Open Peer Review Journal

Has poor station quality biased U.S. temperature estimates?

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

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

Abstract

Two independent surveys have found that about 70% of the thermometer stations in the U.S. Historical Climatology Network (USHCN) dataset are currently poorly or badly sited. Previous investigations into how this poor siting has affected estimates of U.S. temperature trends have led to apparently contradictory conclusions. However, in this study, these contradictions are resolved, and it is shown that poor station quality has introduced a noticeable warming bias into temperature trend estimates for the U.S.

For the unadjusted station records, this poor siting increased the mean temperature trends by about 32%. When time-of-observation adjustments were applied to the records, this increased temperature trends by about 39%, and so the relative fraction of the trends due to the siting bias decreased. However, the siting biases were still substantial, and increased trends by about 18%.

The step-change homogenization algorithm which had been developed to remove non-climatic biases such as siting biases was shown to be seriously problematic. Instead of correcting the poorly- and badly-sited station records to match the trends of the well-sited stations, it appears to have blended the temperature records of all stations to match the trends of the poorly-sited stations.

It seems likely that similar poor siting biases also exist in global thermometer datasets, and this has probably led to an overestimation of the amount of "global warming" since the 19th century.

Citation:
R. Connolly, and M. Connolly (2014). *Has poor station quality biased U.S. temperature estimates?*, Open Peer Rev. J., 11 (*Clim. Sci.*), ver. 0.1 (non peer reviewed draft). URL: http://oprj.net/articles/climate-science/11

Version: 0.1 (non peer-reviewed)
First submitted: January 8, 2014.
This version submitted: January 31, 2014.

1 Introduction

Two recent surveys[1, 2] of the local surroundings of the thermometer stations in the U.S. Historical Climatology Network (often abbreviated USHCN), have revealed that about 70% of the stations are currently sited in poorly or badly exposed locations. It is well-known that the local environment within a few hundred metres of a thermometer station can lead to unusual "micro-climates", which are unrepresentative of the climate in the surrounding area[3–10]. For instance, nearby trees can reduce sunlight and wind[3, 4] and thermometers kept above asphalt concrete can report temperatures considerably warmer than over soil or grass[5, 6]. As a result, the temperature records of poorly exposed stations are likely to contain non-climatic biases from localised changes in the micro-climate immediately surrounding the thermometer housing.

These "siting biases" or "inadequate station exposure biases" are different from the more widely-studied "urbanization biases" which we discuss in Refs. [11–13]. There are some similarities between both types of bias, e.g., they can both arise as a result of modernization and/or urban development in the area. However, while urbanization bias can physically alter the local climate of a large area, siting biases are strictly confined to the localized micro-climate in the immediate vicinity of the thermometer station. Because of this, the two biases can occur independently of each other, e.g., an urbanized station with a strong urbanization bias may have a very good station exposure, while a rural station with no urbanization bias may have a strong siting bias due to inadequate station exposure. In this study, we will

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

focus on siting biases.

The 1218 stations used for the U.S. Historical Climatology Network were selected according to their spatial coverage, record length, data completeness, and historical stability. For this reason, the U.S. Historical Climatology Network is the main dataset used for calculating monthly or annual temperature trends for the contiguous United States, i.e., all of the United States except Hawai'i and Alaska. It is compiled and maintained by the National Climatic Data Center of the U.S.-based National Oceanic and Atmospheric Administration (NOAA) - referred to henceforth as the National Climatic Data Center.

The U.S. Historical Climatology Network is also included in the National Climatic Data Center's Global Historical Climatology Network (often abbreviated GHCN), which as we discuss elsewhere[13] is the main weather record dataset used for estimating *global* temperature trends. The stations in the U.S. network account for nearly 17% of the stations in the global network. But, more importantly, they account for the vast majority of the stations in the global network which are both rural *and* have relatively long, complete station records. For instance, 219 of the 225 stations (i.e., 97.3%) in the Global Historical Climatology Network that are identified as rural in terms of both night-light brightness and associated populations, and have data for at least 95 of the last 100 years are in the U.S. network. These long, rural records are the ones least likely to be affected by urbanization bias - a systematic bias which we argue in Refs. [11–13] has introduced an artificial warming trend into weather station-based global temperature trend estimates.

For these reasons, the U.S. network is an important part of any analysis of global temperature trends. Hence, problems in the reliability of the U.S. Historical Climatology Network have implications not just for regional U.S. temperature trend estimates, but also for global temperature trend estimates. In this study we attempt to estimate the net sign and magnitude of the non-climatic biases introduced into the U.S. temperature records by inadequate station exposure.

Our analysis is based on the results of the Surface Stations project carried out by Watts et al. [1], and suggests that inadequate station exposures have indeed introduced a noticeable warming bias into U.S. temperature trends. It is probable that similar biases exist in global temperature trend estimates. Several studies have previously attempted to construct estimates of the bias from the Surface Stations results, but have each reached different conclusions[2, 14–17]. We will attempt to rationalise the apparent contradictions between the different analyses (including ours).

The layout of this article is as follows. In Section 2, we will summarise the results of the Surface Stations project, and review the current literature on how station quality can influence the station temperature trends. In Section 3, we will present our analysis of the poor station quality problem using the Surface Stations results. In Section 4, we will compare our analysis with the previous studies of the Surface Stations results, and discuss the reasons for the different conclusions between these studies. In Section 5, we will make some practical suggestions and recommendations that we believe could substantially improve the quality of the available temperature record datasets. Finally, we will offer some concluding remarks in Section 6.

2 Literature review

2.1 Motivation for the Surface Stations project

The stations in the U.S. Historical Climatology Network were taken from a larger network called the NOAA Cooperative Observer Program Network, (henceforth, the "COOP Network"), which is a volunteer-run weather observation program for the U.S. Since the stations are mostly volunteer-run, sometimes the official recommendations provided to the observers by NOAA National Weather Service on how to maintain the station are overlooked. This has led some researchers to speculate that the exposure of some of the thermometer shelters may be inadequate. For instance, Robinson, 1990[18] noticed that some weather observers had dramatically altered the exposure of their thermometer shelter when they switched to using electronic thermometers. He cautioned that this may have biased the temperature records.

In 2002, Davey & Pielke, 2005[8] carried out onsite inspections of 57 COOP stations (including 10 Historical Climatology Network stations) in eastern Colorado. They found that many of the stations were poorly exposed. Some stations were located beside air conditioner exhausts, some were surrounded by trees and/or buildings and some were set up over a gravel surface instead of grass. Vose et al., 2005[19] and Peterson, 2006[20] argued that the National Climatic Data Center's homogenization adjustments which were applied to some of the releases of the U.S. Historical Climatology Network had already removed (or at least reduced) any such biases. But, Pielke et al., 2007b[10] questioned the reliability of the adjustments.

Vose et al., 2005[19] had also claimed that Davey & Pielke's study was over too small a part of the U.S. to assume it was a widespread problem. However, Mahmood et al., 2006 showed that inadequate station siting was also a systematic problem for stations in Kentucky[6]. So, in 2007, Watts decided to extend Davey & Pielke's study. Together with a group of more than 650 volunteers, he began visual and photographic on-site inspections of all of the U.S. Historical Climatology Network stations, with his Surface Stations project. In Section 2.2, we will summarise their findings and in the following sections we will discuss the implications of these findings.

2.2 Summary of the Surface Stations findings

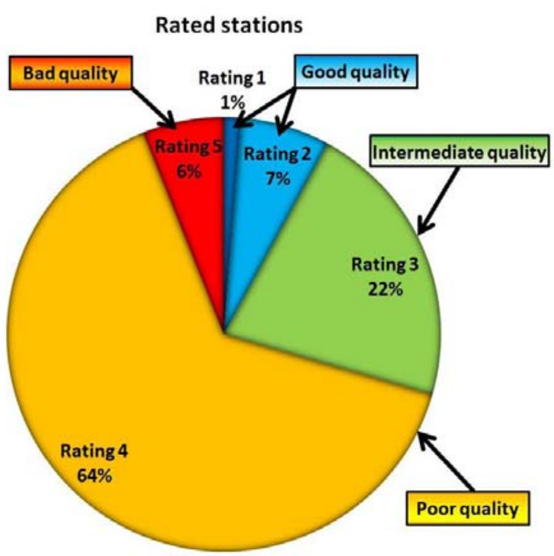

Figure 1: *The relative percentage of different exposure ratings in the U.S. Historical Climatology Network stations that have been rated by the Surface Stations project.*

The Surface Stations project used the rating scheme followed by the National Climatic Data Center when they were establishing a high quality weather station network called the United States Climate Reference Network (USCRN) in 2002. This scheme was described in a NOAA technical document[21] and was based on a scheme proposed by Leroy, 1999[22].

The main parameter by which the stations were classified was the distance of the thermometer sensor from artificial heating sources (or reflecting surfaces), such as buildings, concrete surfaces or parking lots.

Rating 1 There are no artificial heating sources within at least 100 m of the sensor.

Rating 2 There are artificial heating sources within 30 to 100 m of the sensor.

Rating 3 There are artificial heating sources within 10 to 30 m of the sensor.

Rating 4 There are artificial heating sources less than 10 m from the sensor.

Rating 5 There are artificial heating sources located next to, or below, the sensor.

The higher the rating number, the more likely it is that temperature readings have been biased by unrepresentative micro-climate conditions. Ratings 1 and 2 have excellent or good site exposure, and so are unlikely to be biased by micro-climate conditions, while measurements at Rating 5 sites are likely to be dominated by micro-climate biases. If a station is strongly influenced by micro-climate changes, then this reduces the reliability of its temperature records for considering climatological trends. Hence, for this analysis we define stations with ratings of either 1 or 2 to be *"good quality"*, those with a rating of 3 as *"intermediate quality"*, those with a rating of 4 as *"poor quality"* and those with a rating of 5 as *"bad quality"*.

Leroy, 1999 also recommended that all stations should be located on flat and horizontal ground, surrounded by a clear surface with a slope of less than 19°. He recommended that the stations should also be far from large bodies of water, unless they are representative of the area, in which case the station should still be located at least 100 m away[22]. However, he did not indicate how to modify the ratings if stations did not meet those requirements, so these factors do not appear to have been included in the Surface Stations rating system.

Additional requirements for a station to be classified with Rating 1 were that the surrounding grass/low vegetation ground cover is less than 10 cm high, and that the sensor stops being shaded once the sun reaches an elevation of 3° or lower. If either of these requirements are broken, then the best rating the station can receive is Rating 2. However, in this analysis, since only about 1% of the stations have Rating 1, we group both Ratings 1 and 2 together, and so do not consider the vegetation and shading factors.

The Surface Stations group have archived ratings for 1007 out of the 1218 stations (82.7%) which we downloaded on 8th March 2012 from their website at `http://www.surfacestations.org`. We show the relative percentages of these ratings in Figure 1, and their locations in Figure 2. Examples of stations from each of the four subsets are shown in Figure 3. It can be seen that only about 8% of the stations have a good quality rating (1 or 2), and the majority of the rated stations have poor or bad quality ratings of 4 or 5 (70%).

For a few of the unrated stations, the Surface Stations team were able to determine that the rating changed substantially in recent years, and so, they were not assigned a rating. For instance, Malad City, ID (105559) would currently have a rating of 4, but only since 2008, while the Houma, LA (164407) station would have had a rating of 3 during the period 2004-2007. This confirms that station qualities change over time, and so it should be recognised that the current Surface Stations ratings are only based on the station quality they had at the time of the surveys. Obviously, it would be preferable if similar assessments were also available for early periods. Nonetheless, we will see that it is still possible to make useful retrospective assessments of the poor station quality problem from just the current ratings.

In response to the Surface Stations findings, NOAA National Weather Service carried out independent reassessments of 276 of the stations. Their reassessments confirmed that the Surface Stations ratings were reasonably accurate, and that poor siting is indeed a systemic problem in the U.S. Historical Climatology Network[2].

Watts, 2009 provided photographs of many of the stations and it can be seen that, in many cases, station thermometers were situated near (or over) asphalt roads or parking lots, and beside buildings, sometimes beside the exhaust fans of air conditioning units, amongst many other problems[1]. This explains why so many of the stations received a bad rating. All of these problems could easily have biased the station records, and so it is important to reliably account for any such biases.

2.3 Previous assessments of the Surface Stations findings

When the Surface Stations project had rated most of the stations, Watts published a photographic report illustrating the remarkably high occurrence of poor quality siting in the U.S. Historical Climatology Network[1]. He discussed how it was plausible that this poor quality siting had introduced artificial warming trends into many of the station records, and that this would have introduced warming biases into the regional U.S. temperature trends which had been calculated from the Historical Climatology Network, e.g., Ref. [23]. He also suggested that similar problems could exist for the rest of the Global Historical Climatology Network. He did not attempt to quantify this proposed bias, however.

At the time of writing, at least five studies (aside from our own) have attempted to quantify the extent of this bias, by using the Surface Stations results[2, 14–17]. However, before we discuss these studies, it is important to briefly consider the different temperature datasets available for the stations.

The National Climatic Data Center provide three different releases of their U.S. Historical Climatology Network datasets[24], which differ in the degree to which they have been adjusted for potential non-climatic biases.

One of their releases is essentially unadjusted, but has undergone a set of quality control checks to remove individual monthly values which appear erroneous, e.g., monthly temperatures that are far greater than (or far less than) the seasonal average for that station. We will refer to this release as the *"Unadjusted"* dataset.

In a second release, they have also applied specific adjustments to the temperature records for each station to account for documented changes in the time of day that the observers made their measurements[25, 26]. We will refer to this release as the *"Time of observation adjusted"* dataset.

For their third release, they also carry out a series of station-station inter-comparisons to identify and remove station-specific non-climatic step-changes[27]. We will refer to this release as the *"Time of observation and step-change adjusted"* dataset. The National Climatic Data Center also apply inter-station interpolation to "fill in" any missing gaps in the station records for the third release.

Although there is now a general acceptance that the findings of the Surface Stations project are accurate, and that the majority of the stations in the

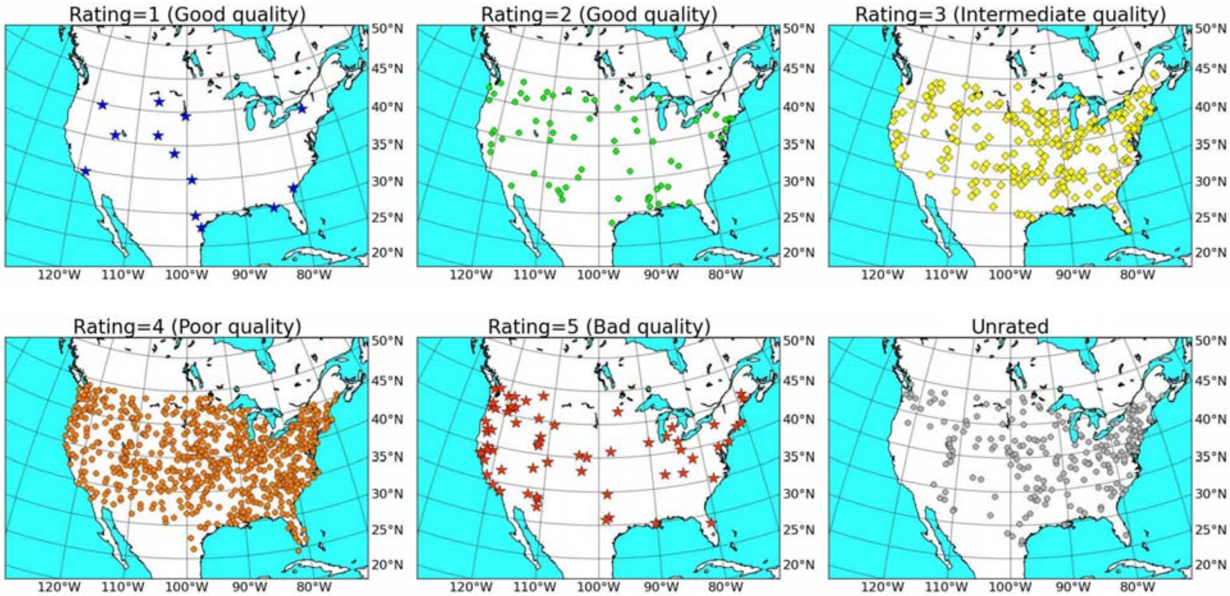

Figure 2: *Location of stations in the U.S. Historical Climatology Network with different siting quality ratings, as well as the remaining unrated stations.*

U.S. Historical Climatology Network are of a poor or bad quality, there has been considerable debate over exactly what biases, if any, this poor siting has actually introduced into the reported U.S. temperature trends.

