataset was *'sufficiently "rural" '*. However, as we discussed in Section 3, urbanization bias seems to be an even greater problem for the Global Network. So, let us now assess the reliability of the MW09 approach for the Global Network.

Table 7 lists the average urbanization of the neighbours used for homogenizing the stations in the Global Network, as before - except that we did not need to use the GRUMP dataset, since we could use the Global Historical Climatology Network urbanization metrics. We can see that urban blending is a serious concern for the Global Network. On average, only 11.9 out of the 40 neighbours (29.75%) used for homogenizing the stations are *fully rural*. As for the U.S. Network, this number increases when we consider the *fully rural* subset, and decreases for the *fully urban* subset. But, in all cases, urban blending is a potential problem. We can assess the problem in more detail, by considering a couple of case studies.

First, let us consider the case of the *fully rural* station, Valentia Observatory, Ireland. This is one of the eight *fully rural* Global Network stations with data for at least 95% of the last century that we discussed in Section 3.2.

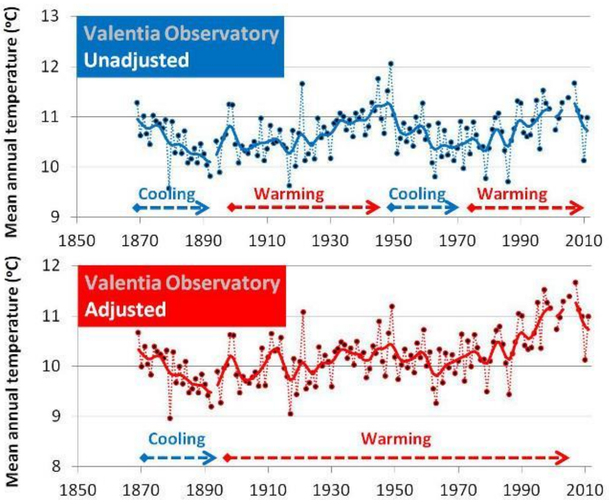

Figure 32: *Temperature trends for the Valentia Observatory station before* (top) *and after* (bottom) *homogenization. Solid lines correspond to 11 point binomial smoothed version of the annual data. The labelled "warming"/"cooling" periods are only qualitatively estimated, and are merely provided to illustrate the approximate differences between the two versions.*

Figure 32 shows its annual temperature record before (top panel) and after (bottom panel) homogenization. The *Unadjusted* record suggests an alternation between periods of warming and periods of cooling, each lasting a few decades. Recent temperatures do not appear particularly unusual. However, in the *Adjusted* record, the trends have changed to an almost continuous warming trend.

In effect, the MW09 homogenization has reduced the cooling trends and increased the warming trends. Perhaps these trends were non-climatic and the

Subset	**Neighbours:**		
	Fully urban	Intermediate	Fully rural
Fully urban	15.6 ± 6.3	17.2 ± 4.1	7.2 ± 5.0
Intermediate	10.8 ± 6.6	18.1 ± 4.2	11.2 ± 6.8
Fully rural	6.9 ± 6.1	16.5 ± 4.8	16.5 ± 7.9
All stations	10.7 ± 7.2	17.3 ± 4.4	11.9 ± 7.7

Table 7: *Average number of* fully urban, intermediate *and* fully rural *neighbours used for homogenizing the Global Network stations with the Menne & Williams, 2009[49] algorithm. The ranges correspond to the standard deviation, which incorrectly assumes a Gaussian distribution, and hence should be treated cautiously.*

MW09 algorithm was simply correcting for non-climatic biases. However, urban blending could also cause this, since urbanization bias introduces a warming trend to station records.

We could rule out the possibility of urban blending if the neighbouring stations used for calculating these adjustments were mostly rural. However, Figure 33 shows the location of the 40 neighbours used for homogenizing Valentia Observatory - only 8 of them are *fully rural* (20%). So, urban blending is a strong possibility.

If we consider Figure 34, the possibility of urban blending becomes greater. We can see that most of the *fully rural* neighbours have very short records, and that stations with the longest and most complete records are the *fully urban* stations. This suggests that the adjustments were *mostly* calculated using *fully urban* station records, which are likely to be affected by urbanization bias.

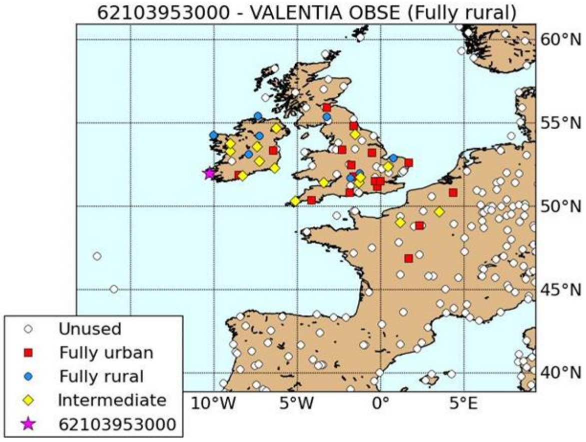

Figure 33: *Valentia Observatory, Ireland and the neighbours used for homogenizing its record, according to the Menne & Williams, 2009 algorithm. The magenta star corresponds to the Valentia Observatory station.*

This shows how the MW09 algorithm can easily introduce urban blending into the records of *fully rural* stations (scenario 3 of the urban blending problem). However, what about the other two scenarios? Could urban blending prevent the MW09 algorithm from removing urbanization bias from urban stations? To test this, let us consider our second case study - the *fully urban* station at Buenos Aires, Argentina.