Menne et al., 2010 suggested that the National Climatic Data Center's homogenization procedure[24] had already accounted for any biases which poor siting may have introduced. Moreover, they suggested that if there was any residual bias, it was probably a slight *cooling* bias[2], rather than the warming bias Watts, 2009 had suggested.

Muller et al., 2013[15] claimed that the linear trends of the *Unadjusted* records for stations with Ratings 1, 2 or 3 were comparable to those of stations with Ratings 4 or 5, and that there was not much difference between estimates constructed from the Ratings 1-3 and Ratings 4-5 subsets of the USHCN. Therefore, they concluded that poor siting did not have much effect on the temperature trends.

Martinez et al., 2012 used the Surface Stations ratings in their analysis of temperature trends for the state of Florida (USA)[16]. For their study, they used the *Time-of-observation and step-change adjusted* dataset. As they were only studying the trends for Florida, their study only involved 22 Historical Climatology Network stations. So, they were cautious about drawing definitive conclusions on the effects of poor station quality on temperature trends. Nonetheless, they found the linear trends were different for the subsets of the worst rated (4 & 5) and best rated (1 & 2) over the two periods they considered - 1895-2009 and 1970-2009. From this they concluded that station quality *does* influence temperature trends. However, they were unclear as to the sign of this influence. For the 1895-2009 period, the poor quality stations showed a greater warming trend in the mean monthly temperatures than the good quality stations, while for the 1970-2009 period, the reverse applied.

Fall et al., 2011[14] agreed with Menne et al., 2010 that the National Climatic Data Center's homogenization adjustments reduced much (although not all) of the difference between the good quality and poor quality subsets. But, they argued that the *Unadjusted* and *Time-of-observation adjusted* trends of the poor quality stations showed a substantial warming bias relative to the good quality stations. They also argued that poor exposure led to biases in Diurnal Temperature Range (DTR) trends.

The Diurnal Temperature Range is the difference between the mean maximum daily temperatures and the mean minimum daily temperatures. Although it is not as widely studied as the mean temper-

(a) *Fallon, Nevada (Good quality)*

(b) *Boulder, Colorado (Intermediate quality)*

(c) *Napa State Hospital, California (Poor quality)*

(d) *Santa Rosa, California (Bad quality)*

Figure 3: *Examples of stations from each of the four subsets - Good quality (Fallon; ID = 262780, Rating 2); Intermediate quality (Boulder; ID = 050848, Rating 3); Poor quality (Napa State Hospital; ID = 046175, Rating 4); Bad quality (Santa Rosa; ID = 047965, Rating 5). Photographs were downloaded from* http://surfacestations.org/*. The photographer for these four stations was Anthony Watts.*

ature trends, there has been considerable interest in Diurnal Temperature Range trends, partly because it is thought they can provide insight into the cause of mean temperature trends, e.g., Refs. [28–30]. Fall et al. calculated that the worst-sited stations implied a Diurnal Temperature Range trend of $-0.4°C$/century, whereas the best-sited stations had essentially no long-term trend.

Watts et al. (in preparation, 2012)[17] argued that the original Leroy et al., 1999[21, 22] rating system used by Watts, 2009 was not rigorous enough, and they re-evaluated the station exposures using the recommendations of Leroy et al., 2010[31]. When they applied this new rating system to the stations, they found a greater difference between the poor quality and good quality stations than before (for the *Unadjusted* records), with the poorly-sited stations showing more warming. They questioned the reliability of the National Climatic Data Center's homogenization adjustments, and suggested that a combination of poor station exposure, urbanization bias and unreliable homogenization adjustments had led to a spurious doubling of U.S. mean temperature trends over the period 1979-2008.

2.4 Exposure problems for the early instrumental period

Before we discuss our re-analysis of the Surface Stations findings, it is worth discussing a related station exposure problem which has also received some discussion in the literature lately. Namely, there have been several attempts to estimate the extent to which changes in the types of thermometer shelters used by weather observers have biased 18th, 19th and early 20th century temperature records, relative to modern records[7, 32–44].

During the 20th century, a lot of the thermometers used by weather observers were housed in outdoor shelters like the Stevenson screen[1]. In recent decades, many stations have switched to using automated thermometer systems. Although there may have been instrumental biases from the switch in instrumentation during the recent move to automation[45–47] and some of the station observers also seem to have reduced the quality of the station exposure[1, 2, 18], the new automated stations are also outdoor shelters. However, before the introduction of the Stevenson-type screens, thermometers had quite different station exposures. For instance, in the early 1700s, English weather observers were encouraged to keep their thermometers unscreened and *indoors* in well-ventilated, north-facing, fireless rooms[7].

Chenoweth, 1992[32] and 1993[33] noted that many U.S. observers in the late 19th century and early 20th century were using similar unscreened, north-facing, thermometers, and even when screens were introduced, there was a wide range of different types of screens. In many cases, the siting of the screen was often inappropriate, e.g., attached to a wall[32], and would have received a bad or poor rating under the Surface Stations project. In some cases, thermometers located at railway stations, frequently registered artificially high temperatures when trains arrived at the station, but also sometimes gave minimum temperatures that were too low if the arriving train shook the thermometer[32].

Parker, 1994[34] compiled some information on the thermometer exposures *recommended* by different countries between the mid-19th and mid-20th centuries, and found that the recommended thermometer exposures often varied dramatically over time. Unfortunately, most of the available information is rather limited, and establishing the *actual thermometer exposures used* has been contentious. For example, there has been considerable debate over when the use of Stevenson screens first became widespread in Australia[35–39].

It is quite likely that these changes in thermometer shelters introduced biases of some sort (sometimes called "shelter biases", "screen biases" or "early instrumental period biases"). However, establishing what those biases would have been is difficult. The biases introduced into daily averages by different thermometer exposures could depend on a number of factors, e.g., the observation time and averaging method used; the degree of urbanization of the site; the materials the nearby buildings were constructed from; the station siting of both the old and new exposures.

Some studies have attempted to estimate these biases by comparing temperature measurements recorded simultaneously at the same site by thermometers with different shelters[32–34, 40, 43, 44]. In some cases, these studies were carried out during the actual change-over, i.e., the late-19th/early 20th century[33, 34, 40]. However, in other cases, they are modern attempts to recreate the transition[32, 33, 40, 43, 44].

We note that these experiments are not as straightforward as commonly assumed. For instance, Chenoweth, 1992 calculated different estimates of the screen bias of a north-facing window thermometer relative to a thermometer in a Cotton Region Shelter (a similar shelter to the Stevenson screen), depending on which Cotton Region Shelter he used. For his estimates, he had two different shelters, one located in his backyard, and the other in a nearby field. From the photographs in Chenoweth, 1992, it appears to us that one of the shelters would have had a good quality Surface Stations rating, while the other would have had a poor quality rating[32]. This suggests that any screen biases would have depended on the siting of the old and new thermometers.

We also note that many of the stations with long records are currently more urbanized than they would have been in the 18th/19th centuries. Since urbanization bias as well as other land use changes are known to affect the diurnal temperature range at a station[29], it is possible that estimates of different screen biases carried out using modern field tests may over/underestimate the actual biases. For instance, we used Google Earth to analyse aerial photographs of the study sites used by Brunet et al., 2011[44] (see the Supplementary Information for the Google Earth station location files and some aerial photographs).

[1] Invented in the 19th century by Sir Thomas Stevenson, the father of the well-known author, Robert Louis Stevenson.

Both of the sites Brunet et al. used appear to be highly urbanized, and the local environment has clearly undergone dramatic changes since the 19th century. For this reason, it is likely that the current "screen bias" is different than it would have been in the 19th century/early 20th century when the Montsouri shelters (which Brunet et al., 2011 were assessing) were actually in use. We note that the bias appears to be greater at the more urbanized of the two stations (Murcia). From the aerial photographs, it also appears that both sites would receive a poor or bad Surface Stations rating, however, it is possible that this was also the case for the historic sites.

In contrast, the site used for the Böhm et al., 2010 study[43] is located in a relatively rural location, at a Benedictine monastery in Kremsmünster, Austria. However, while it is a useful comparative study, their estimates of the screen bias are not necessarily representative.

Their old window thermometer is located in a tall astronomical tower with panoramic views of the surrounding town. Meanwhile, the new shelter is in a heavily shaded garden in front of the building, surrounded on all four sides by either tall buildings or trees.

Shelters exposed in the shade will tend to register lower maximum temperatures than properly exposed shelters[32]. So, if their new shelter was too heavily shaded, then this would have exaggerated Böhm et al.'s estimate of the screen bias. Analysis of the location of the new shelter suggests it does suffer from shading problems. For instance in the 8th August 2012 aerial Google Earth photograph for the location (see Supplementary Information), the shelter was in shadow, and it was also in shadow in the photograph shown in Böhm et al., 2010, which was taken on 21st March 2007[43].

Another potential problem in estimating the shelter biases is that the observation times and averaging methods used by observers have also changed over time. As we will discuss in Section 4.4, different observation times and averaging methods can lead to different estimates of the daily mean temperature at a station. So, the biases introduced by changes in the thermometer shelter used would have depended on which averaging method the observers were using.

In particular, we note that the apparent biases often seem to be greatest for the minimum and maximum daily temperatures, and could have been smaller for some of the observation times which might have been used. Many modern weather observers use minimum-maximum thermometers and approximate the daily mean temperature by calculating the mean of the maximum and minimum temperatures reached during the previous 24 hours. But, especially in the 18th and 19th centuries, many observers would have measured the temperatures at specific times in the day, and estimated the daily average using those measurements. When the different shelters were being developed in the 19th century, there was considerable awareness of the biases that different thermometer exposures introduced[33, 34, 40]. So, it is plausible that, in some cases, the averaging formulae used were partially chosen in an attempt to minimise the screen biases.

For all these reasons, we find the current estimates of these screen biases are still incomplete, and require more careful studies. Despite this, several researchers believe that screen biases have led to a significant overestimation of pre-20th century summer temperatures[34, 37, 39–44]. This is considered a particular problem for the longest thermometer records, which are mostly European. One of the main reasons for this belief appears to be because temperature proxies for the same areas suggest colder 18th century temperatures than the thermometer records[41–43]. We review this *"convergence problem"* issue in a separate paper[48], and a detailed comparison of temperature proxies and thermometer records is beyond the scope of this article. However, we will note that it seems ironic that, while some researchers are arguing that early thermometer records are unreliable because the temperature proxies show colder temperatures, other researchers are arguing that temperature *proxies* are unreliable for the late 20th century, because they do not show the warm temperatures of the thermometer records[49, 50], i.e., the so-called *"divergence problem"*.

3 Our reanalysis

For our analysis of the siting biases in the U.S. Historical Climatology Network, we first downloaded the station ratings from the Surface Stations website (on 8th March 2012). Ratings were available for 1007 out of the 1218 stations (82.7%). We then divided the stations with ratings into four subsets, i.e., good quality, intermediate quality, poor quality and bad quality, as described in Section 2.2. For each of these subsets we calculated the gridded mean temperature trends using the three different U.S. Historical Climatology Network datasets, i.e., the *Unadjusted*, *Time-*

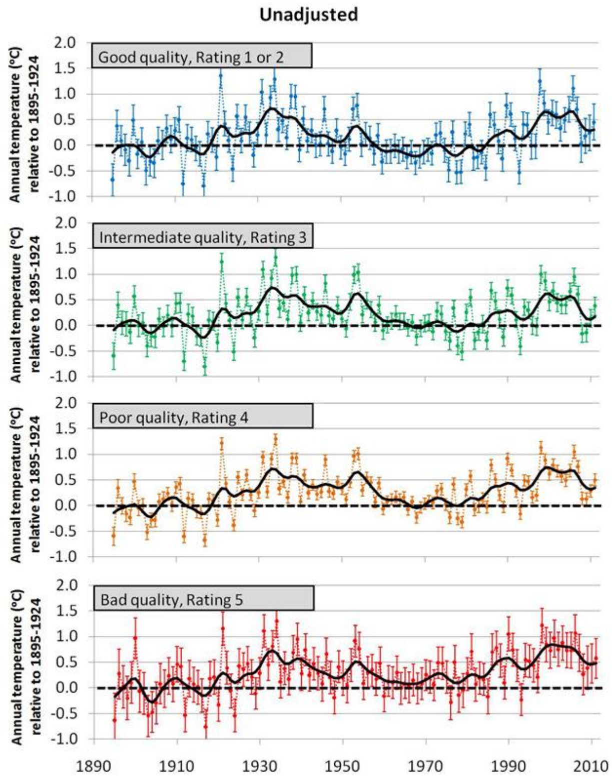

Figure 4: *Mean gridded trends for each of the four subsets using the* Unadjusted *dataset. Solid lines correspond to the 11 point binomial smoothed versions of the annual values. Confidence errors correspond to twice the standard error of the annual means.*

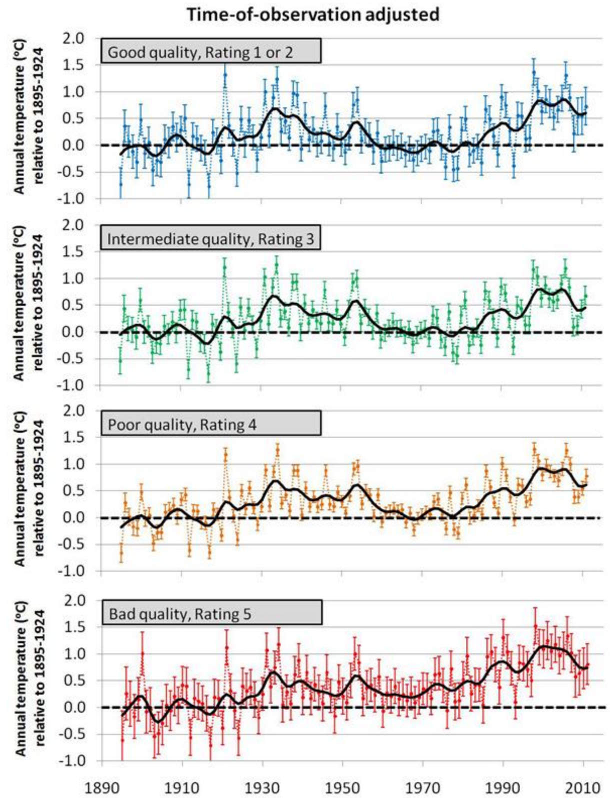

Figure 5: *Mean gridded trends for each of the four subsets using the* Time-of-observation adjusted *dataset. Solid lines correspond to the 11 point binomial smoothed versions of the annual values. Confidence errors correspond to twice the standard error of the annual means.*

of-observation adjusted and *Time-of-observation and step-change adjusted* datasets.

To calculate the gridded mean trends we adopted a similar procedure to the one we used in Refs. [11–13]. Namely, we first converted all of the station annual temperature records for a given subset into temperature anomalies relative to the mean station temperature during a 30 year baseline period, 1895-1924. As we will discuss in Section 4.1, this differs from the 1961-1990 baseline period we used in Refs. [11–13], because we wanted to better study the divergence between subsets over time. Stations which did not have at least 5 years of data during this period were dropped from our analysis. This also differs from the 15 year minimum we used in Refs. [11–13]. As we will discuss in Section 4.1, this was because we wanted to reduce the number of stations dropped from our analysis. 87 of the 1218 stations ($\sim$ 7%) were dropped from our analysis for this reason.

Stations were then assigned to $5° \times 5°$ grid boxes. For each year, the mean temperature anomalies for each of the grid boxes were calculated by determining the simple mean of the temperature anomalies of all the stations in that box with data for that year.

The mean U.S. temperature anomaly for each year was then calculated as the area-weighted mean of all of the grid box means. Standard errors of the mean were also calculated using the method described in the Supplementary Information.

The mean U.S. temperature trends for each of the four subsets, using each of the three datasets are shown in Figures 4, 5 and 6.

The different estimates of U.S. temperature trends all have a lot of similarities, but there are also substantial differences between them. In terms of the similarities, one striking feature is the pronounced alternation between "warming" periods and "cooling"

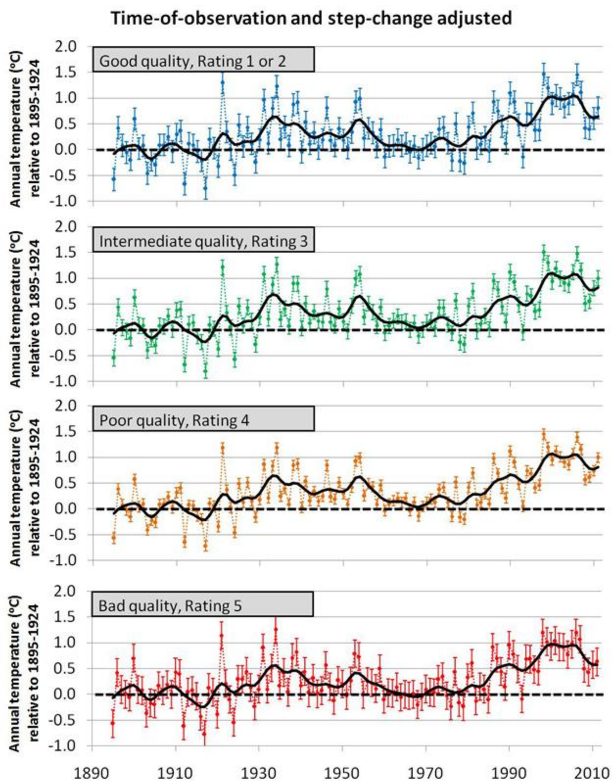

Figure 6: *Mean gridded trends for each of the four subsets using the* Time-of-observation and step-change adjusted *dataset. Solid lines correspond to the 11 point binomial smoothed versions of the annual values. Confidence errors correspond to twice the standard error of the annual means.*

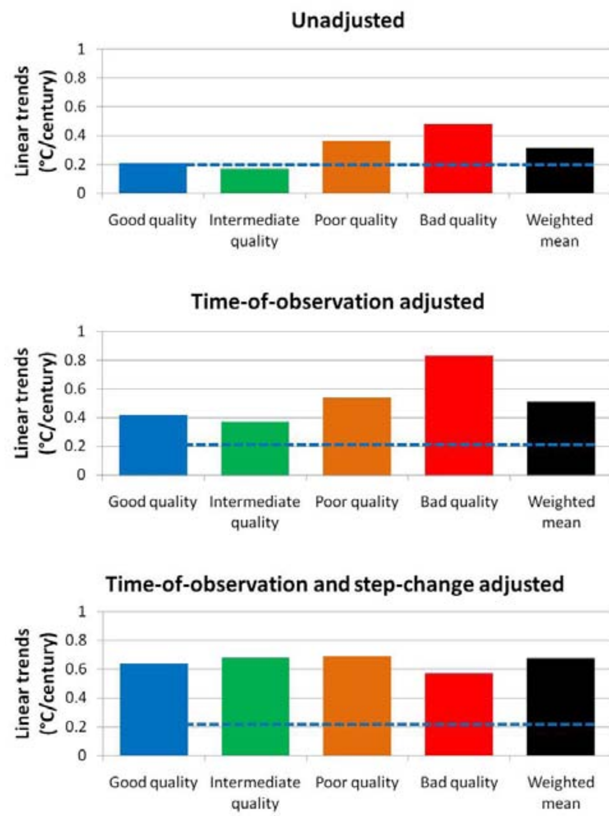

Figure 7: *Bar charts showing the 1895-2011 linear trends of each of the subsets, and the weighted mean of the subsets, for each of the three datasets. The dashed blue lines correspond to the linear trend of the* Unadjusted *good quality subset.*

periods, each lasting several decades. Since the start of the estimates in 1895, there seem to have been two warm periods (1920s-1930s and 1990s-2000s) and two cool periods (1900s-1910s and 1960s-1970s).

One of the main differences between the estimates is in how the warm periods and cool periods compare to each other. For the good quality and intermediate quality, *Unadjusted* subsets, the early warm period appears comparable to the recent warm period, and the early cool period appears comparable to the recent cool period. However, as the quality of the subsets decreases, or as the datasets become more heavily adjusted, the relative warmth of the early warm period appears to decrease, and the recent cool period appears warmer.

It is worth noting that, if the less heavily-adjusted, good quality subsets are reliable, then the recent warming in the U.S. does not appear unusual in the context of the overall record. This would contradict the popular perception that, because of man-made global warming, the recent warm period is the hottest on record in the U.S.[51–53]. The 1930s coincided with a period of drought in parts of the U.S. leading to the economically disastrous "Dust Bowl era"[54], so it is quite plausible that this period of drought also corresponded to a warm period. However, there are many non-climatic biases present in long-term temperature records like the U.S. Historical Climatology Network. So, **if** the adjustments applied by the National Climatic Data Center have successfully removed these biases, then this would make the adjusted datasets more reliable. For this reason, we will consider the implications of all three datasets. We will assess the reliability of the National Climatic Data Center's adjustments in Sections 4.4 and 4.5.

The temperature trends of the estimates show strongly non-linear behaviour. So, as we will discuss in Section 4.2, calculating the linear trends for the estimates is a very crude method for describing the temperature trends. Nonetheless, it offers a simple metric which allows us to make rough comparisons between the different estimates.

Table 1 lists the 1895-2011 linear trends for each of the estimates, as calculated by linear least squares fitting. The linear trends are also shown graphically in Figure 7. The corresponding r^2 fitting parameters are also shown in Table 1, and it can be seen that most of the linear trends are very poor fits. This is as expected, since the trends are quite non-linear. So, the linear trend values should be treated cautiously.

We first note that all of the subsets have a long-term "warming" trend, i.e., they all have positive linear trends for the 1895-2011 period. However, as we mentioned above, this is not surprising, or particularly informative. The 1890s-1910s were a relatively cool period and the 1980s-2000s were a relatively warm period, so it would be expected that all estimates should have a positive trend.

Second, the net effect of each of the adjustments applied by the National Climatic Data Center is to substantially increase the linear trends of the estimates. For example, the *Unadjusted* good quality subset has a linear trend of $+0.21°C$/century, but when the time-of-observation adjustments are applied, this doubles to $+0.42°C$/century, and it increases by a similar amount after the step-change adjustments to $+0.64°C$/century. Similar increases occur for all subsets. The only exception is that the step-change adjustments reduce the linear trend of the bad quality subset.

The next thing we note is that there are indeed substantial differences between the linear trends of the different subsets. For the *Unadjusted* and *Time-of-observation adjusted* datasets, the good quality and intermediate quality subsets have the lowest linear trends and the trend consistently increases from the intermediate quality to the poor quality to the bad quality subsets. This suggests that the biases due to inadequate station exposure introduce non-climatic warming trends into the temperature records.

The step-change adjustments dramatically reduce the differences between subsets. There are at least two schools of thought on why this occurs. Some researchers have argued that this is because the step-change adjustments have managed to remove the siting biases at each station[2]. However, a second explanation is that the reduction arises because the siting biases have been blended or "homogenized" together, rather than removed. In Section 4.5, we argue in favour of this second argument. Watts has also favoured this second argument in on-line commentary on his website[55].

In Table 1, we also list the weighted mean linear trend of the rated stations, which is calculated by weighting the trend of each subset by the percentage of stations in that subset. Although these weighted mean trends are less than for the poor and bad quality subsets, they are still greater than the linear trends for the good quality subset. This suggests that the mean U.S. temperature trends are indeed significantly biased by inadequate station exposure.

Surprisingly, the linear trends for the *Unadjusted* and *Time-of-observation adjusted* datasets are actually slightly smaller for the intermediate quality subset than for the good quality subset. This appears to contradict the expectation that the bias should continuously decrease in going from the worst quality subsets to the best quality subsets.

As Muller et al., 2013 suggest, it is plausible that the intermediate quality stations are of a high enough quality that they are unbiased[15]. After all, the only difference between Ratings 2 and 3 is the distance of the station from heating sources, and this distance is at least 10m for Rating 3 stations. So, it might be that 10m is a sufficient distance for a station to be unaffected by inadequate station exposure. In that case, since the sample size of the intermediate quality subset is considerably larger than the good quality subset (see Figure 1), it could be that the trends of the intermediate subset are the most reliable. However, this is not clear, and on physical grounds, we know that the good quality subset is the least likely of the subsets to have biases due to inadequate station exposure. Therefore we will assume that the good quality subset is the best representation of the "unbiased" subset, in terms of station exposure. This gives slightly lower estimates of the biases for the *Unadjusted* and *Time-of-observation adjusted* datasets.

On this basis, we estimate the siting biases for each subset by subtracting the good quality temperature trends of the appropriate dataset from the subset temperature trends. These difference trends are shown in Figure 8. The 1895-2011 linear trends and r^2 fitting parameters are listed in Table 2.

Again, the difference trends show substantial non-linearity. But, it can be seen from Figure 8, that the linear fits do at least offer a rough approximation of the 1895-2011 trend. On this basis, we can cal-

	Unadjusted		*Time-of-observation adjusted*		*Time-of-observation and step-change adjusted*	
Subset	$°C$/century	r^2	$°C$/century	r^2	$°C$/century	r^2
Good (8%)	+0.21	0.03	+0.42	0.10	+0.64	0.23
Intermediate (22%)	+0.17	0.02	+0.37	0.09	+0.68	0.25
Poor (64%)	+0.36	0.10	+0.54	0.19	+0.69	0.28
Bad (6%)	+0.48	0.16	+0.83	0.37	+0.57	0.21
Weighted mean	+0.31		+0.51		+0.67	

Table 1: *Linear trend estimates (°C/century), calculated by linear least squares fitting, of the four subsets and three different datasets shown in Figures 4, 5 and 6. r^2 shows the fitting coefficients, which theoretically can vary from 0 (poor fit) to 1 (perfect fit)*

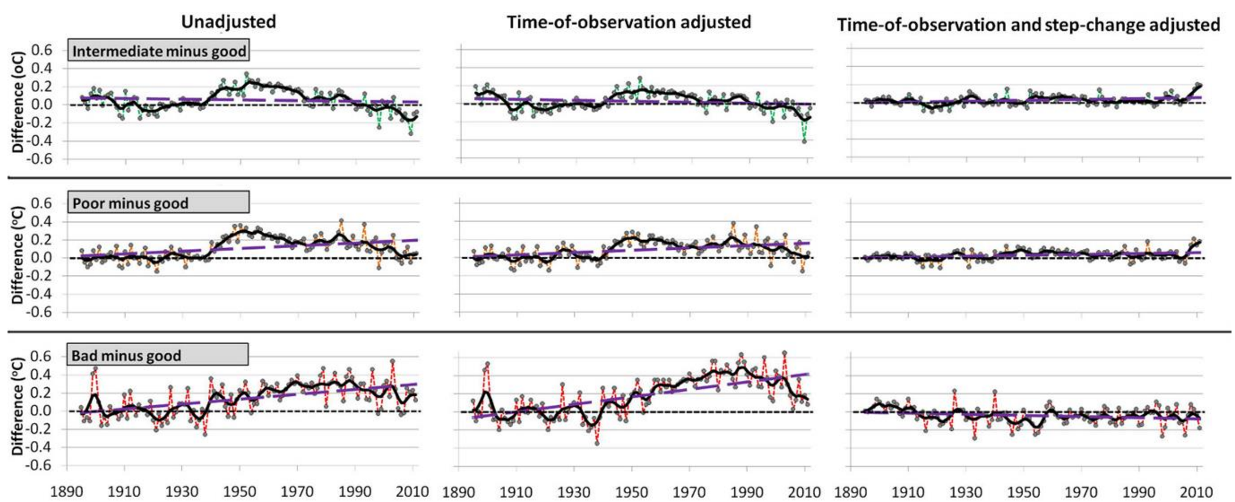

Figure 8: *Annual differences between the good quality subset and each of the other subsets for the three datasets. Solid lines correspond to the differences between the 11 point binomial smoothed versions. The coloured dashed lines correspond to the 1895-2011 linear trends of the differences.*

culate rough estimates for the siting bias of each of the datasets from either Tables 1 or 2. For Table 1, we can do so by subtracting the linear trend of each subset from the linear trend for the good quality subset. This gives the same values as in Table 2, since the linear trend of the differences is equivalent to the difference between the linear trends.

The biases in the *Unadjusted* dataset seem to be about +0.15°C/century for the poor quality subset and +0.27°C/century for the bad quality subset (see Tables 1 or 2). For the set of all the rated U.S. Historical Climatology Network stations, i.e., the weighted mean, the siting bias seems to be about +0.10°C/century. As a percentage of the linear trends, this means that siting biases account for approximately 42% of the poor quality trends, 56% of the bad quality trends and 32% for the entire rated network.

The magnitude of the biases in the *Time-of-observation adjusted* dataset are similar, i.e., +0.13°C/century for the poor quality subset, +0.41°C/century for the bad quality subset and +0.09°C/century for the entire rated network. However, since the time-of-observation adjustments consistently increase the linear trends of all subsets by about +0.2°C/century, the fraction of the trends due to siting bias is less. That is, the percentage of the trends due to siting biases is reduced to 24% of the poor quality subset, 49% of the bad quality subset and 18% of the full rated network. Having said that, the time-of-observation adjustments increase the trends of the bad quality subset by more than any of the other subsets (+0.35°C/century), which increases the difference between the good and bad

	Unadjusted			*Time-of-observation adjusted*			*Time-of-observation and step-change adjusted*		
Subset	$^{\circ}C/100$y	r^2	Bias	$^{\circ}C/100$y	r^2	Bias	$^{\circ}C/100$y	r^2	Bias
Intermediate - good	-0.04	0.01	-24%	-0.05	0.03	-14%	+0.05	0.09	+7%
Poor - good	+0.15	0.16	+42%	+0.13	0.16	+24%	+0.05	0.09	+7%
Bad - good	+0.27	0.27	+56%	+0.41	0.40	+49%	-0.06	0.05	-11%

Table 2: *Linear trends ($^{\circ}C$/century) of the* differences *between the good quality subsets and the other subsets, calculated by least squares fitting from the data shown in Figure 8, for each of the datasets. r^2 shows the fitting coefficients, which vary from 0 (poor fit) to 1 (perfect fit). "Bias" shows the ratio of the linear trend of the difference relative to the linear trend of the subset, as a percentage.*

quality subsets, i.e., our estimate of the siting biases. So, this suggests that siting biases still account for about 49% of the trends of the bad quality subset.

The percentage of the trends in the *Time-of-observation adjusted* subsets that are due to the time-of-observation adjustments are as follows: Good quality = 50% adjustments; intermediate quality = 54% adjustments; poor quality = 33% adjustments; bad quality = 42% adjustments; entire rated network = 39% adjustments. We will consider the time-of-observation adjustments in Section 4.4.

On average, the step-change adjustments also substantially increase the linear trends, by about $+0.16^{\circ}C$/century (for the entire rated network). However, the step-change adjustments are quite different for each of the subsets. In particular, the increase in trend is greatest for the good and intermediate quality subsets ($+0.22^{\circ}C$/century and $+0.31^{\circ}C$/century, respectively), and the adjustments actually decrease the trend for the bad quality subsets (by $-0.26^{\circ}C$/century).

As a result, the differences between the subsets are substantially reduced, and the estimates of the siting biases for the *Time-of-observation and step-change adjusted* dataset are dramatically reduced. Indeed, since the trend of the bad quality subset is now less than the good quality subset, it could be argued, as Menne et al., 2010 did[2], that the bias becomes a slight "cooling" bias ($-0.06^{\circ}C$/century) for the worst sited stations. The linear trend of the good quality subset is still less than for the entire rated network ($+0.64^{\circ}C$/century compared to $+0.67^{\circ}C$/century - see Table 1). So, nominally, it could be argued that there is still a bias of about $+0.03^{\circ}C$/century, but this is less than 5% of the linear trend of the entire rated network.

The problem in assessing the biases in the *Time-of-observation and step-change adjusted* dataset arises from uncertainties over the reliability of the step-change adjustments. The step-change adjustments were developed in an attempt to identify and remove any non-climatic step-change biases in the temperature records[27], e.g., station relocations or changes in instrumentation. So, a plausible explanation for the reduction in the differences between the subsets is that the adjustments have succeeded in removing the non-climatic siting biases, as Menne et al., 2010 contend[2]. However, as Watts notes in on-line commentary[55], another plausible explanation is that the adjustments have blended together the biases from each of the subsets, leading to all subsets being biased by about the same amount. In Section 4.5, we argue that the latter explanation is more likely.

In any case, it appears that siting biases have increased the temperature trends of the entire rated network by about 32% for the *Unadjusted* dataset and about 18% for the *Time-of-observation adjusted* dataset. This is a substantial non-climatic bias, which contradicts the claims of others, e.g., Muller et al., 2013[15]. The linear trends of the bad quality subset are particularly heavily biased by siting bias (56% for the *Unadjusted* dataset and 49% for the *Time-of-observation adjusted* dataset), although it is true that the bad quality subset only makes up 6% of the U.S. Historical Climatology Network (Figure 1).