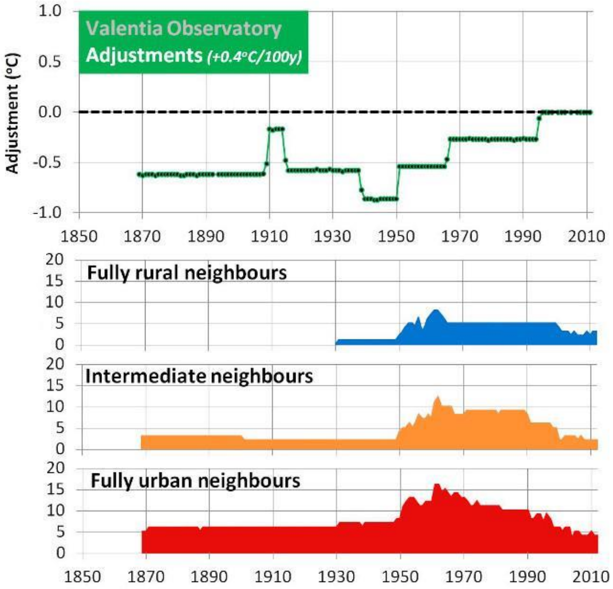

Figure 34: *Adjustments applied to the Valentia Observatory record in January 2013, and the neighbours which were used for calculating these adjustments.*

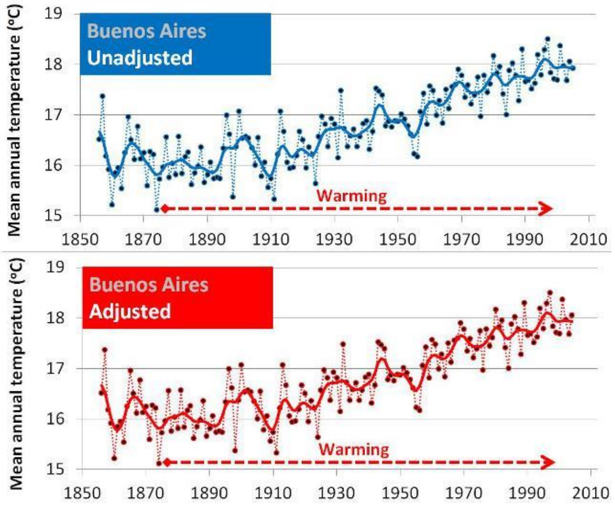

Figure 35: *Temperature trends of the Buenos Aires station before (top) and after (bottom) homogenization. Solid lines correspond to 11 point binomial smoothed version of the annual data.*

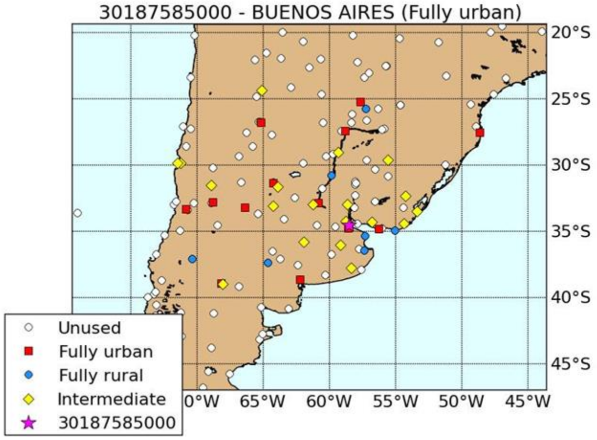

Figure 36: *Buenos Aires, Argentina and the neighbours used for homogenizing its record, according to the Menne & Williams, 2009 algorithm. The magenta star corresponds to the Buenos Aires station.*

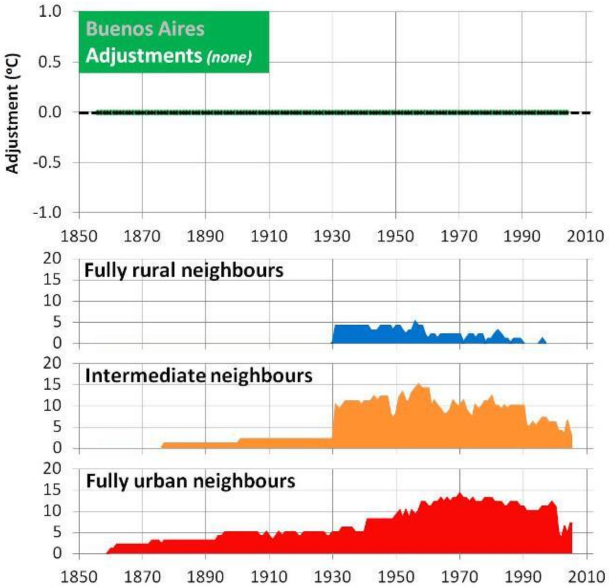

Figure 37: *Adjustments applied to the Buenos Aires record in January 2013, and the neighbours which were used for calculating these adjustments.*

Buenos Aires has witnessed dramatic urbanization over the last century, and the city is known to have a strong urban heat island, e.g., see Figuerola & Mazzeo, 1998[116]. Therefore, if the National Climatic Data Center are correct in their claim that the MW09 algorithm removes most of the urbanization bias from station records, then the homogenization adjustments for the Buenos Aires station should be quite substantial.

Figure 35 shows its annual temperature record before (top panel) and after (bottom panel) homogenization using the MW09 algorithm. Before homogenization, the record suggests a strong, almost continuous, warming trend, which is typical of a record strongly affected by urbanization bias. However, after homogenization, the record remains exactly the same.

Figures 36 and 37 explain why. Only 7 of the 40 neighbours (17.5%) are *fully rural*, and as for Valentia Observatory, their records are relatively short. Again, as for Valentia Observatory, the longest and most complete station records are for *fully urban* stations. Hence, the urban blending problem prevents the MW09 algorithm from removing *any* urbanization bias from the Buenos Aires record.