4 Discussion and comparison with the previous analyses

Initially, it might appear that there are serious contradictions between each of the attempts to use the Surface Stations results to estimate the biases introduced to U.S. temperature trends by inadequate station exposure. However, in this section, we will attempt to reconcile these apparent contradictions, and establish reasonable estimates for the actual siting biases present in the U.S. Historical Climatology Network. In Section 2.3, we discussed the findings of the Watts, 2009[1]; Menne et al., 2010[2]; Muller et al., 2013[15]; Martinez et al., 2012[16]; and Watts et al. (in preparation, 2012)[17], while in Section 3, we discussed the findings of our analysis.

All six of these studies have taken slightly different methodological approaches to analysing the Surface Stations results. So, in Section 4.1, we will first summarise these differences, and assess what influence they might have had on the conclusions of the different studies.

A major problem with analysing the Surface Stations results is that both the temperature trends of the individual subsets and the difference trends between the subsets are non-linear (as we discussed in Section 3). This means that using linear trend analysis is an overly simplistic approach. Nonetheless all six of the studies (including ours) use linear trend analysis. In Section 4.2, we will highlight some of the problems involved in using this approach.

Probably the most difficult challenge in estimating the siting biases lies in the fact that there are many other non-climatic biases present in the U.S. Historical Climatology Network. This means that it is difficult to establish how much of the differences between subsets are actually due to the siting biases, and how much are due to other factors. In Section 4.3, we will consider how one such factor (urbanization bias) could influence the siting bias estimates.

A popular approach to minimising this problem has been to rely on datasets that have been statistically "homogenized" in an attempt to reduce the magnitude of these non-climatic biases. However, this approach will only be successful if the homogenization techniques work. So, it is important to assess the reliability of these techniques. In Section 4.4, we will consider the time-of-observation adjustments, while in Section 4.5, we will assess the step-change adjustments.

4.1 Use of different subsets and averaging approaches

In this section, we will consider some of the main methodological differences between each of the Surface Stations studies, and estimate what impact these differences have on the conclusions of the studies.

- Different approaches to regional averaging

Menne et al., 2010 calculated their regional trends for the contiguous U.S. by first assigning all of the stations in each subset to $0.25° \times 0.25°$ grid boxes, then calculating the mean monthly anomalies for each of those boxes. The grid anomalies were then area-weighted to yield regional anomalies for each of the subsets[2].

Instead of gridding the stations, Fall et al., 2011 divided the contiguous U.S. into nine climatic regions, then calculated the mean monthly anomalies of the different subsets for each of those regions. These regional means were then weighted by the areas of the regions, and averaged to give regional averages for the entire contiguous U.S.[14]. Watts et al. (in preparation, 2012) also adopted this approach[17].

Our approach is probably intermediate between those two approaches. Like Menne et al., 2010, we used an area-weighted uniform gridding. However, in keeping with our global analyses in other papers[11–13], we used a much larger grid size ($5° \times 5°$). This larger grid size still yields about 40 grid boxes for the contiguous U.S. which is more than the 9 regions used by Fall et al. and Watts et al.

The Martinez et al., 2012 study just considered a single state (Florida) which only contained 22 stations. So, they did not attempt gridding and just calculated the monthly means of all the stations in each subset. The area of the state of Florida is less than one of our $5° \times 5°$ grid boxes, so this is probably reasonable.

Fall et al., 2011 argue that the net differences between their approach and Menne et al.'s is very small when averaged over the contiguous U.S.[14]. So, it seems likely that the different gridding approaches used by us, Menne et al., Fall et al., Watts et al. and Martinez et al. are all reasonably equivalent, and therefore have probably not introduced much of a difference between the analyses.

Having said that, Muller et al., 2013 did not use a gridding approach, but instead used a particular statistical interpolation approach. In a separate study, we note that this approach to calculating regional (and global) temperature trends appears to dampen the differences between different station records[11]. We suspect that this is one reason why they were unable to detect much of a difference between their subsets, since other studies[14, 17] (including ours) *were* able to detect significant differences between the subsets in the *Unadjusted* datasets used by Muller et al.

- Different rating data sets

We downloaded the station ratings from the Surface Stations website on 8th March 2012. The ratings available then were the ones used by Fall et al., 2011[14], and were determined from surveys carried out by the Surface Stations volunteers between 2nd June, 2007 and 23rd February, 2010. Ratings were available for 1007 out of the 1218 stations (82.7%). This was also the dataset used by Muller et al., 2013 and Martinez et al., 2010[16]. However, the other studies used slightly different versions of the dataset.

Watts, 2009[1] was based on a preliminary analysis (2nd November, 2009) which only included ratings for 865 stations. However, although the Menne et al., 2010 study was carried out in response to the Watts, 2009 study, it was based on an even earlier (18th April, 2008) provisional dataset, in which only 525 stations had been rated (43.1%), and these ratings had not undergone quality control. According to Watts, Menne et al. were in the process of asking him for the more complete dataset used in the Watts, 2009 study, but stopped when Watts asked to be involved in their study[55].

For the Watts et al. (in preparation, 2012) studies, further refinements and additional station surveys were available (15th June, 2011 - 1st July, 2012)[17]. But, since the Watts et al. study is still in preparation, the newer dataset had not been published on the Surface Stations website at the time of writing.

Although Menne et al., 2010 only used a provisional dataset, their results apparently broadly agree with those of the Fall et al., 2011 and Muller et al., 2013 studies[14, 15]. This suggests that, while their sample may have been somewhat biased by the "low-lying fruit problem", as Watts warned[55], the preliminary dataset was probably sufficiently complete for an approximate estimate. For this reason, it seems that the different sample sizes used by the studies did not majorly influence their results.

In addition, as a prelude to the Menne et al. study, NOAA National Weather Service Forecast Office personnel investigated 276 of the 525 stations (52.6%) and confirmed those ratings from the Surface Stations project[2]. Hence, the Menne et al., 2010 study offers an independent confirmation of the poor quality of the station sitings.

The Watts et al. (in preparation, 2012) study uses the most up-to-date version of the dataset. However, they used a new rating system described by Leroy, 2010[31], which is apparently more rigorous. As a result, they were only able to calculate ratings for 779 of the stations under the new system. This is still a larger sample than the Menne et al., 2010 study, and Watts et al. claim that the new system is able to more accurately distinguish between good and poor quality stations[17]. The new system increases the percentage of stations identified as Rating 1 (6% compared to 1% under the old system); Rating 2 (14% compared to 7%); Rating 3 (32% compared to 22%) and Rating 5 (12% compared to 6%), but decreases the percentage of stations identified as Rating 4 (36% compared to 64%).

Since the ratings under the new system have not been published yet, we have not been able to quantify the effects such a change would have on our analysis. But, in general, if particular aspects of poor siting lead to greater biases, then a rating system which more rigorously distinguishes between these aspects should provide more accurate estimates of the biases. We would expect the magnitude of our estimated biases to increase under such a system, since, under a less rigorous system, some stations which are relatively unbiased would have been misidentified as having a worse quality, while other stations which are relatively biased would have been misidentified as having a better quality.

Watts et al., (in preparation, 2012) argue that the new Leroy, 2010 rating system leads to an increase in the differences between subsets[17], which suggests it is an improved system. If so, it is encouraging that the new system more than doubles the percentage of stations identified as being of good quality, because it suggests that it may be possible to isolate a relatively large subset of high quality stations for estimating U.S. temperature trends, after all.

- Different sub-setting approaches

As can be seen from Figure 1, the good and bad quality subsets have relatively small sample sizes. In particular, only 1% of the stations are of Rating 1, and these stations are not evenly distributed across the U.S. (see Figure 2). It is plausible that these small sample sizes and uneven distributions could introduce statistical artefacts. In an attempt to minimise these potential artefacts, all of the studies have grouped at least some of the ratings together.

In the Menne et al., 2010 study, the small sample size problems were accentuated, because they only had ratings for 525 stations[2]. For that reason, they decided to consider only two subsets - a "good" subset comprising the 71 stations of their sample with Ratings 1 or 2 (13.5%), and a "poor" subset comprising the 454 stations with Ratings 3, 4 or 5 (86.5%). We saw in Section 3 that while the Rating 5 subset shows a large apparent warming bias (in the *Unadjusted* dataset at least), and the Rating 4 subset shows a considerable apparent warming bias, the linear trends of the Rating 3 subset are actually the lowest of all of the subsets, and that the apparent biases are "cooling". So, by grouping Ratings 3, 4 and 5 together, the Menne et al. study would have reduced the differences between their two subsets, making it harder to detect the siting biases. In addition, their "good" subset only included 71 stations, which were not evenly distributed across the country. So, their analysis might still have been affected by the problem of inadequate sample sizes.

Muller et al., 2013 argued from the fact that the Rating 3 stations had the lowest linear trends, that they were not affected by siting biases[15]. Hence, for the main part of their analysis, they constructed two subsets - an "OK" subset comprising stations with Ratings 1, 2 and 3 and a "bad" subset comprising stations with Ratings 4 and 5. However, it must be acknowledged that their preliminary analysis involved a separate histogram of linear trends for all five subsets, and they also included an additional two subsets for their Table 1 (which listed the mean linear trends of different subsets) - "good" and "poor" using the same grouping as Menne et al.

There was only two Rating 3 stations available for the Martinez et al., 2012 study[16], and a total of just 22 stations, so they created two subsets - one from the 5 stations of Ratings 1 and 2 and the other from the 13 stations of Ratings 4 and 5 (2 stations were unrated). As we will discuss in Section 4.3, and is apparent from Figure 1 and Table 1 of Martinez et al., 2012[16], there are a number of climatic differences between the 22 stations in the Martinez et al. study. So, it is probable that much of the apparent differences between the subsets may be due to confounding factors other than siting bias. Martinez et al., 2012 were conscious of this and cautioned against relying on their comparisons too much[16].

For our analysis, we wanted to avoid grouping too many stations with different ratings together, but considered the sample size of the Rating 1 stations to be too small for a statistical analysis. Hence, we grouped the Rating 1 and Rating 2 stations together (as the "good quality" subset) and considered four subsets, as described in Section 2.2. This means that we were able to distinguish between the poor quality and bad quality subsets, for instance. However, it does mean it is probable that some of the apparent differences between our subsets is related to the small sample sizes and uneven distribution of the smaller subsets, i.e., the good quality and bad quality subsets, which only had 80 and 66 stations, respectively.

Fall et al., 2011 also used our subsetting approach. But, in addition, they attempted to assess the effects of the uneven distribution of the good quality and bad quality subsets, by carrying out a second analysis testing the effects of using "proxies" for the good and bad quality stations[14]. They constructed a replacement "proxy network" for both their good and bad quality subsets by substituting each of the Rating 1 or 2 stations (for the good subset) and the Rating 5 stations (for the bad subset) with the nearest stations with a Rating of 3 or 4. They then compared these proxy subsets with the full subset of all stations with Ratings 3 or 4. They found that the 1979-2008 linear trends of the two proxy networks were different from those of the full subset. This confirms that at least some of the apparent differences between the subsets is due to uneven distribution and small sample sizes[14].

Watts et al. (in preparation, 2012) used a different rating system, which divided the stations into more evenly distributed subsets[17]. So, even though the total number of rated stations was less than for our analysis, the statistical problems of small sample size and uneven distribution should have been considerably reduced. At the time of writing, Watts et al. had not yet archived these new ratings, but it is likely that, when they are available, they would lead to more reliable estimates of the siting biases.

- Different homogenization steps of the temperature records

As we mentioned in Section 2.3, the National Climatic Data Center provide three different releases of their U.S. Historical Climatology Network dataset. The station records in two of their releases have been adjusted in an attempt to remove the presence of non-climatic biases[24]. We saw from Figures 4, 5 and 6, as well as Table 1, that most of these adjustments have the effect of increasing the long-term trends of all of the subsets.

If the adjustments are of similar magnitudes in all subsets, then this should not overly affect the estimates of the siting biases, since the differences between subsets would remain similar. However, we saw in Section 3 that, while the time-of-observation adjustments applied to each subset are quite similar, the step-change adjustments dramatically reduce the differences between the subsets. As a result, the estimates of the siting biases from the *Time-of-observation and step-change adjusted* dataset will be much smaller than those determined from the other datasets.

The step-change adjustments were designed for removing non-climatic biases[27]. So, it is plausible that they may have removed (or reduced) the magnitude of the siting biases in the records. In this case, the convergence of the station records would represent an improvement in their reliability, suggesting the lower estimates are more realistic. However, if the step-change adjustments are inadequate then this convergence would actually represent a *reduction* in their reliability. We will assess the reliability of the step-change adjustments in Section 4.5.

Menne et al., 2010 argue that the step-change adjustments could have removed many of the siting biases present in the unadjusted dataset[2]. They therefore claim that the *Time-of-observation and step-change adjusted* dataset is the most reliable, and hence their estimates of the siting biases are very low. Also assuming that the adjustments are reliable, Martinez et al., 2012 only considered the *Time-of-observation and step-change adjusted* dataset[16].

In contrast, Muller et al., 2013 only considered the *Unadjusted* dataset[15]. However, they carried out their own independent data homogenization adjustments. Like, Fall et al., 2011[14], we considered all three releases in our analysis. We assess the validity of the adjustments in Section 4.4 and 4.5.

Watts et al. (in preparation, 2012) was sceptical of the reliability of the adjustments, because they noticed that the adjustments consistently increased the 1979-2010 linear warming trends of *all* subsets, even when they further sub-divided the subsets by removing stations likely to be strongly affected by non-climatic warming biases, i.e., urbanized stations and airport stations[17]. As a result, they considered the *Unadjusted* dataset to be more reliable than the *Time-of-observation and step-change adjusted* dataset, and used the former for their main conclusions. They did not analyse the *Time-of-observation adjusted* dataset, as they considered such an analysis beyond the scope of their study[17].

Since the step-change adjustments reduce the differences between subsets, in general, studies which favour the *Time-of-observation and step-change adjusted* dataset should tend to obtain lower estimates of the siting biases. However, we do note that Muller et al., 2013 obtained a low estimate, even though they used the *Unadjusted* dataset[15], while Martinez et al., 2012 found evidence of strong siting biases, even though they used the *Time-of-observation and step-change adjusted* dataset[16].

- Use of different baseline anomaly periods.

A common approach to constructing gridded temperature trends is to convert each of the temperature records into "temperature anomalies" relative to a fixed *baseline period*, e.g., 1961-1990. This means that, for each station, the mean temperature of that station over this period was subtracted from all of its annual temperatures. This "common anomaly" approach[56] has the advantage that gridded averages are not overly affected by changes in the numbers of "cold" (e.g., high altitude) and "warm" (e.g., low latitude) stations which have data for a given year. Typically, a 30 year period is chosen, as this is generally considered long enough to reduce the noise of inter-annual temperature changes, but short enough that most records will have enough data in the period.

Unfortunately, the common anomaly approach has the disadvantage that it artificially increases the apparent agreement between stations near the anomaly period (e.g., 1961-1990). This can be easily understood by recognising that if the anomaly period is shortened to one year, then the gridded average temperature anomaly for that year will be exactly $0.0°C$, by definition. A consequence of this is that the differences between any two subsets may be artificially reduced near the anomaly period.

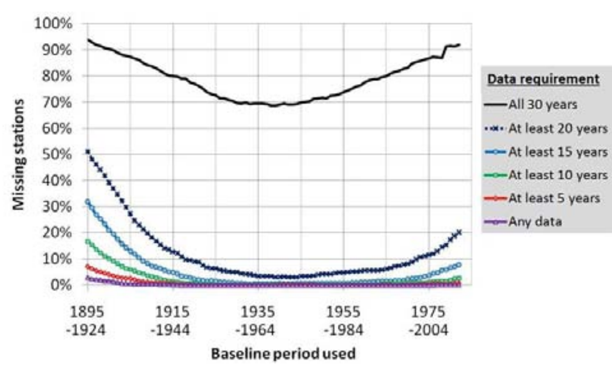

Figure 9: *The percentage of U.S. Historical Climatology Network stations which have to be dropped from analysis when different 30 year baseline periods and data requirements are used.*

To reduce the effect of this statistical artefact, we chose the earliest possible 30 year period for our analysis, i.e., 1895-1924. This means that the difference between the subsets should be low at the start of the records, and any long-term divergences between the subsets should become more pronounced towards the end of the records.

A difficulty in using an early anomaly period is that many of the shorter records will not have enough data during that period to calculate an anomaly mean. From Figure 9 it can be seen that very few stations have complete records for the entire 30 year period, regardless of the period chosen[2]. However, if we are content to allow stations to be missing some data in the anomaly period, anomaly averages can be calculated for most of the stations during a lot of different baseline periods.

In our analyses in other studies, we typically use the 1961-1990 period and a requirement of at least 15 years data[11–13]. However, for this analysis, the requirement of at least 15 years, would have meant discarding more than 30% of the stations for the 1895-1924 period (Figure 9). As a result, we reduced this restriction to a minimum of 5 years during the 1895-1924 period. This reduced the number of stations which had to be discarded to 87, or about 7% of the available stations.

Each of the other studies used different anomaly periods. Menne et al., 2010[2] and Martinez et al., 2012[16] used the 1971-2000 period. Watts et al. (in preparation)[17] and Fall et al., 2011 used 1979-2008, but Fall et al. also considered a 36 year period, 1895-1930[14]. Muller et al., 2013 used a 31 year period, 1950-1980[15].

It is plausible that these different anomaly periods could alter the subset trends. To test this possibility, we repeated our analysis using the 1961-1990 period (which is the same one we use in our other papers[11–13]). The mean gridded temperature trends of each of the subsets using the two different baseline periods are shown in Figure 10. As there are a large number of plots shown, we have just shown the 11-point binomial smoothed trends, for clarity. The annual variability and standard errors of the mean for the non-smoothed plots are similar to those shown in Figures 4, 5 and 6.

Individually, the mean trends of each of the subsets are essentially the same under both baseline periods. The only major difference is that they have been rescaled to a different baseline. This is confirmed by Table 3, where we have calculated the 1895-2011 *linear trends* for each of the equivalent subsets. Although there are some differences between the trends calculated using the different baseline periods, they are quite small. Some of these differences are probably due to fact that some of the stations (87 or $\sim 7\%$) did not have enough data during the 1895-1924 period, and so were not included in the gridded means.

1895-2011 linear trends ($°C$/century)		
	Baseline period	
Subset	1895-1924	1961-1990
Unadjusted		
Good quality	0.21	0.24
Intermediate quality	0.17	0.21
Poor quality	0.36	0.34
Bad quality	0.48	0.55
Time-of-observation adjusted		
Good quality	0.42	0.41
Intermediate quality	0.37	0.39
Poor quality	0.54	0.51
Bad quality	0.83	0.91
Time-of-observation and step-change adjusted		
Good quality	0.64	0.64
Intermediate quality	0.68	0.69
Poor quality	0.69	0.69
Bad quality	0.57	0.57

Table 3: *1895-2011 linear trends for each of the subsets using different baseline periods, calculated by linear least squares fitting.*

Instead, the main differences arise when we are *comparing* different subsets. We can see that, as expected from our discussion above, all plots converge near the baseline period. This creates the false impression of extra agreement between subsets near the baseline period compared to other periods. We can see this extra agreement is merely a statistical artefact, because when we change the baseline period from 1895-1924 to 1961-1990, the apparent period of greatest "agreement" also shifts.

A less obvious statistical artefact occurs for the *Time-of-observation and step-change adjusted* subsets. For this dataset, the National Climatic Data Center has applied a series of pairwise step-change adjustments to each of the station records in an attempt to reduce the amount of non-climatic step biases. We will discuss these adjustments further in Section 4.5, but here it is worth noting that these adjustments are carried out by adjusting the earlier portions of the records relative to the most recent portions. In other words, the annual adjustments for each record converge towards zero for the most recent

[2]The situation is better for the *Time-of-observation and step-change adjusted* dataset, since many of these data gaps have been filled in using inter-station and intra-station interpolation.

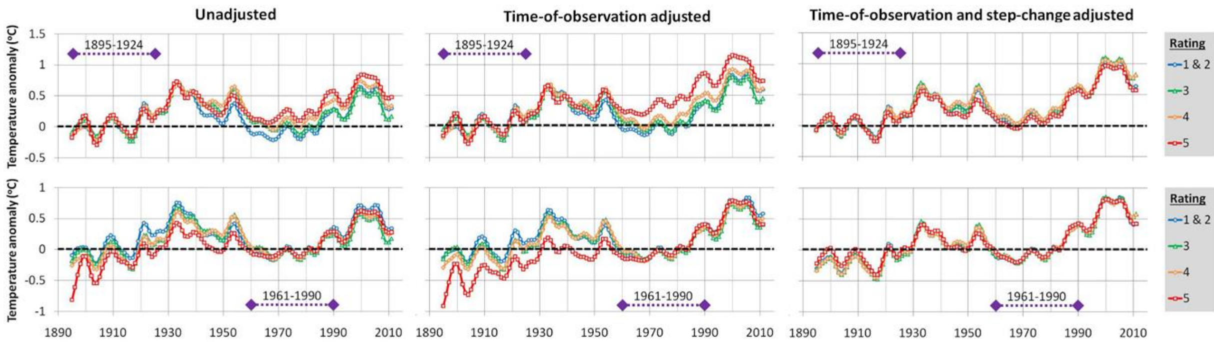

Figure 10: *Comparison of the 11 point binomial smoothed temperature trends of each of the subsets, using two different anomaly baseline periods - top panels: 1895-1924 and bottom panels: 1961-1990.*

year[24]. Since the adjustments can be of either sign, this means that applying these adjustments should lead to a partial convergence between all adjusted station records for the most recent year (in this case, 2011).

Since the biases of the poor and bad quality subsets are often estimated by considering the *differences between subsets*, it is important to recognise these artificial statistical convergences will reduce the apparent differences. In particular, the apparent biases (i.e., the differences between subsets) will be substantially reduced during whatever baseline period is chosen. In addition, for the *Time-of-observation and step-change adjusted* datasets, the biases will be artificially reduced for recent years. For this reason, it is important to estimate the extent of the poor station quality biases over the entire 1895-now period, even if it is believed that the biases were mostly introduced in recent decades[1, 2]. Of the previous analyses, only Fall et al., 2011 considered the effects of the first artificial convergence, i.e., the effect of changing baseline periods[14]. The artificial convergence introduced by the National Climatic Data Center's step-change adjustments does not appear to have been considered before now.

4.2 Over-reliance on linear trend analysis

Wunsch, 1999[57] and Percival & Rothrock, 2005[58] have illustrated how random, trend-less data series frequently produce apparent "linear trends", but that these are spurious. This is well known in the field of time series analysis, but Wunsch, 1999 correctly notes that it is frequently overlooked by researchers studying climate trends[57]. We would like to stress another point, which while also well-known in the field of regression analysis, is often overlooked by researchers considering temperature trends - namely, that, even if there are genuine trends in the data, if the trends are non-linear, calculating their "linear trends" can often be misleading.

All five of the previous studies attempting to quantify the significance of the Surface Stations results use linear trend analysis[2, 14–17], and even the preliminary qualitative analysis of Watts, 2009 uses linear trends to discuss one of their figures (Figure 23 of Watts, 2009)[1]. We also presented much of our discussion in terms of linear trend analysis. But, as we discussed in Section 3, the data we are discussing is quite non-linear in nature. So, it is important to consider the limitations of linear analysis.

There certainly is considerable convenience in calculating the linear trend for a data series. It provides a relatively simple, single value for describing an entire data series, and is included as a standard tool in most statistical data analysis packages, and even spreadsheet software. However, the problem occurs when it is used on data series that have substantial non-linearities in their trends. A "linear trend" can nominally be calculated for *any* two-dimensional series of data points, regardless of whether the series is linear or not, but if the series is non-linear then its "linear trend" might be completely meaningless. Aside from the fact that, by framing their analysis in terms of linear trends, researchers may become biased into prematurely expecting the trends to have an underlying "linear behaviour", this means that non-linear trends will be overlooked.

Some researchers take a more sophisticated approach to linear trend analysis, by calculating the significance of the linear fits, e.g., the Martinez et al., 2012 study used three separate statistical tests for assessing the significance of their linear trends[16]. However, while this is preferable to using linear trends without testing, it is still an inadequate approach, if the data series show non-linear trends. This is particularly the case, when the number of data points in the series is low (a point Martinez et al., 2012 concede[16]). For instance, if a data series contains only two points, then the "linear fit" of the points will nominally be a "perfect fit". Of course, in reality, this tells us nothing about whether the data series is genuinely linear or not.

We have intentionally plotted the trends of the twelve subsets in Figures 4, 5 and 6 without including their "linear trends", as we have found that whenever a plot is shown with a linear trend, many readers will then mentally "see" a linear trend in the data, even if the actual trends is totally non-linear, e.g., cyclical. This is perhaps an artefact of the way the human brain has evolved a high tendency towards *false pattern recognition*[57, 59]. In any case, the danger of this tendency can be illustrated by comparing the subsets in any of those figures. All of our subsets have positive linear trends when considered over the entire 1895-2011 period (see Table 1). So, if a researcher relied on those linear trends for their comparison, they might find all of the subsets to be fairly similar, in that they all "showed warming". However, it can be seen from Figure 8 that there are actually substantial differences between the subsets. For instance, for the time-of-observation adjusted datasets in Figure 5, the good quality subset appears to have alternated between periods of "warming" *and* "cooling", each lasting several decades. In contrast, the bad quality subset suggests an almost continuous "warming" trend. Both subsets have a positive linear trend for the period 1895-2011, but for the good quality subset, this is merely a consequence of the 1890s-1900s being cooler than the 1990s-2000s, i.e., if a different starting point, and period length was chosen, both the sign and the magnitude of the "linear trend" could change.

Figure 11 shows some of the different linear trends which can be calculated for the time-of-observation adjusted good exposure subset by merely varying the start year and the length of time over which the trend is calculated. Depending on the start year, periods of both "cooling" and "warming" can be found. In addition, the magnitude of the trends depends strongly on the number of years included in the calculation. For instance, the maximum and minimum trends for

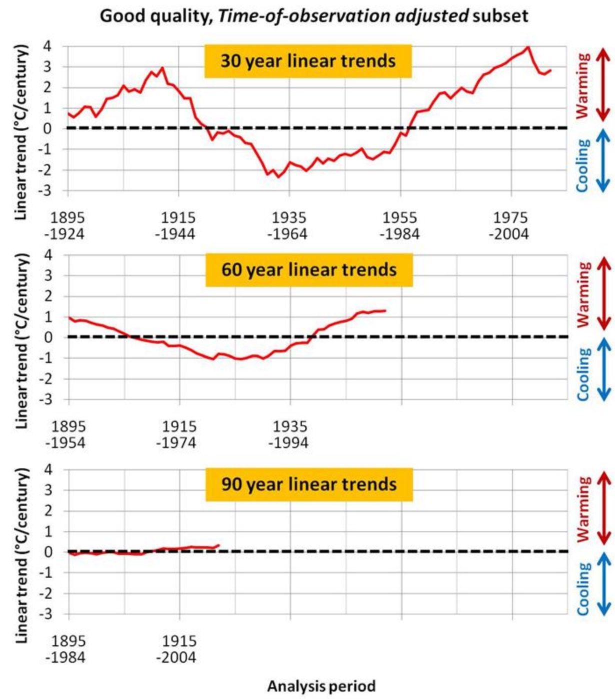

Figure 11: *Linear trends, using different starting years and period lengths, for the time-of-observation adjusted good quality subset.*

the 30 year calculations are an order of magnitude greater than the maximum and minimum trends for the 90 year calculations (+3.96 and -2.33 $^{\circ}C$/century compared to +0.33 and -0.11 $^{\circ}C$/century). If the temperature trends were truly linear then the values of the linear trends should not depend on either the start year or the number of years included.

The *difference* between the bad quality and good quality subsets is slightly better characterised by a linear fit - see Figure 12. Although the shorter time-scale linear fits show considerable variability, the mean values for each of the time-scales are of the same order of magnitude (see Table 4). In addition, aside for a few of the 30 year estimates, the trends are all of the same sign, i.e., indicating a warming of the bad quality stations relative to the good quality stations. This slightly better fit is also confirmed by the better fitting coefficients (compare the r^2 values in Tables 1 and 2).

However, even with these better fits, the 30 year estimates still show considerable variability over the entire period. The bias appears to be greatest for the 1934-1963 period. Although most periods suggest a warming bias, for three short periods, the bias ap-

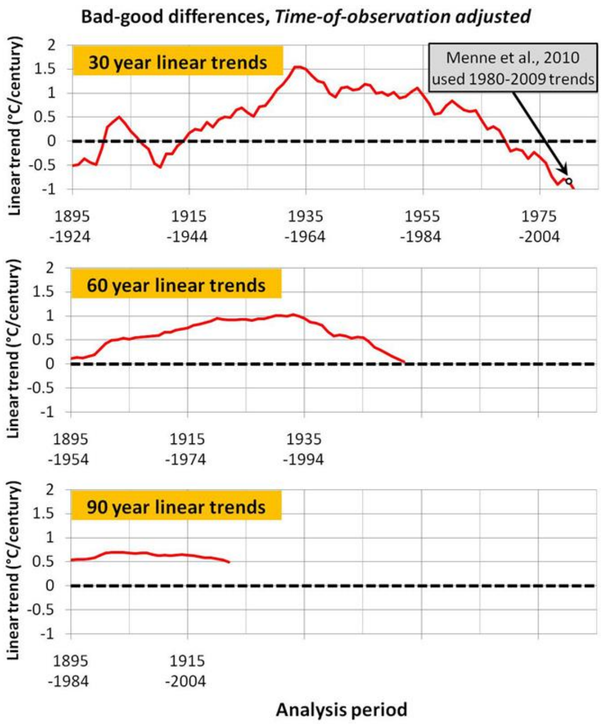

Figure 12: *Linear trends, using different starting years and period lengths, for the difference between the good and bad quality time-of-observation adjusted subsets. For comparison with Menne et al., 2010[2], the 1980-2009 linear trend is highlighted.*

pears to be a cooling one, i.e., the trends beginning during the period 1895-1900, 1907-1914 or 1970-1982 (ending in 1924-1929, 1936-1943 and 1999-2011 respectively) are negative. We note that much of the analysis by Menne et al., 2010 was based on 30 year linear trends during one of these negative trend periods - their analysis was on the 1980-2009 linear trends[2]. This might have contributed to their conclusion that poor station siting leads to a cooling bias, i.e., the opposite of our conclusions.

It can be seen that linear trends provide an overly simplistic description of non-linear time series, such as the subset temperature trends and difference trends we are considering in this study. Therefore, we do not recommend relying heavily on linear analysis in this case. Most of the previous analyses have relied very heavily on linear analysis, which limits the validity of their conclusions.

Having said that, linear trends do offer crude approximations of the overall trends, and can provide a helpful method for **crudely** comparing subsets and biases. Hence, we recognise they can be useful, *provided that the non-linear nature of the data is adequately discussed.*

	Linear trends (°C/century)		
Period of fit	Mean	Minimum	Maximum
30 years	+0.39	-1.14	+1.55
60 years	+0.63	+0.05	+1.02
90 years	+0.62	+0.49	+0.70
Whole range	+0.41	N/A	N/A

Table 4: *Mean, maximum and minimum linear trends of the difference between the bad and good subsets, for each of the time-scales used in Figure 12.*

4.3 Neglecting urbanization bias

In a series of three papers, we discuss how urbanization bias has introduced artificial warming trends into many of the station records[11–13]. Although this problem seems to be more serious for the non-U.S. component of the Global Historical Climatology Networks, it is still a significant problem for the U.S. component.

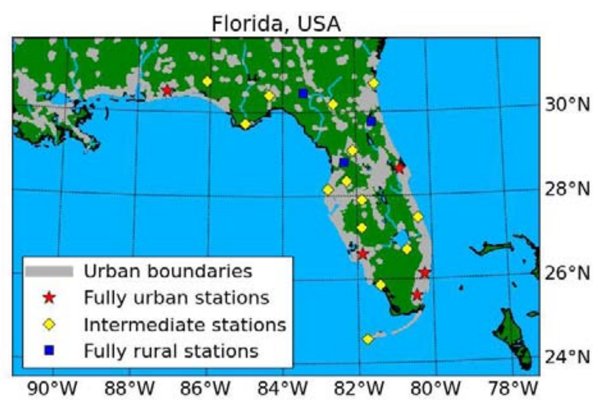

Figure 13: *Location of the Florida stations used by the Martinez et al., 2012 study[16], and the degree of urbanization of those stations.*

As the U.S. Historical Climatology Network is part of the Global Historical Climatology Network, we can use the estimates of station urbanization provided with the Global Historical Climatology Network dataset[60], to estimate the extent of urbanization in the U.S. network. The National Climatic Data Center provide two estimates of urbanization for their Global Historical Climatology Network stations - one based on the population in the general area of the station and one based on the 1994/1995 night light brightness at the station. Under each estimate, a station can be "rural", "sub-urban" or "urban". As we did in Ref. [13], we will here define a station as "fully rural" if it is identified as "rural" by both estimates, "fully urban" if it is identified as "urban" by both estimates, and otherwise "intermediate".