5 Conclusions

In this paper, we considered the extent to which urbanization bias is likely to be affecting the two Historical Climatology Network datasets. We found that both datasets are indeed significantly affected by urbanization.

The U.S. Network is relatively rural. Presumably this is aided by the fact that Karl et al., 1988 had actively attempted to select mostly rural stations when compiling the dataset[4]. However, only about 23% of the stations are *"fully rural"* - by which we mean rural in terms of both low neighbouring population and low associated night-light brightness. So, urbanization bias is potentially a problem for more than three quarters of the stations.

About 10% of the U.S. Network stations are *fully urban*. These are definitely affected by urbanization bias. We found that for the *Unadjusted* dataset, the subset of these *fully urban* stations show a warming trend of about $0.7°C$/century relative to the *fully rural* subset. In the *Time-of-Observation* adjusted version of the U.S. Network (the *"Partially adjusted"* dataset), this difference was partially reduced, which suggests that some of the apparent urban-rural difference is due to different observation practices between the subsets. However, the urban-rural difference is still substantial (about $0.5°C$/century) for the *Time-of-observation* adjusted version.

Compared to the U.S. Network, the Global Network appears to be a highly urbanized dataset. Although nearly a third (1992 out of 6062) of the stations in the dataset are *fully rural*, most of these stations have records with only a few decades of data. Of the 173 stations with data for at least 95 of the last 100 years, only 8 (4.6%) are *fully rural*.

Worse still, only one of the eight stations is from the southern hemisphere, and five of the stations are all from a similar region (Europe). This is too sparse a distribution to claim a global coverage.

There have been claims in the literature[13, 15] that the Menne & Williams, 2009 homogenization algorithm[49] which is applied to the homogenized versions of the two datasets substantially reduces the urbanization bias problem. We found these claims to be unfounded, and we believe that the homogenization has probably *reduced* the reliability of the dataset, rather than improved it.

The step-change adjustments proposed by Menne & Williams, 2009 are inappropriate for removing trend biases such as urbanization bias. But, the algorithm seems to be particularly unsuccessful when a large number of stations are affected by urbanization bias, as seems to be the case for the Global Network.

Even in the relatively rural U.S. Network, the step-change homogenization only reduced the difference between the *fully rural* and *fully urban* subsets from about 0.5°C/century to about 0.3°C/century. So, the homogenization did *not* succeed in removing all of the urbanization bias in the un-homogenized dataset. Hausfather et al., 2013 also found a similar result in their analysis of urbanization bias in the U.S. Network, although they were more optimistic about the reliability of the homogenized dataset in their conclusions[31].

There is a more serious problem with the homogenization algorithm, however. Much of the apparent "reduction" in the urban-rural difference seems to be a result of "urban blending", i.e., the spreading of urbanization bias into rural neighbours through homogenization. We illustrated the problem of urban blending in the Global Network by considering the homogenization adjustments applied to a *fully rural* station (Valentia Observatory, Ireland) and a *fully urban* station (Buenos Aires, Argentina).

The Menne & Williams, 2009 algorithm introduced a substantial "warming" adjustment to the *fully rural* Valentia Observatory station, but this adjustment did not appear to be based on the trends of its rural neighbours. Rather, it appeared to be mostly based on the trends of its urban neighbours. In other words, the record of the rural record was adjusted to better match the records of its urban neighbours.

On the other hand, with the *fully urban* Buenos Aires station, no adjustment was applied. This was a station located in an area which has seen a very pronounced urbanization over the last century, and is known to have a strong urban heat island[116]. However, the Menne & Williams, 2009 algorithm failed to remove any urbanization bias. This appears to be because there were only a few *fully rural* stations used as neighbours, and the average overlap of their records with the Buenos Aires record was only 21.5 years.

We note that rural trends for the U.S. Network suggest an alternation between warming periods and cooling periods since the start of the dataset in 1895. Although the U.S. Network suggests a warming period since the 1970s, it followed a cooling period from the 1940s-1970s. As a result, the recent warm period in the U.S. seems comparable to the 1920s-1940s warm period. That is, recent U.S. temperatures do *not* appear to be unusual or unprecedented, despite several claims to the contrary[65–67].

Because there is such a severe shortage of *fully rural* stations with long records in the Global Network, we were unable to determine exactly what the true global temperature trends since the 19th century have been. But, it seems very likely that the oft-cited claims of unusual "global warming"[5] have been substantially exaggerated by urbanization bias, at least.

Acknowledgements

No funding was received for this research.

We would like to thank Don Zieman for some useful feedback on an early draft of this article. We would also like to thank Dr. Bianca Schön-Phelan for assistance in translating the Wikipedia.de article on the Hohenpeißenberg station, and Paul Whelan for some help during a preliminary (2010) analysis of the Global Historical Climatology Network dataset.

References

[1] R. Connolly and M. Connolly. "Urbanization bias I. Is it a negligible problem for global temperature estimates?" 28 (Clim. Sci.). Ver. 0.1 (non peer reviewed draft). 2014. URL: http://oprj.net/articles/climate-science/28.

[2] R. Connolly and M. Connolly. "Urbanization bias II. An assessment of the NASA GISS urbanization adjustment method". 31 (Clim. Sci.). Ver. 0.1 (non peer reviewed draft). 2014. URL: http://oprj.net/articles/climate-science/31.

[3] I. D. Stewart and T. R. Oke. "Local climate zones for urban temperature studies". *Bull. Amer. Meteor. Soc.* 93 (2012), pp. 1879–1900. DOI: 10.1175/BAMS-D-11-00019.1.