Using these definitions, of the 1218 U.S. Historical Climatology Network stations, 277 (22.7%) are fully rural, 842 (69.1%) are intermediate and 99 (8.1%) are fully urban. As we discuss in Ref. [13], this is relatively rural compared to the averages for the Global Historical Climatology Network. However, only 22.7% of the stations are fully rural, so it is likely that a substantial fraction of the stations are affected by at least some urbanization bias.

Since urbanization bias generally leads to an artificial warming trend, it can easily affect estimates of the separate bias due to the station siting. For instance, because the Martinez et al., 2012 study focused on a single state (Florida), only two of the stations were rated 5, and only five were rated 1 or 2. As can be seen from Figure 13, only three of the 22 Florida stations considered by Martinez et al., 2012 are identified as fully rural (blue squares). So, it is likely that much of the differences between the different stations are actually due to different amounts of urbanization bias. All three of the fully rural stations (Federal Point, Inverness 3 SE and Madison) had the same Surface Stations rating of 4, and so it would not have been possible for them to just use the fully rural stations for investigating the effects of the different ratings.

Figure 14: *Locations of good quality stations, showing urbanized regions. Urban boundaries were determined using the GRUMP dataset[61].*

The National Climatic Data Center have argued that their data homogenization should have removed much of the urbanization bias from the *Time of observation and step-change adjusted* dataset[24, 62]. However, as we discuss in Ref. [13], their step-change homogenization method is inadequate for removing urbanization bias, especially if the neighbours they use for homogenization are also affected by urbanization bias. Indeed, Pielke et al., 2007b note that such step-change homogenization algorithms provide inaccurate results when applied to stations with trend biases, such as urbanization bias[10].

For the COOP neighbours used by the National Climatic Data Center for homogenizing the Historical Climatology Network, we cannot use the Global Historical Climatology Network estimates. However, we can estimate their urbanization by assigning the station co-ordinates to the gridded urbanization estimates from the GRUMP dataset (see Ref. [61] for details on the GRUMP dataset). This approach identifies 45.0% of the COOP neighbours as being "urban" (for comparison, 55.6% of the U.S. Historical Climatology Network stations are urban with this approach). In other words, nearly half of the neighbours used for homogenizing the stations are urban. Since, this ratio is higher in urbanized areas, it is likely that the step-change homogenization approach used for generating the *Time-of-observation and step-change adjusted* dataset has led to "urban blending" into many of the rural stations which are near urbanized areas.

We can see from Figure 14 that a large fraction of the good quality stations are in urbanized areas. The sample size of the good quality stations is already quite low (8%), so there are probably not enough stations to reliably estimate U.S. regional trends by considering just the stations which are both of a good quality and rural. If the siting quality of the Global Historical Climatology Network is similar to that of the U.S. network, then the problem of separating the two confounding biases is likely to be even more challenging for the more urbanized global network.

It may be difficult to obtain a large enough sample of stations that are unaffected by either urbanization bias or inadequate station exposure, for directly assessing genuine climatic trends. Still, if the extent of urbanization bias is comparable in each subset, then it may still be possible to reliably estimate the bias introduced by inadequate station exposure, separately. This is because, we can then assume the urbanization bias will be roughly the same magnitude in each of the subsets. In other words, while the temperature trends of the subsets may be affected by urbanization bias, the *differences between the trends* should not be overly affected.

Table 5 shows the degree of urbanization of each of the subsets. There are some differences between the subsets. For example, both the good quality (Ratings 1 & 2) and the bad quality (Rating 5) stations have a relatively high percentage of fully urban stations. Similarly, there is a relatively high percentage of fully rural stations of Rating 3, although a quite low percentage for Rating 5. However, in general, the ratios seem to be *roughly* similar to the average for the entire U.S. Historical Climatology Network (i.e., the bottom row in Table 5). Therefore, it is probably still a reasonable approximation, *provided* we recognise that the estimates of the station exposure biases will be affected by the differences in urbanization bias between subsets. In some cases, these urbanization differences may be quite large, as in the Martinez et al., 2012 study.

	Degree of urbanization		
Rating	Fully rural	Intermediate	Fully urban
1 & 2	20.0%	61.3%	18.8%
3	27.2%	62.7%	10.1%
4	23.3%	69.8%	6.9%
5	8.2%	73.8%	18.0%
Unrated	21.8%	75.4%	2.8%
Average	22.7%	69.1%	8.1%

Table 5: *The percentage urbanization of each of the Surface Stations subsets.*

4.4 Are the time-of-observations adjustments reliable?

One of the main homogeneity adjustments applied to the U.S. Historical Climatology Network records is to correct for non-climatic biases introduced by documented changes in the time at which observers made their measurements at individual stations. Karl et al., 1986 calculated that if observers were estimating daily mean temperatures from a single daily reading of a minimum-maximum thermometer, then the time of day in which they made their observation would have significantly affected the mean monthly temperatures which they obtained[25]. They calculated statistical estimates for this "time-of-observation bias", which vary seasonally (from month to month), and regionally, e.g., the magnitude of the biases tends to increase with latitude because of the larger daily temperature range.

With this in mind, when the National Climatic Data Center was compiling the U.S. Historical Climatology Network, they collected the observation times associated with each station, whenever they were reported. They then applied the corresponding Karl et al., 1986 adjustments to each of the station records for two of their releases (i.e., the *Time-of-observation adjusted* and *Time-of-observation and step change adjusted* datasets).

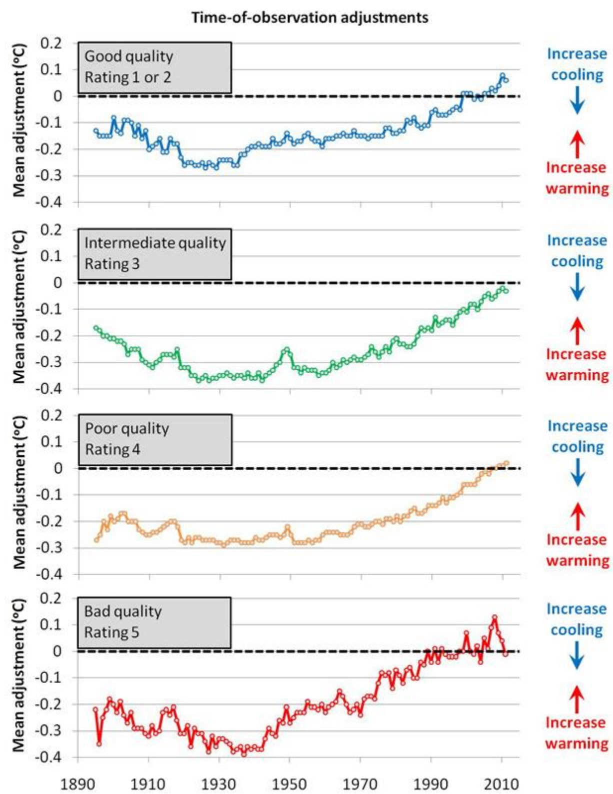

Figure 15: *Mean time-of-observation adjustments applied by the National Climatic Data Center to the gridded subsets.*

Since the U.S. station histories indicate a general reduction in the number of evening observers and increase in the number of morning observers, the time-of-observation adjustments have the net effect of introducing a post-1930s warming trend into the U.S. temperature trends[3] - see Figure 15. In general, each of the homogeneity adjustments developed by the National Climatic Data Center for the U.S. Historical

[3] Although, the pre-1930s adjustments introduce a cooling trend.

Climatology Network coincidentally seem to introduce “more warming” (e.g., see our analysis in Ref. [13]). This seems to have led to some cynicism about their reliability[1, 17, 63]. In particular, Balling & Idso, 2002 criticised the time-of-observation adjustments and argued that the unadjusted data appeared to better match satellite and radiosonde estimates of surface temperature changes[63]. In response, Vose et al., 2003[26] updated Karl et al., 1986’s study using a larger dataset. They found similar results to Karl et al., 1986, and so concluded that their original time of observation bias adjustments were reasonably accurate.

A detailed assessment of the time-of-observation bias adjustments used by the National Climatic Data Center is beyond the scope of this article. We will note that our preliminary calculations suggest that their adjustments are reasonable, provided the archived station histories for the U.S. Historical Climatology Network are accurate, and that the time-of-observation changes do not involve other station changes. For this reason, we will assume that the *Time-of-observation adjusted* dataset is more reliable than the *Unadjusted* dataset.

The adjustments are of a relatively large magnitude, however, e.g., doubling the linear trends of the good, intermediate and bad quality subsets - see Table 1. So, a more careful assessment of their reliability may be warranted. In particular, it might be worth checking whether the documented time-of-observation changes also coincided with other non-climatic changes, e.g., changes in instrumentation or station location. One way to assess the adjustments would be to compare the differences between the temperature record and the records of the station’s neighbours before and after adjustments.

With this in mind, the estimates of the biases introduced from inadequate station exposure which are determined from the *Time-of-observation adjusted* dataset should be more reliable than those determined from the *Unadjusted* dataset. If the time-of-observation adjustments are similar for all stations, on average, then these estimates should be the same, in which case it would be irrelevant which dataset was used. However, it can be seen from Figure 15 that there are slight differences in the mean adjustments applied to each of the subsets. Part of this may be a statistical artefact, due to some of the subsets having relatively small sample sizes. But, it is plausible that some of the factors influencing station quality may also be related to observer practice, in which case there may be a genuine difference between subsets.

The main difference between the subsets is that the adjustments during the 1930s are slightly greater for the bad quality subset. This increases the long-term trend of the difference between the bad quality and good quality subsets (see Figure 8), and therefore the estimate of the bias, as can be seen from Table 2. There is also a slight reduction in the estimate of the bias for the poor quality subset.

Watts et al. (in preparation, 2012) did not use the *Time-of-observation adjusted* dataset for their estimates of the biases, favouring the *Unadjusted* dataset[17]. As explained above, if the mean time-of-observation adjustments are similar for all subsets, then this should not matter. They used a different rating system which created a more evenly divided group of subsets, than the one we used. So, it is plausible that this could have reduced the differences between the adjustments applied to each subset. If this is the case, then the application of *Time-of-observation* adjustments should not substantially alter the findings of Watts et al., but this would need to be tested.

As this article is focused on the reliability of the U.S. station records, the issue of time-of-observation biases in other networks is a subject for a separate study. However, we do think it is worth mentioning that systematic changes in observation time are *not* confined to the U.S. For instance, in August 1961, stations in the People’s Republic of China changed from taking measurements at 01, 07, 13 and 19h local mean solar time to 02, 08, 14 and 20h Beijing time, i.e., local mean solar time at 120°E [64]. On 1st July 1961, the observation times at all Canadian airport weather stations were changed from 12h UTC (for maximum temperatures) and 00h UTC (for minimum temperatures) to 06h UTC[65]. In Japan, stations generally have a time-of-observation of midnight (00h JST). But, before 1940, it was 22h JST and during the period 1953-1963, minimum temperatures were observed separately at 09h JST[66].

For this reason, it is likely that time-of-observation adjustments are also necessary for other countries. In the case of the U.S., the bias seems to have led to a significant artificial cooling trend over the 20th century[26]. Similar biases, either cooling *or* warming, may also exist for many of the stations in the Global Historical Climatology Network, and they may have led to systematic biases for individual countries. For instance, in Australia, a change in the official observation time in 1964 to 0900 local time appears to have introduced an artificial warm bias in the records relative to the pre-1964 periods[67]. Also, the U.S. time-of-observation adjustments suggest a pre-1930s warming bias (see Figure 15). Until the signs and magnitudes of these biases can be reliably determined for the Global Historical Climatology Network stations, the records for those stations should be treated cautiously.

If future researchers attempt to track down observation times for stations outside of the U.S., it may also be important to check if time of observation adjustments have already been applied to the station records. Jones et al., 1986[56] suggest that time-of-observation bias corrections may have already been applied to some records, before incorporation in datasets, e.g., "*Such adjustments were used in* [the World Weather Records dataset] *for the United States up to 1940 or 1950 (depending on the station)*"[56]. The U.S. Historical Climatology Network was compiled directly from the COOP archive[25]. So, it is likely that their source records had not been corrected for time of observation. However, the Global Historical Climatology Network dataset was compiled from pre-existing datasets (including the World Weather Records), so it is possible that some of its station records have already had time-of-observation corrections, while others have not.

4.5 Are the step-change adjustments reliable?

From Figure 8, it is clear that the differences between the subsets are considerably reduced in the *Time-of-observation and step-change adjusted* dataset, compared to the *Unadjusted* and *Time-of-observation adjusted* datasets. In this sense, the step-change adjustments have led to a greater "homogeneity" between individual station records. However, this does *not* actually tell us whether the more homogeneous records are more or less climatically representative.

As we mentioned in Section 3, one explanation for the greater homogeneity is that the step-change adjustments have succeeded in substantially reducing the magnitude of the non-climatic biases in the station records. If this is the case, then Menne et al., 2010 are correct in concluding that the station quality problems identified by the Surface Stations project do not seriously affect the *Time-of-observation and step-change adjusted* estimates of U.S. temperature trends[2]. But, another explanation is that the greater homogeneity arises because the non-climatic biases have been averaged between neighbouring stations through "blending" of the biases, rather than removed.

In other words, the records with the most non-climatic biases will have some of their biases correctly removed by the process, but the records with the least non-climatic biases will have the non-climatic biases of their neighbours introduced by the process[55]. If this is the case, then Watts et al. (in preparation, 2012) are correct in concluding that the homogenization process has reduced the reliability of the good quality station records, and that the non-homogenized records are therefore more reliable, even if they have not been corrected for non-climatic biases[17]. Essentially, Watts et al. are arguing that, while the *Unadjusted* records probably contain non-climatic biases, on average, they are still more reliable than the *Time-of-observation and step-change adjusted* records.

As we discussed in Section 4.4, we are assuming the *Time-of-observation adjusted* dataset is more climatically representative than the *Unadjusted* dataset favoured by Watts et al. (in preparation, 2012). However, we also saw from Figure 15 that the net effect of these adjustments is broadly similar for all of the subsets. Indeed, the largest adjustments were for the bad quality subset, and as a result, had the effect of *increasing* the bias estimates for the bad quality subset. Hence, the main question that needs to be resolved is whether the step-change adjustments improve or reduce the reliability of the records.

With this in mind, it is worth carefully considering how the step-change adjustments are carried out. When the Surface Stations project began, the National Climatic Data Center were still using the Karl & Williams, 1987 algorithm[68] for their step-change homogenization of the U.S. Historical Climatology Network dataset. However, in 2009, they updated the dataset (see Menne et al., 2009[24]), and began using the newer Menne & Williams, 2009 algorithm[27].

There are several important differences between the two algorithms, e.g., the Karl & Williams, 1987 algorithm only tested for non-climatic biases when it was known from the station histories that a significant change had occurred, while the Menne & Williams, 2009 algorithm can test for both documented and undocumented changes. However, for the purposes of the following discussion, both algorithms are equivalent. Pielke et al., 2007b[10]'s criticism of the Karl & Williams, 1987 algorithm is similar to ours, and so also applies to the current Menne & Williams, 2009 algorithm.

Both algorithms work by comparing the temperature records of each "target" station to the records of a large number of neighbouring stations (20 for Karl & Williams, 1987; 40 for Menne & Williams, 2009). In both algorithms, the years of potential station changes are identified from the station history files. But, the Menne & Williams, 2009 algorithm also looks for additional undocumented station changes, by comparing the target to each neighbour in turn, and noting the approximate years of any apparent discrepancies between the stations. If an apparent discrepancy occurs at around the same time in several neighbour-target comparison, then it is assumed that this corresponds to a station change, or "change-point" in the target station.

It should be noted that, if similar biases occur around the same time in enough neighbours, then this could cause the algorithm to incorrectly identify the target station as having the bias. This means that if biases are relatively frequent amongst the neighbours, any target stations which are genuinely unbiased will be mistakenly treated as non-climatic "outliers", while the biased records will be mistakenly treated as the climatic trends. Stations which are heavily biased should be correctly identified as "outliers", and their biases would be reduced, however only enough to match the average of the neighbours - which would themselves be biased. This would lead to a blending (or "homogenization") of the biases amongst all the stations in the vicinity, whether biased or not.

To estimate the sign and magnitude of any non-climatic biases introduced by the proposed change-points, the target station is again compared to the neighbours, one at a time. The temperature record of each neighbour is subtracted from the target's record, to construct a "difference series". The mean temperature difference of the segment after the proposed change-point is compared to the mean temperature difference of the segment before the proposed change-point. If the difference between the two means is greater than a certain threshold value, then it is considered as a possible step-change bias. The different estimates of this proposed bias are compared for each neighbour, and if there is sufficient agreement between the estimates, then all of the years up to the year of the change-point are adjusted by that value[27, 68].

As deGaetano, 2006[69] and Pielke et al., 2007b[10] note, and as we discuss in Ref. [13], these step-change adjustments are inadequate for correcting trend biases, such as urbanization bias. The Menne & Williams, 2009 algorithm does actually consider trend biases when identifying potential change-points[27]. However, since they then treat the identified biases as step-change biases, the algorithm is still inadequate. For instance, let us suppose that the algorithm correctly identifies the start of a trend bias, and the trend bias is roughly linear in nature. In that case the *mean* value of the bias will only be half of the bias at the end of the segment, and the adjustment will be incomplete.

Pielke et al., 2007b[10] tested the effectiveness of the Karl & Williams, 1987[68] homogeneity adjustments algorithm. They simulated 1000 temperature records, and introduced artificial step change biases and trend biases into these simulated records. When they applied the Karl & Williams algorithm to these records, they found the algorithm overestimated the magnitude of positive step changes, and underestimated the magnitude of negative step changes if the station being homogenized also had a warming trend bias[10]. This confirms that the statistical "aliasing" effect described above is a problem when station records are affected by trend biases.

An additional problem of trend biases is that, if a neighbouring station suffers from a trend bias, then this will increase (or decrease) the estimates of any potential step-change adjustments.

Next, we shall consider how these algorithms operate on the U.S. Historical Climatology Network datasets. However, before doing so, it is worth considering an additional change the National Climatic Data Center adopted with their 2009 update[24]. When they switched to using the Menne & Williams, 2009 algorithm, they also switched to using the entire COOP Network for their station neighbours, rather than just using the Historical Climatology Network stations.

There are at least three serious problems with this decision:

1. Because the National Climatic Data Center do not currently provide easy access to the COOP Network dataset, their decision makes it harder for researchers to replicate and/or assess the reliability of their step-change adjustments. In our case, we were only able to carry out our analysis of the COOP neighbours because we had downloaded a 2011 version of the COOP 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/).

2. The Historical Climatology Network stations were partially chosen on the basis that they had relatively long and complete station records[70]. As a result, the average period of overlap between neighbouring Historical Climatology Network stations is much greater (78±12 years) than the overlap between Historical Climatology Network stations and their nearest COOP stations (29 ± 6 years) (see our discussion in Ref. [13]). This shorter length means that the homogenization algorithm is less likely to correctly identify the non-climatic biases in the records being homogenized. The problem is also accentuated by the Menne & Williams, 2009 algorithm, because they preferentially select neighbours which have a higher correlation with the target station[27]. This is a problem because poorly correlated station records are more likely to *appear* well-correlated, if the period of overlap between the records is short, as we illustrate in Ref. [13].

3. The Surface Stations project only investigated the U.S. Historical Climatology Network stations, and as a result, Surface Stations ratings are not available for the 90% of COOP stations which are not in the Historical Climatology Network.

On this last point, it seems reasonable to assume that the COOP stations have a similar statistical distribution of ratings to the Historical Climatology Network, i.e., something similar to Figure 1. The Historical Climatology Network was constructed from COOP stations, so it is likely that it has a similar distribution of ratings. Indeed, Fall et al., 2011 note that several station observers were unaware whether their station was part of the Historical Climatology Network or not[14]. This suggests that the differences between COOP stations and Historical Climatology Network stations were not even clear to the observers. However, this does not tell us specifically which neighbours are of good, intermediate, poor or bad quality.

Let us now consider the effects of the Menne & Williams, 2009 step change adjustments on the U.S. Historical Climatology Network stations. From our discussion throughout this article, we can make several predictions as to what these effects would be.

We saw in Section 4.3 that a substantial percentage of the U.S. stations are at least partially urbanized, and hence many of the COOP neighbours used for homogenizing the Historical Climatology Network stations would be affected by urbanization bias. For this reason, we would expect that stations with urbanized neighbours would be affected by urban blending.

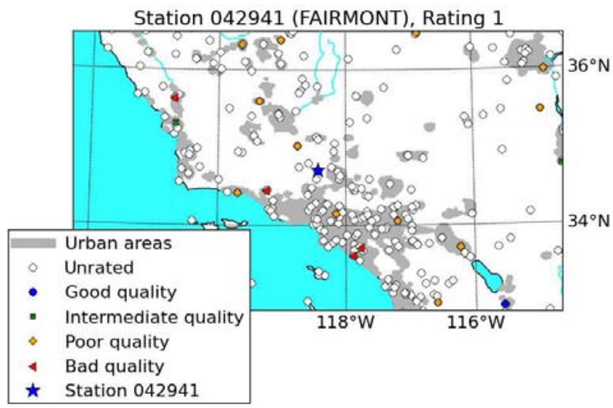

Figure 16: *Locations of the neighbouring COOP stations in the vicinity of a good quality station, Fairmont.*

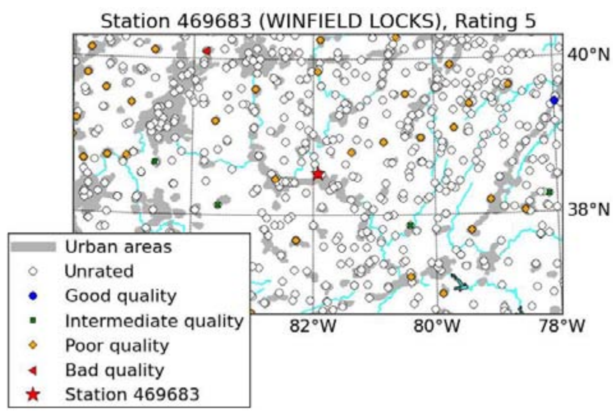

Figure 17: *Locations of the neighbouring COOP stations in the vicinity of a bad quality station, Winfield Locks.*

The urbanization (warming) bias of the most heavily affected stations would be slightly reduced, but in a similar manner, many of the rural stations with no urban bias would have an artificial warming trend introduced by urban blending. Our analysis in Ref. [13] suggests that this is the case. As mentioned earlier, the application of step-change adjustments, such as the Menne & Williams, 2009 algorithm, to trend biases substantially underestimates the biases because of statistical "aliasing" effects[10, 13, 69]. Hence, we would expect that, on average, the net effect of the adjustments would be the introduction of an artificial warming trend into all subsets that is roughly half the magnitude of the urbanization bias present in the *Time-of-observation adjusted* dataset.

Figures 16 and 17 show the COOP neighbouring stations in the vicinity of a typical rating 1 and rating 5 station, respectively. As expected, most of the neighbouring stations are unrated. But, in neither of the cases, are many of the rated neighbours of a good quality. If we assume that the COOP stations have a similar distribution of ratings to the Historical Climatology Network, then statistically, we would expect that only about 8% of the neighbours will be of a good quality. In other words, the good quality stations are likely to be considered as statistical outliers. The 6% of bad quality stations would also be considered statistical outliers, in a similar manner.

We would therefore expect that the trends of the "outlier" stations will be adjusted to better match those of the most common stations. Hence, we would expect that, while the warming trends of the bad quality stations would be partially reduced to better match those of the poor quality stations, the trends of the good quality would *incorrectly* be increased to better match those of the poor quality stations. The net effect of these adjustments would be to reduce the differences between the subsets.

As we mentioned in Section 4.1, the step-change adjustments are calculated retrospectively, i.e., when a non-climatic bias is identified, the most recent temperatures are assumed to be "correct", and the earlier temperatures are adjusted to match. A consequence of this is that the net adjustments of all subsets will converge towards zero for the most recent years, introducing an artificial convergence between the subsets.

If these predictions are accurate, then the step-change adjustments would actually reduce the reliability of the records. At any rate, they would lead to unreliable estimates of the station exposure biases. The net adjustments applied to each subset are shown in Figure 18. They are certainly consistent with all of the predictions described above. For this reason, we suggest that the bias estimates from the *Unadjusted* and *Time-of-observation adjusted* datasets are currently the most accurate. This would invalidate the conclusions of Menne et al., 2010, who assumed that the estimates from the *Time-of-observation and step-change adjusted* dataset were the most reliable[2].

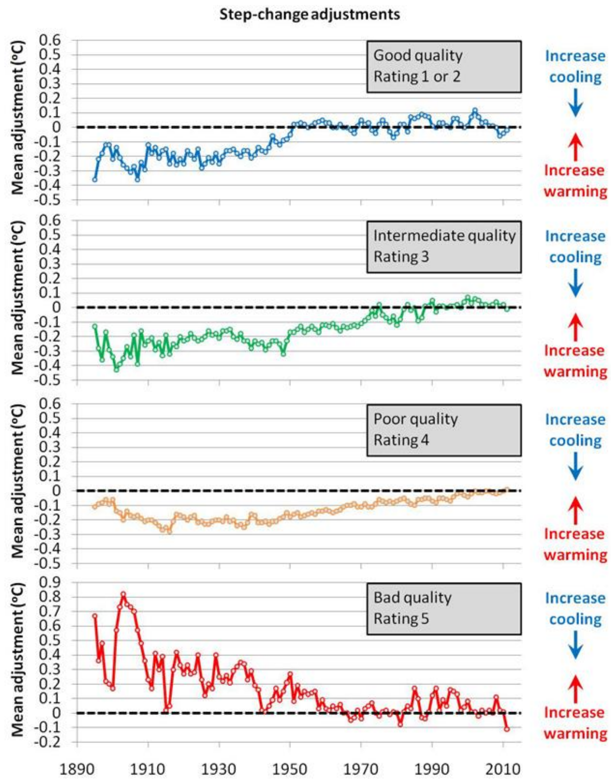

Figure 18: *Mean step-change adjustments applied by the National Climatic Data Center to the gridded subsets using the Menne & Williams, 2009[27] algorithm.*

Having said that, we must also acknowledge that there are also many other non-climatic biases in station records, e.g., station moves and changes in instrumentation. Many of these biases are probably the step-change biases which the Menne & Williams, 2009 algorithm was designed to remove[27]. Naïvely, it might be expected that, statistically, these biases should "cancel" each other out over time. However, non-climatic biases are often of the same sign and can lead to long-term apparent trends in the station records[71, 72].

Runnalls & Oke, 2006 suggest it is possible that this could have a tendency to introduce an apparent warming trend[72], in which case, a successful step-change adjustment algorithm would lead to net "cooling" adjustments. But, they point out that it could also lead to the opposite, in which case, successful step-change adjustments would lead to net "warming" adjustments. In other words, it is plausible that at least *some* of the "warming" introduced to the Historical Climatology Network by the step-change adjustments *may* be genuine. However, since it is expected for the reasons described above that the step-change adjustments will mistakenly introduce substantial artificial warming trends, it seems unwise to make any assumptions about the correct sign of the remaining non-climatic biases, yet.

5 Recommendations to improve the reliability of climate records

We saw that the majority of thermometer stations in the U.S. Historical Climatology Network currently suffer from inadequate siting, and that this has introduced artificial warming trends into estimates of U.S. temperature trends. It is quite likely that a significant fraction of stations in the global network are also affected by poor-siting problems. Indeed, from preliminary assessments[4] of the Global Historical Climatology Network stations from the rest of the world, there seem to be a mixture of poor quality and good quality stations.

With this in mind, we believe it would be useful if a global extension of the Surface Stations project were carried out so that the magnitude of the siting biases in the current estimates of global temperature trends could be determined. Since such a project would require considerable organization and international collaboration, it might be useful to carefully discuss how to assess the stations. In their initial assessments, the Surface Stations team used Leroy, 1999[22]'s rating system, but Watts et al. (in preparation, 2012) suggest that the rating system proposed by Leroy, 2010[31] is more appropriate. Decisions on whether to use one of these two systems, or another should be agreed at the start of the project. It would be desirable to choose a rating system that would allow comparable estimates of past station exposures to be at least approximated from the site documentation which is sometimes included by station histories[10].

According to Watts et al. (in preparation, 2012)[17], the Surface Stations team did not collect enough information to adequately rate the shading and ground cover associated with stations and their ratings were instead predominantly based on heat source/sink proximity. Changes in shading and/or ground cover, e.g., from growth/removal of neighbouring trees, can substantially alter micro-climate[3, 4]. So, future investigations should also investigate these factors.

It would probably be sensible if the team involved in the original Surface Stations project were consulted before this type of project is begun on a global scale, as they already have several years of valuable experience in carrying out station quality assessments.

Such a global project would probably have been too overwhelming a task during the 1990s, when the Global Historical Climatology Network was being first compiled. But, with modern improvements in the globalization of communication, data archiving and data sharing such a project is probably now feasible.

It would also be desirable if repeat station surveys could be carried out every few years, as Leroy, 2010 recommends[31]:

> *The rating of each site should be reviewed periodically as environmental circumstances can change over a period of time. A systematic yearly visual check is recommended: if some aspects of the environment have changed, a new classification process is necessary.*
>
> *A complete update of the site classes should be done at least every 5 years.*

A major difficulty in resolving the issue of how station exposure have biased long term temperature trends is that the Surface Stations project only provides us with information about the *current* station quality. Repeat surveys every few years, as suggested above, would reduce this issue, going forward. But, if we want to continue to use station data for earlier periods, we should probably attempt to collect more information on the quality of these stations for earlier times.

Pielke et al., 2007b[10] note that for the early and middle part of the 20th century, COOP observers were encouraged to provide hand-drawn schematics illustrating the site exposure associated with their stations. Although these schematics were generally not as comprehensive as the descriptions compiled by the Surface Stations project, and station moves were not always documented[10], it may be possible that partial estimates of the historical station exposures can be reconstructed from these schemat-

[4] See the "surfacestation" posts on the Tallbloke's Talkshop blog, the following posts on Roger Pielke Sr.'s blog: 2011/08/11, 2011/08/16, 2011/08/17, 2011/08/23, 2011/08/29, 2011/09/08, 2011/09/16,2011/09/27 and 2011/09/28, or these Watts Up With That posts on the Sydney, Australia station: 2008/07/02 and 2013/01/14.

ics. In the 1980s, these schematics were replaced with more cryptic, and less informative, shorthand notation so that they could be entered into simple computer forms. Nonetheless, when combined with the earlier schematics and photographic evidence of the current site exposure, it may be possible to reconstruct reasonable approximations of the site exposures during that period.

We saw in Section 4.5 that the step-change homogenization algorithm currently used by the National Climatic Data Center is unable to correctly deal with the siting biases in the U.S. Historical Climatology Network. In Ref. [13], we also showed that this algorithm breaks down when a large number of stations are affected by urbanization bias (or any non-climatic *trend* biases). All automated statistical homogenization algorithms are by their nature imperfect, since they result in a mixture of true positives, true negatives, false positives and false negatives. However, while the Menne & Williams, 2009 algorithm appears to perform relatively well on synthetic temperature records[24, 73], it appears to be seriously problematic when dealing with the extensive non-climatic biases present in the U.S. Historical Climatology Network, and presumably also the Global Historical Climatology Network.

We appreciate that it is important to remove these non-climatic biases, as well as others, e.g., station moves, time-of-observation biases or changes in instrumentation. A considerable amount of work has already been carried out in assessing these step-change homogenization techniques, e.g., see Refs. [69, 73, 74]. However, the current step-change homogenization approaches still seem to introduce at least as many problems as they remove. For this reason, it seems unwise to apply these approaches to the temperature records in an automated manner, as is currently done.

We suggest that, if researchers want to continue to use these automated homogenization approaches, they should only be used for *identifying* potential step changes (and possibly estimating their magnitudes), and flagging them for the researcher to manually check, on a case-by-case basis. Records of the researcher's justifications for accepting/rejecting individual adjustments should be kept, and provided to users of the data, who want to make their own assessments. This would require more work on the part of the researcher, e.g., a typical station record may contain about 6 such flags per 100 years[27], and regional station networks may contain hundreds or thousands of station records. But, from our discussion in Section 4.5, it seems manual checks are still necessary.

Rather than focusing on developing automated statistical methods for identifying undocumented non-climatic biases, maybe we should first attempt to collect (and digitize, where necessary) as much information on individual station histories as possible. When we have a more complete knowledge of the potential non-climatic biases for which documentation is available, and have developed a more accurate knowledge of the impact of the documented biases, we may be in a better position to assess how to best treat the undocumented biases.