[4] T. R. Karl, H. F. Diaz, and G. Kukla. "Urbanization: Its detection and effect in the United States climate record". *J. Clim.* 1 (1988), pp. 1099–1123. DOI: 10.1175/1520-0442(1988)001<1099:UIDAEI>2.0.CO;2.

[5] K. E. Trenberth et al. "3. Observations: Surface and atmospheric climate change". In: *Climate change 2007: The physical science basis. Contribution of Working Group 1 to the Third Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by S. Solomon et al. Cambridge University Press. Cambridge. New York., 2007, 1056pp. URL: http://www.ipcc.ch/.

[6] G. C. Hegerl et al. "7. Understanding and attributing climate change". In: *Climate change 2007: The physical science basis. Contribution of Working Group 1 to the Third Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by S. Solomon et al. Cambridge University Press. Cambridge. New York., 2007, 1056pp. URL: http://www.ipcc.ch/.

[7] N. Stern. "The economics of climate change: The Stern review". In: Cambridge University Press. Cambridge. New York., 2007, 712pp. ISBN: 9780521700801.

[8] E. Jansen et al. "6. Palaeoclimate". In: *Climate change 2007: The physical science basis. Contribution of Working Group 1 to the Third Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by S. Solomon et al. Cambridge University Press. Cambridge. New York., 2007, 1056pp. URL: http://www.ipcc.ch/.

[9] R. Connolly and M. Connolly. "Global temperature changes of the last millennium". 16 (Clim. Sci.). Ver. 0.1 (non peer reviewed draft). 2014. URL: http://oprj.net/articles/climate-science/16.

[10] J. M. Mitchell. "On the causes of instrumentally observed secular temperature trends". *J. Meteor.* 10 (1953), pp. 244–261. DOI: 10.1175/1520-0469(1953)010<0244:OTCOIO>2.0.CO;2.

[11] R. S. Vose et al. "The Global Historical Climatology Network: Long-term monthly temperature, precipitation, sea level pressure, and station pressure data". Carbon Dioxide Information Analysis Center. Oak Ridge National Laboratory, Oak Ridge, TN 37831. 1992. URL: http://cdiac.ornl.gov/epubs/ndp/ndp041/ndp041.html.

[12] T. C. Peterson and R. S. Vose. "An overview of the Global Historical Climatology Network temperature database". *Bull. Amer. Meteor. Soc.* 78 (1997), pp. 2837–2849. DOI: 10.1175/1520-0477(1997)078<2837:AOOTGH>2.0.CO;2.

[13] J. H. Lawrimore et al. "An overview of the Global Historical Climatology Network monthly mean temperature dataset, Version 3". *J. Geophys. Res.* 116 (2011), p. D19121. DOI: 10.1029/2011JD016187.

[14] D. R. Easterling et al. "United States Historical Climatology Network (U.S. HCN) monthly temperature and precipitation data". ORNL/CDIAC-87, NDP-019/R3. Carbon Dioxide Information Analysis Center. Oak Ridge National Laboratory, Oak Ridge, TN 37831. 1996. URL: http://cdiac.ornl.gov/r3d/ushcn/ushcn.html.

[15] M. J. Menne, C. N. Jr. Williams, and R. S. Vose. "The U.S. Historical Climatology Network monthly temperature data, version 2". *Bull. Amer. Meteor. Soc.* 90 (2009), pp. 993–1007. DOI: 10.1175/2008BAMS2613.1.

[16] M. J. Menne, C. N. Jr. Williams, and R. S. Vose. "United States Historical Climatology Network (USHCN) Version 2 serial monthly dataset". NOAA National Climatic Data Center. 2011. URL: http://www.ncdc.noaa.gov/oa/climate/research/ushcn/.

[17] R. Rohde et al. "Berkeley Earth temperature averaging process". *Geoinfor. Geostat.* 1 (2013). DOI: 10.4172/gigs.1000103.

[18] P. Brohan et al. "Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850". *J. Geophys. Res.* 111 (2006), p. D12106. DOI: 10.1029/2005JD006548.

[19] R. Rohde et al. "A new estimate of the average Earth surface land temperature spanning 1753 to 2011". *Geoinfor. Geostat.* 1 (2013). DOI: 10.4172/gigs.1000101.

[20] F. Fujibe and K. Ishihara. "Possible urban bias in gridded climate temperature data over the Japan area". *SOLA* 6 (2010), pp. 61–64. DOI: 10.2151/sola.2010-016.

[21] P. D. Jones et al. "Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010". *J. Geophys. Res.* 117 (2012), p. D05127. DOI: 10.1029/2011JD017139.

[22] K. M. Lugina et al. "Monthly surface air temperature time series area-averaged over the 30-degree latitudinal belts of the globe, 1881-2005". In: *Trends: A compendium of data on global change*. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, TN, U.S.A. DOI: 10.3334/CDIAC/cli.003.

[23] T. M. Smith et al. "Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880-2006)". *J. Clim.* 21 (2008), pp. 2283–2296. DOI: 10.1175/2007JCLI2100.1.

[24] T. C. Peterson et al. "First difference method: Maximising station density for the calculation of long-term global temperature change". *J. Geophys. Res. D* 103 (1998), pp. 25967–25974. DOI: 10.1029/98JD01168.

[25] J. Hansen et al. "Global surface temperature change". *Rev. Geophys.* 48 (2010), RG4004. DOI: 10.1029/2010RG000345.

[26] R. McKitrick. "A critical review of global surface temperature data products". Soc. Sci. Res. Network. August 5, 2010 version. URL: http://ssrn.com/abstract=1653928.

[27] P. D. Jones and T. M. L. Wigley. "Estimation of global temperature trends: what's important and what isn't". *Clim. Change* 100 (2010), pp. 59–69. DOI: 10.1007/s10584-010-9836-3.

[28] R. A. Muller. "The case against global-warming skepticism". Wall Street Journal. October 21. 2011. URL: http://online.wsj.com/article/SB10001424052970204422404576594872796327348.html.

[29] R. A. Muller et al. "Earth atmospheric land surface temperature and station quality in the United States". *Submitted to journal* (2011). URL: http://berkeleyearth.org/.

[30] U.S. Census Bureau. "United States Summary: 2010; Population and housing unit counts. CPH-2-1. Issued September 2012". *Census of population and housing, 2010* (2012). URL: http://www.census.gov/prod/www/decennial.html.

[31] Z. Hausfather et al. "Quantifying the effect of urbanization on U.S. Historical Climatology Network temperature records". *J. Geophys. Res.* in press (2013). DOI: 10.1029/2012JD018509.

[32] E. Kalnay and M. Cai. "Impact of urbanization and land-use change on climate". *Nature* 423 (2003), pp. 528–531. DOI: 10.1038/nature01675.

[33] United Nations. "World urbanization prospects: The 2009 revision". *Population Division* (2010). URL: http://esa.un.org/unpd/wup/index.htm.

[34] T. R. Karl and C. N. Jr. Williams. "An approach to adjusting climatological time series for discontinuous inhomogeneities". *J. Clim. Appl. Meteor.* 26 (1987), pp. 1744–1763. DOI: 10.1175/1520-0450(1987)026<1744:AATACT>2.0.CO;2.

[35] T. R. Karl et al. "A model to estimate time of observation bias associated with monthly mean maximum, minimum and mean temperatures for the United States". *J. Clim. Appl. Meteor.* 25 (1986), pp. 145–160. DOI: 10.1175/1520-0450(1986)025<0145:AMTETT>2.0.CO;2.

[36] C. Gillham et al. "Near enough for a sheep station". Ken's Kingdom blog. 2012. URL: http://kenskingdom.wordpress.com/2012/03/13/near-enough-for-a-sheep-station/.

[37] C. A. Davey and R. A. Sr. Pielke. "Microclimate exposures of surface-based weather stations: Implication for the assessment of long-term temperature trends". *Bull. Amer. Meteor. Soc.* 86 (2005), pp. 497–504. DOI: 10.1175/BAMS-86-4-497.

[38] A. Watts. "Is the U.S. surface temperature record reliable?" The Heartland Institute. Chicago, IL. 2009. URL: http://surfacestations.org/.

[39] R. G. Quayle et al. "Effects of recent thermometer changes in the Cooperative Station Network". *Bull. Amer. Meteor. Soc.* 72 (1991), pp. 1718–1723. DOI: 10.1175/1520-0477(1991)072<1718:EORTCI>2.0.CO;2.

[40] K. G. Hubbard and X. Lin. "Reexamination of instrument change effects in the U.S. Historical Climatology Network". *Geophys. Res. Lett.* 33 (2006), p. L15710. DOI: 10.1029/2006GL027069.

[41] R. C. Hale et al. "Land use/land cover change effects on temperature trends at U.S. Climate Normals stations". *Geophys. Res. Lett.* 33 (2006), p. L11703. DOI: 10.1029/2006GL026358.

[42] Wettersäulen in Europa. "Säntis Wetterstation (in German)". Photographic history of the Säntis weather station. Accessed: 2013-02-08. (Archived by WebCite® at http://www.webcitation.org/6EHP6vrIE). URL: http://www.wettersaeulen-in-europa.de/direct.htm?/saentis/saentis.htm.

[43] M. Begert, T. Schlegel, and W. Kirchhofer. "Homogeneous temperature and precipitation series of Switzerland from 1864 to 2000". *Int. J. Climatol.* 25 (2005), pp. 65–80. DOI: 10.1002/joc.1118.

[44] P. D. Jones et al. "A grid point surface air temperature data set for the Northern Hemisphere". In: *Technical Report #022 (TR022)*. U. S. Department of Energy, Office of Energy Research, Office of Basic Energy Sciences, Carbon Dioxide Research Division. Washington, D.C. 20545. URL: http://www.cru.uea.ac.uk/st/.

[45] Australian Government Bureau of Meteorology. "History of Lord Howe Island Meteorological Office". Accessed: 2013-07-23. (Archived by WebCite® at http://www.webcitation.org/6IKPJw8IM). URL: http://www.bom.gov.au/nsw/lord_howe/history.shtml.

[46] Wikipedia contributors. "Meteorologisches Observatorium Hohenpeißenberg (in German)". Accessed: 2013-07-11. (Archived by WebCite at http://www.webcitation.org/6I26nJr4z). URL: http://de.wikipedia.org/wiki/Meteorologisches_Observatorium_Hohenpei%C3%9Fenberg.

[47] R. McGreevy. *Irish Times* (April 11th, 2011). URL: http://www.irishtimes.com/newspaper/ireland/2011/0411/1224294394389.html.

[48] R. Connolly and M. Connolly. "Has poor station quality biased U.S. temperature trend estimates?" 11 (Clim. Sci.). Ver. 0.1 (non peer reviewed draft). 2014. URL: http://oprj.net/articles/climate-science/11.

[49] M. J. Menne and C. N. Jr. Williams. "Homogenization of temperature series via pairwise comparisons". *J. Clim.* 22 (2009), pp. 1700–1717. DOI: 10.1175/2008JCLI2263.1.