While station histories were collected during the compilation of the U.S. Historical Climatology Network, the same was *not* done for the Global Historical Climatology Network. When Peterson et al., 1997 were compiling the Global Historical Climatology Network[60], they were mainly combining pre-compiled temperature (and precipitation) datasets. These datasets often just contained mean monthly temperature (and/or precipitation) values, and so Peterson et al. decided not to consider detailed station histories and metadata.

Although it is true that standards differ between meteorological organisations, and the level of documentation may vary from station to station, it is likely that there is a considerable amount of useful station documentation out there, which could be useful. We suggest that, if the station observers and/or meteorological organisations associated with given stations are contacted directly, relatively detailed station histories may be available. If a global equivalent of the Surface Stations project were to be organised the collection of these details could be carried out simultaneously.

Better station histories and documentation could allow researchers to dramatically increase the reliability of the global climate network. For instance, if time-of-observation biases are as serious a problem for U.S. stations as is claimed[25, 26], then it is likely that similar problems exist for other networks in the Global Historical Climatology Network. Station histories could provide information on the observation times and averaging methods used at different stations, which could help researchers assess these problems.

As we discuss in Refs. [11–13], many of the long station records in the Global Historical Climatology Network contain a large number of unexplained data gaps. These also are frequently associated with large temperature changes. The station observers, or station histories may be able to explain those gaps, and also whether any temperature changes between the gaps are likely to be climatic or not. In some cases, it may be possible to update the temperature records, if extra data is available. We note that it may be useful to also collect daily measurements too, if possible, as daily measurements often provide more insight than monthly measurements[75].

In terms of how such a global project could be organized, we note that, in the U.S. alone, approximately \$2.5 billion/year is currently allocated for *"federal research on global change and climate change"* through the United States Global Change Research Program. In such terms, a relatively small outlay of funding into systematic methodical on-site assessments and updates of station records and station histories would drastically improve the reliability of our climate records. Even in the absence of direct funding, the success of the Surface Stations project as well as other climate-related volunteer projects, such as the Old Weather and Climate Prediction projects, suggests that there could be sufficient interest in *"citizen science"* research[76] for a global extension of the original Surface Stations study of U.S. stations. If this approach was taken, it would be useful if the project received explicit approval and cooperation from the various international meteorological agencies whose stations would be considered. The backing of the World Meteorological Organization would probably help such a project.

Finally, we note that government-funded, high quality, carefully located and sited, modern climate networks like the National Climatic Data Center's U.S. Climate Reference Network (USCRN), if properly managed, have the potential to offer future researchers climate records which would be unaffected by many of the main non-climatic biases plaguing current studies, e.g., urbanization bias, siting bias, time-of-observation bias and possibly instrumental bias[77, 78]. The U.S. Climate Reference Network already has a decade of data[77], and if similar networks were setup and maintained in enough other countries, then it may possible in the future to compile enough stations for global climate studies. We recognise that this would not be of immediate benefit to current researchers, but are reminded of the proverb: *"The best time to plant a tree is twenty years ago. The second best time is now."*

6 Conclusions

Recent surveys of the current thermometer station exposure of the weather stations in the U.S. Historical Climatology Network have identified that about 70% of stations are poorly or badly sited[1, 2]. As a result, it is likely that many of the thermometer records currently used for estimating climate trends (both regional U.S. temperature trends and global temperature trends) have been biased by changes in the localised micro-climate in the immediate vicinity of the thermometer stations, rather than representing purely climatic trends. However, attempts to quantify these siting biases have until now been somewhat contradictory and controversial[1, 2, 14–17].

In this study, we estimate that siting biases have artificially increased the mean temperature trends of the *Unadjusted* station records by about 32%, with the subset of the worst-sited stations being increased by about 56%. When time-of-observation adjustments were applied to the station records, this led to a mean increase in temperature trends of about 39%, as has been previously noted[26, 63]. For this reason, although the siting biases remained similar in magnitude, the *relative* fraction of the temperature trends in the *Time-of-observation adjusted* station records that was due to siting biases was reduced - an artificial warming of about 18% for all rated stations, and about 49% for the worst-sited stations. This still represents a substantial bias.

When the *Time-of-observation adjusted* station records were also homogenized with step-change adjustments using the Menne & Williams, 2009 algorithm[27], the differences between all subsets of stations were substantially reduced. There was no longer much difference in temperature trends between the good quality stations and the bad quality stations. Previously it had been argued that this was because the step-change adjustments had removed most of the siting biases in the station records[2]. However, our analysis shows that the Menne & Williams, 2009 adjustments actually lead to considerable blending/mixing of siting biases between stations, rather than their removal. As a result, the current *Time-of-observation and step-change adjusted* version of the U.S. Historical Climatology Network is unreliable, and its use should be discontinued.

It seems likely that similar siting biases also exist in the Global Historical Climatology Network, which has been frequently used for estimating global temperature trends. This suggests that current estimates of the amount of "global warming" since the 19th century (e.g., Ref. [23]) have been significantly overestimated.

Siting biases are not the only non-climatic biases affecting the thermometer stations being used for studying climatic trends. For instance, in Refs. [11–13] we discuss the widespread problem of urbanization bias in current thermometer records. We agree that thermometer station records could be an invaluable resource for evaluating global and regional temperature trends. However, many of the problems identified by earlier researchers, such as Mitchell, 1953[79] have still not been adequately resolved. In Section 5, we offer a number of recommendations which we believe would lead to more reliable climate records.

Acknowledgements

No funding was received for this research.

We are grateful to Anthony Watts and the other volunteers who worked on the Surface Stations project for generating and archiving the station exposure ratings for the U.S. Historical Climatology Network.

We would like to thank Don Zieman for some useful comments.

References

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

[2] M. J. Menne, C. N. Jr. Williams, and M. A. Palecki. "On the reliability of the U.S. surface temperature record". *J. Geophys. Res.* 115 (2010), p. D11108. DOI: 10.1029/2009JD013094.

[3] D. A. Robinson. "Gathering climatic data of the highest quality" (1997), pp. 305–308.

[4] 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/.

[5] H. Yilmaz et al. "Determination of temperature differences between asphalt concrete, soil and grass surfaces of the city of Erzurum, Turkey". *Atmósfera* 21 (2008), pp. 135–146. URL: http://www.scielo.org.mx/scielo.php?pid=S0187-62362008000200002&script=sci_arttext.

[6] 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.

[7] G. Manley. "Central England temperatures: monthly means 1659-1973". *Quart. J. R. Met. Soc.* 100 (1974), pp. 389–405. DOI: 10.1256/smsqj.42510.

[8] 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.

[9] R. A. Sr. Pielke et al. "Unresolved issues with the assessment of multidecadal global land-surface temperature trends". *J. Geophys. Res.* 112 (2007), D24S08. DOI: 10.1029/2006JD008229.

[10] 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.

[11] 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.

[12] 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.

[13] R. Connolly and M. Connolly. "Urbanization bias III. Estimating the extent of bias in the Historical Climatology Network datasets". 34 (Clim. Sci.). Ver. 0.1 (non peer reviewed draft). 2014. URL: http://oprj.net/articles/climate-science/34.

[14] S. Fall et al. "Analysis of the impacts of station exposure on the U.S. Historical Climatology Network temperatures and temperature trends". *J. Geophys. Res.* 116 (2011), p. D14120. DOI: 10.1029/2010JD015146.

[15] R. A. Muller et al. "Earth atmospheric land surface temperature and station quality in the contiguous United States". *Geoinfor. Geostat.* 1 (2013). DOI: 10.4172/2327-4581.1000107.

[16] C. J. Martinez, J. J. Maleski, and M. F. Miller. "Trends in precipitation and temperature in Florida, USA". *J. Hydrol.* 452-453 (2012), pp. 259–281. DOI: 10.1016/j.jhydrol.2012.05.066.

[17] 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/.

[18] 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.

[19] R. S. Vose et al. "Comments on 'Microclimate exposures of surface-based weather stations: Implication for the assessment of long-term temperature trends'". *Bull. Amer. Meteor. Soc.* 86 (2005), pp. 504–506. DOI: 10.1175/BAMS-86-4-504.

[20] T. C. Peterson. "Examination of potential biases in air temperature caused by poor station locations". *Bull. Amer. Meteor. Soc.* 87 (2006), pp. 1073–1089. DOI: 10.1175/BAMS-87-8-1073.

[21] NOAA/NESDIS. "Climate Reference Network (CRN) Site Information Handbook". NOAA/NESDIS CRN Ser. X030, CRN Rep. NOAA-CRN/OSD-2002- 0002R0UD0, National Climatic Data Center, Asheville, N.C., U.S.A. 2002. URL: http://www1.ncdc.noaa.gov/pub/data/uscrn/documentation/program/X030FullDocumentD0.pdf.

[22] M. Leroy. "Classification d'un site". Note Technique #35. Direction des Systmes dObservation, Météo-France, Trappes, France. 1999. URL: http://www.ccrom.org/ccrom/IMG/pdf/note_technique35-2.pdf.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] T. R. Karl et al. “A new perspective on recent global warming: Asymmetric trends of daily maximum and minimum temperature”. *Bull. Amer. Meteor. Soc.* 74 (1993), pp. 1007–1023. DOI: 10.1175/1520-0477(1993)074<1007:ANPORG>2.0.CO;2.

[29] 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.

[30] R. S. Vose, D. R. Easterling, and B. Gleason. “Maximum and minimum temperature trends for the globe: An update through 2004”. *Geophys. Res. Lett.* 32 (2005), p. L23822. DOI: 10.1029/2005GL024379.

[31] M. Leroy. “Siting classification for surface observing stations on land”. JMA/WMO Workshop on Quality Management in Surface, Climate and Upper-air Observations in RA II (Asia). Tokyo, Japan 27-30 July 2010. 2010. URL: http://www.jma.go.jp/jma/en/Activities/qmws_2010/CountryReport/CS202_Leroy.pdf.

[32] M. Chenoweth. “A possible discontinuity in the U.S. historical temperature record”. *J. Clim.* 5 (1992), pp. 1172–1179. DOI: 10.1175/1520-0442(1992)005<1172:APDITU>2.0.CO;2.

[33] M. Chenoweth. “Nonstandard thermometer exposures at U.S. Cooperative weather stations during the late nineteenth century”. *J. Clim.* 6 (1993), pp. 1787–1797. DOI: 10.1175/1520-0442(1993)006<1787:NTEAUC>2.0.CO;2.

[34] D. E. Parker. “Effects of changing exposure of thermometers at land stations”. *Int. J. Climatol.* 14 (1994), pp. 1–31. DOI: 10.1002/joc.3370140102.

[35] W. S. Hughes. “Comment on D. E. Parker, ’Effects of changing exposure of thermometers at land stations’”. *Int. J. Climatol.* 15 (1995), pp. 231–234. DOI: 10.1002/joc.3370150208.

[36] D. E. Parker. “Effects of changing exposure of thermometers at land stations: A reply”. *Int. J. Climatol.* 15 (1995), p. 235. DOI: 10.1002/joc.3370150209.

[37] N. Nicholls et al. “Historical thermometer exposures in Australia”. *Int. J. Climatol.* 16 (1996), pp. 705–710. DOI: 10.1002/(SICI)1097-0088(199606)16:6<705::AID-JOC30>3.0.CO;2-S.

[38] W. S. Hughes. “Comment on ’Historical thermometer exposures in Australia’ by N. Nicholls et al.” *Int. J. Climatol.* 17 (1997), pp. 197–199. DOI: 10.1002/(SICI)1097-0088(199702)17:2<197::AID-JOC113>3.0.CO;2-R.

[39] B. Trewin. “Another look at Australia’s record high temperature”. *Aust. Met. Mag.* 46 (1997), pp. 251–256. URL: http://www.bom.gov.au/amm/docs/1997/trewin.pdf.

[40] P. Ø. Nordli et al. “The effect of radiation screens on Nordic time series of mean temperature”. *Int. J. Climatol.* 17 (1997), pp. 1667–1681. DOI: 10.1002/(SICI)1097-0088(199712)17:15<1667::AID-JOC221>3.0.CO;2-D.

[41] A. Moberg et al. “Were southern Swedish summer temperatures before 1860 as warm as measured?” *Int. J. Climatol.* 23 (2003), pp. 1495–1521. DOI: 10.1002/joc.945.

[42] D. Frank et al. “Warmer early instrumental measurements versus colder reconstructed temperatures: shooting at a moving target”. *Quat. Sci. Rev.* 26 (2007), pp. 3298–3310. DOI: 10.1016/j.quascirev.2007.08.002.

[43] R. Böhm et al. “The early instrumental warm-bias: a solution for long central European temperature series 1760-2007”. *Clim. Change* 101 (2010), pp. 41–67. DOI: 10.1007/s10584-009-9649-4.

[44] M. Brunet, J. Asin, J. Sigró, et al. “The minimization of the screen bias from ancient Western Mediterranean air temperature records: an exploratory statistical analysis”. *Int. J. Climatol.* 31 (2011), pp. 1879–1895. DOI: 10.1002/joc.2192.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] R. Wilson et al. “A matter of divergence: Tracking recent warming at hemispheric scales using tree ring data”. *J. Geophys. Res.* 112 (2007), p. D17103. DOI: 10.1029/2006JD008318.

[50] R. D’Arrigo et al. “On the “divergence problem” in northern forests: A review of the tree-ring evidence and possible causes”. *Glob. Planet. Change* 60 (2008), pp. 289–305. DOI: 10.1016/j.gloplacha.2007.03.004.

[51] 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/.

[52] 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.

[53] 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/.

[54] R. L. Baumhardt. “Dust Bowl Era”. In: *Encyclopedia of Water Science*. Marcel Dekker, Inc. New York., 2003, pp. 187–191.

[55] A. Watts. “Rumours of my death have been greatly exaggerated” (2010). URL: http://wattsupwiththat.com/2010/01/27/rumours-of-my-death-have-been-greatly-exaggerated/.

[56] P. D. Jones et al. “Northern Hemisphere surface air temperature variations: 1851-1984”. *J. Clim. Appl. Meteor.* 25 (1986), pp. 161–179. DOI: 10.1175/1520-0450(1986)025<0161:NHSATV>2.0.CO;2.

[57] C. Wunsch. “The interpretation of short climate records, with comments on the North Atlantic and Southern Oscillations”. *Bull. Amer. Meteor. Soc.* 80 (1999), pp. 245–255. DOI: 10.1175/1520-0477(1999)080<0245:TIOSCR>2.0.CO;2.

[58] D. B. Percival and D. A. Rothrock. ““Eyeballing” trends in climate time series: A cautionary note”. *J. Clim.* 18 (2005), pp. 886–891. DOI: 10.1175/JCLI-3300.1.

[59] M. Shermer. “Turn me on, dead man”. *Sci. Am.* 292 (May) (2005), p. 37. URL: http://www.scientificamerican.com/article.cfm?id=turn-me-on-dead-man.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] D. A. Portman. “Identifying and correcting urban bias in regional time series: Surface temperature in China’s northern plains”. *J. Clim.* 6 (1993), pp. 2298–2308. DOI: 10.1175/1520-0442(1993)006<2298:IACUBI>2.0.CO;2.

[65] 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.

[66] F. Fujibe et al. “Long-term changes of temperature extremes and day-to-day variability in Japan”. *Pap. Met. Geophys.* 58 (2007), pp. 63–72. DOI: 10.2467/mripapers.58.63.

[67] B. Trewin. “A daily homogenized temperature data set for Australia”. *Int. J. Climatol.* In press (2012). DOI: 10.1002/joc.3530.

[68] 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.

[69] 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.

[70] 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.

[71] J. R. Lanzante. “Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data”. *Int. J. Clim.* 16 (1996), pp. 1197–1226. DOI: 10.1002/(SICI)1097-0088(199611)16:11<1197::AID-JOC89>3.0.CO;2-L.

[72] 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.

[73] 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.

[74] 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.

[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] C. Franzoni and H. Sauermann. “Crowd science: The organization of scientific research in open collaborative projects”. *SSRN* (November 14, 2012). DOI: 10.2139/ssrn.2167538. URL: http://ssrn.com/abstract=2167538.

[77] H. J. Diamond et al. “U.S. Climate Reference Network after one decade on operations: Status and assessment”. *Bull. Amer. Meteor. Soc.* in press (2012). DOI: 10.1175/BAMS-D-12-00170.

[78] K. P. Gallo. “Evaluation of temperature differences for paired stations of the U. S. Climate Reference Network”. *J. Clim.* 18 (2005), pp. 1629–1636. DOI: 10.1175/JCLI3358.1.

[79] 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.