[50] ACIA. "Arctic Climate Impact Assessment". In: Cambridge University Press, UK., 2005, 1042pp. URL: http://www.acia.uaf.edu.

[51] I. V. Polyakov et al. "Variability and trends of air temperature and pressure in the maritime Arctic, 1875-2000". *J. Clim.* 16 (2003), pp. 2067–2077. DOI: 10.1175/1520-0442(2003)016<2067:VATOAT>2.0.CO;2.

[52] S. I. Kuzmina et al. "High northern latitude surface air temperature: comparison of existing data and creation of a new gridded data set 1900-2000". *Tellus* 60A (2008), pp. 289–304. DOI: 10.1111/j.1600-0870.2008.00303.x.

[53] J. C. Comiso et al. "Accelerated decline in the Arctic sea ice cover". *Geophys. Res. Lett.* 35 (2008), p. L01703. DOI: 10.1029/2007GL031972.

[54] G. S. Callendar. "The artificial production of carbon dioxide and its influence on temperature". *Quart. J. Roy. Meteor. Soc.* 64 (1938), pp. 223–240. DOI: 10.1002/qj.49706427503.

[55] G. J. Kukla et al. "New data on climatic trends". *Nature* 270 (1977), pp. 573–580. DOI: 10.1038/270573a0.

[56] K. M. Hinkel and F. E. Nelson. "Anthropogenic heat island at Barrow, Alaska, during winter: 2001-2005". *J. Geophys. Res.* 112 (2007), p. D06118. DOI: 10.1029/2006JD007837.

[57] T. C. Peterson et al. "Global rural temperature trends". *Geophys. Res. Lett.* 26 (1999), pp. 329–332. DOI: 10.1029/1998GL900322.

[58] E. Aguilar et al. "Guidelines on climate metadata and homogenization". WCDMP-53. WMO-TD No. 1186. World Meteorological Organization, Geneva, Switzerland. 2003. URL: http://www.wmo.int/pages/prog/wcp/wcdmp/wcdmp_series/.

[59] J.-F. Ducré-Robitaille, L. A. Vincent, and G. Boulet. "Comparison of techniques for detection of discontinuities in temperature series". *Int. J. Clim.* 23 (2003), pp. 1087–1101. DOI: 10.1002/joc.924.

[60] A. T. DeGaetano. "Attributes of several methods for detecting discontinuities in mean temperature series". *J. Clim.* 19 (2006), pp. 838–853. DOI: 10.1175/JCLI3662.1.

[61] J. Reeves et al. "A review and comparison of changepoint detection techniques for climate data". *J. Appl. Meteor. Clim.* 46 (2007), pp. 900–915. DOI: 10.1175/JAM2493.1.

[62] R. Lund and J. Reeves. "Detection of undocumented changepoints: A revision of the two-phase regression model". *J. Clim.* 15 (2002), pp. 2547–2554. DOI: 10.1175/1520-0442(2002)015<2547:DOUCAR>2.0.CO;2.

[63] K. E. Runnalls and T. R. Oke. "A technique to detect microclimatic inhomogeneities in historical records of screen-level air temperature". *J. Clim.* 19 (2006), pp. 959–978. DOI: 10.1175/JCLI3663.1.

[64] D. R. Easterling and T. C. Peterson. "A new method for detecting undocumented discontinuities in climatological time series". *Int. J. Clim.* 15 (1995), pp. 369–377. DOI: 10.1002/joc.3370150403.

[65] J. Gillis. "Not even close: 2012 was hottest ever in U.S." New York Times on 8 January 2013. Accessed: 2013-03-12. (Archived by WebCite® at http://www.webcitation.org/6F4adRVHa). URL: http://www.nytimes.com/2013/01/09/science/earth/2012-was-hottest-year-ever-in-us.html?_r=0.

[66] B. Walsh. "2012 was the hottest year in U.S. history. And yes - it's climate change". Time magazine, Ecocentric blog post on 8 January 2013. Accessed: 2013-03-12. (Archived by WebCite® at http://www.webcitation.org/6F4ayyon3). URL: http://science.time.com/2013/01/08/2012-was-the-hottest-year-in-u-s-history-and-yes-its-climate-change/.

[67] K. Than. "2012: hottest year on record for continental U.S." National Geographic News. 9 January 2013. Accessed: 2013-03-12. (Archived by WebCite® at http://www.webcitation.org/6F4blW8i2). URL: http://news.nationalgeographic.com/news/2013/01/130109-warmest-year-record-2012-global-warming-science-environment-united-states/.

[68] W. Ellis. "On the difference produced in the mean temperature derived from daily maxima and minima as dependent on the time at which the thermometers are read". *Quart. J. Roy. Met. Soc.* 16 (1890), pp. 213–218. DOI: 10.1002/qj.4970167605.

[69] E. S. Nichols. "Time limits of the day as affecting records of minimum temperature". *Mon. Wea. Rev.* 62 (1934), pp. 337–343. DOI: 10.1175/1520-0493(1934)62<337:TLOTDA>2.0.CO;2.

[70] W. F. Rumbaugh. "The effect of time of observation on mean temperature". *Mon. Wea. Rev.* 62 (1934), pp. 375–376. DOI: 1520-0493(1934)62<375:TEOTOO>2.0.CO;2.

[71] B. N. Belcher and A. T. deGaetano. "A method to infer time of observation at US Cooperative Observer Network stations using model analyses". *Int. J. Clim.* 25 (2005), pp. 1237–1251. DOI: 10.1002/joc.1183.

[72] L. A. Vincent et al. "Bias in minimum temperature introduced by a redefinition of the climatological day at the Canadian synoptic stations". *J. Appl. Meteor. Clim.* 48 (2009), pp. 2160–2168. DOI: 10.1175/2009JAMC2191.1.

[73] R. C. Jr. Balling and C. D. Idso. "Analysis of adjustments to the United States Historical Climatology Network (USHCN) temperature database". *Geophys. Res. Lett.* 29 (2002), p. 1387. DOI: 10.1029/2002GL014825.

[74] R. S. Vose et al. "An evaluation of the time of observation bias adjustment in the U.S. Historical Climatology Network". *Geophys. Res. Lett.* 30 (2003), p. 2046. DOI: 10.1029/2003GL018111.

[75] A. T. DeGaetano. "A method to infer observation time based on day-to-day temperature variations". *J. Clim.* 12 (1999), pp. 3443–3456. DOI: 10.1175/1520-0442(1999)012<3443:AMTIOT>2.0.CO;2.

[76] A. Watts et al. "An area and distance weighted analysis of the impacts of station exposure on the U.S. Historical Climatology Network temperatures and temperature trends". *In preparation* (2012). URL: http://wattsupwiththat.com/2012/07/29/press-release-2/.

[77] T. C. Peterson. “Assessment of urban versus rural in situ surface temperatures in the contiguous United States: No difference found”. *J. Clim.* 16 (2003), pp. 2941–2959. DOI: 10.1175/1520-0442(2003)016<2941:AOUVRI>2.0.CO;2.

[78] G. Kukla, J. Gavin, and T. R. Karl. “Urban warming”. *J. Clim. Appl. Meteor.* 25 (1986), pp. 1265–1270. DOI: 10.1175/1520-0450(1986)025<1265:UW>2.0.CO;2.

[79] D. R. Cayan and A. V. Douglas. “Urban influences on surface temperatures in the southwestern United States during recent decades”. *J. Clim. Appl. Meteor.* 23 (1984), pp. 1520–1530. DOI: 10.1175/1520-0450(1984)023<1520:UIOSTI>2.0.CO;2.

[80] Columbia University Center for International Earth Science Information Network (CIESIN) et al. “Global Rural-Urban Mapping Project, Version 1 (GRUMPv1): Urban Extents Grid”. Date of download: 6th January 2013. 2011. URL: http://sedac.ciesin.columbia.edu/data/dataset/grump-v1-urban-extents.

[81] J. Hansen et al. “A closer look at United States and global surface temperature change”. *J. Geophys. Res. D* 106 (2001), pp. 23947–23963. DOI: 10.1029/2001JD000354.

[82] C. N. Jr. Williams. “A comparison of the original United States Historical Climatology Network (USHCN) and USHCN v2”. Presentation to “18th Conference on Climate Variaibility and Change”. The 86th AMS Annual Meeting (Atlanta, GA). 31 January. 2006. URL: http://ams.confex.com/ams/Annual2006/techprogram/paper_100746.htm.

[83] R. C. Jr. Balling and R. S. Cerveny. “Long-term associations between winds speeds and the urban heat island of Phoenix, Arizona”. *J. Clim. Appl. Meteor.* 26 (1987), pp. 712–716. DOI: 10.1175/1520-0450(1987)026<0712:LTABWS>2.0.CO;2.

[84] R. A. Sr. Pielke et al. “Problems in evaluating regional and local trends in temperature: An example from eastern Colorado, USA”. *Int. J. Clim.* 22 (2002), pp. 421–434. DOI: 10.1002/joc.706.

[85] K. G. Hubbard et al. “Air temperature comparison between the MMTS and the USCRN temperature systems”. *J. Atmos. Oceanic Tech.* 21 (2004), pp. 1590–1597. DOI: 10.1175/1520-0426(2004)021<1590:ATCBTM>2.0.CO;2.

[86] R. Mahmood, S. A. Foster, and D. Logan. “The Geoprofile metadata, exposure of instruments, and measurement bias in climatic record revisited”. *Int. J. Clim.* 26 (2006), pp. 1091–1124. DOI: 10.1002/joc.1298.

[87] C. Genthon et al. “Atmospheric temperature measurement biases on the Antarctic plateau”. *J. Atm. Oceanic Tech.* 28 (2011), pp. 1598–1605. DOI: 10.1175/JTECH-D-11-00095.1.

[88] R. A. Sr. Pielke et al. “Documentation of uncertainties and biases associated with surface temperature measurement sites for climate change assessment”. *Bull. Amer. Meteor. Soc.* 88 (2007), pp. 913–928. DOI: 10.1175/BAMS-88-6-913.

[89] T. C. Peterson and D. R. Easterling. “Creation of homogeneous composite climatological reference series”. *Int. J. Clim.* 14 (1994), pp. 671–679. DOI: 10.1002/joc.3370140606.

[90] W. M. Wendland and W. Armstrong. “Comparison of maximum-minimum resistance and liquid-in-glass thermometer records”. *J. Atmos. Oceanic Tech.* 10 (1993), pp. 233–237. DOI: 10.1175/1520-0426(1993)010<0233:COMRAL>2.0.CO;2.

[91] R. Heino. “Homogeneity of the long-term urban data records”. *Atm. Environ.* 33 (1999), pp. 3879–3883. DOI: 10.1016/S1352-2310(99)00130-2.

[92] S. A. Changnon and K. E. Kunkel. “Changes in instruments and sites affecting historical weather records: A case study”. *J. Atm. Ocean. Tech.* 23 (2006), pp. 825–828. DOI: 10.1175/JTECH1888.1.

[93] D. G. Baker. “Effect of observation time on mean temperature estimation”. *J. Appl. Meteor.* 14 (1975), pp. 471–476. DOI: 10.1175/1520-0450(1975)014<0471:EOOTOM>2.0.CO;2.

[94] D. A. Robinson. “Gathering climatic data of the highest quality”. Proceedings of the 10th Conference on Applied Climatology. Reno, NV, Am. Meteor. Soc. 1997.

[95] K. I. Scott, J. R. Simpson, and E. G. McPherson. “Effects of tree cover on parking lot microclimate and vehicle emissions”. *J. Arboriculture* 25 (1999), pp. 129–142. URL: http://joa.isa-arbor.com/.

[96] K. Gallo, D. R. Easterling, and T. C. Peterson. “The influence of land use/land cover on climatological values of the diurnal temperature range”. *J. Clim.* 9 (1996), pp. 2941–2944. DOI: 10.1175/1520-0442(1996)009<2941:TIOLUC>2.0.CO;2.

[97] K. P. Gallo et al. “Temperature trends of the U.S. Historical Climatology Network based on satellite-designated land use/land cover”. *J. Clim.* 12 (1999), pp. 1344–1348. DOI: 10.1175/1520-0442(1999)012<1344:TTOTUS>2.0.CO;2.

[98] H. Alexandersson and A. Moberg. “Homogenization of Swedish temperature data. Part I: Homogeneity test for linear trends”. *Int. J. Clim.* 17 (1997), pp. 25–34. DOI: 10.1002/(SICI)1097-0088(199701)17:1<25::AID-JOC103>3.0.CO;2-J.

[99] L. A. Vincent. “A technique for the identification of inhomogeneities in Canadian temperature series”. *J. Clim.* 11 (1998), pp. 1094–1104. DOI: 10.1175/1520-0442(1998)011<1094:ATFTIO>2.0.CO;2.

[100] J. Hansen et al. “GISS analysis of surface temperature change”. *J. Geophys. Res. D* 104 (1999), pp. 30997–31022. DOI: 10.1029/1999JD900835.

[101] L. A. Vincent and D. W. Gullett. “Canadian historical and homogeneous temperature datasets for climate change analysis”. *Int. J. Clim.* 19 (1999), pp. 1375–1388. DOI: 10.1002/(SICI)1097-0088(199910)19:12<1375::AID-JOC427>3.0.CO;2-0.

[102] L. A. Vincent et al. “Homogenization of daily temperatures over Canada”. *J. Clim.* 15 (2002), pp. 1322–1334. DOI: 10.1175/1520-0442(2002)015<1322:HODTOC>2.0.CO;2.

[103] T. C. Peterson et al. “Homogeneity adjustments of in situ atmospheric climate data: A review”. *Int. J. Clim.* 18 (1998), pp. 1493–1517. DOI: 10.1002/(SICI)1097-0088(19981115)18:13<1493::AID-JOC329>3.0.CO;2-T.

[104] A. C. Costa and A. Soares. “Homogenization of climate data: Review and new perspectives using geostatistics”. *Math. Geosci.* 41 (2009), pp. 291–305. DOI: 10.1007/s11004-008-9203-3.

[105] P. Domonkos. “Efficiency evaluation for detecting inhomogeneities by objective homogenisation methods”. *Theor. Appl. Clim.* 105 (2011), pp. 455–467. DOI: 10.1007/s00704-011-0399-7.

[106] V. K. C. Venema et al. “Benchmarking homogenization algorithms for monthly data”. *Clim. Past* 8 (2012), pp. 89–115. DOI: 10.5194/cp-8-89-2012.

[107] H. Alexandersson. “A homogeneity test applied to precipitation data”. *J. Climatol.* 6 (1986), pp. 661–675. DOI: 10.1002/joc.3370060607.

[108] D. A. Robinson. “The United States Cooperative climate-observing systems: Reflections and recommendations”. *Bull. Amer. Meteor. Soc.* 71 (1990), pp. 826–831. DOI: 10.1175/1520-0477(1990)071<0826:TUSCCO>2.0.CO;2.

[109] R. Lund et al. “Changepoint detection in periodic and autocorrelated time series”. *J. Clim.* 20 (2007), pp. 5178–5190. DOI: 10.1175/JCLI4291.1.

[110] J. Hansen and S. Lebedeff. “Global trends of measured surface air temperature”. *J. Geophys. Res. D* 92 (1987), pp. 13345–13372. DOI: 10.1029/JD092iD11p13345.

[111] P. D. Jones, T. J. Osborn, and K. R. Briffa. “Estimating sampling errors in large-scale temperature averages”. *J. Clim.* 10 (1997), pp. 2548–2568. DOI: 10.1175/1520-0442(1997)010<2548:ESEILS>2.0.CO;2.

[112] A. Toreti et al. “A note on the use of the standard normal homogeneity test to detect inhomogeneities in climatic time series”. *Int. J. Clim.* 31 (2011), pp. 630–632. DOI: 10.1002/joc.2088.

[113] S. N. Rodionov. “A sequential algorithm for testing climate regime shifts”. *Geophys. Res. Lett.* 31 (2004), p. L09204. DOI: 10.1029/2004GL019448.

[114] X. L. Wang. “Comments on “Detection of undocumented changepoints: A revision of the two-phase regression model””. *J. Clim.* 16 (2003), pp. 3383–3385. DOI: 10.1175/1520-0442(2003)016<3383:CODOUC>2.0.CO;2.

[115] C. N. Jr. Williams, M. J. Menne, and P. Thorne. “Benchmarking the performance of pairwise homogenization of surface temperatures in the United States”. *J. Geophys. Res.* in press (2012). DOI: 10.1029/2011JD016761.

[116] P. I. Figuerola and N. A. Mazzeo. “Urban-rural temperature differences in Buenos Aires”. *Int. J. Climatol.* 18 (1998), pp. 1709–1723.