Evidence of urban blending in homogenized temperature records in Japan and in the United States: implications for the reliability of global land surface air temperature data

Genki Katata,[a,b] Ronan Connolly,[c,d] and Peter O'Neill[e].

[a] *Ibaraki University, Ibaraki 310-8512, Japan*

[b] *The Canon Institute for Global Studies (CIGS), Tokyo 100-6511, Japan*

[c] *Center for Environmental Research and Earth Science (CERES), Salem, MA 01970, USA*

[d] *Independent scientist, Dublin, Ireland*

[e] *School of Mechanical and Materials Engineering (retired), University College Dublin, Ireland*

Corresponding author: Genki Katata, katata.genki@canon-igs.org

File generated with AMS Word template 2.0

Early Online Release: This preliminary version has been accepted for publication in *Journal of Applied Meteorology and Climatology*, may be fully cited, and has been assigned DOI 10.1175/JAMC-D-22-0122.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT

In order to reduce the amount of non-climatic biases of air temperature in each weather station's record by comparing it to neighboring stations, global land surface air temperature datasets are routinely adjusted using statistical homogenization to minimize such biases. However, homogenization can unintentionally introduce new non-climatic biases due to an often-overlooked statistical problem known as "urban blending" or "aliasing of trend biases". This issue arises when the homogenization process inadvertently mixes urbanization biases of neighboring stations into the adjustments applied to each station record. As a result, urbanization biases of the original unhomogenized temperature records are spread throughout the homogenized data. To evaluate the extent of this phenomenon, the homogenized temperature data for two countries (Japan and United States) are analyzed. Using the Japanese stations in the widely used Global Historical Climatology Network (GHCN) dataset, it is first confirmed that the unhomogenized Japanese temperature data are strongly affected by urbanization bias (possibly ~60% of the long-term warming). The United States Historical Climatology Network dataset (USHCN) contains a relatively large amount of long, rural station records and therefore is less affected by urbanization bias. Nonetheless, even for this relatively rural dataset, urbanization bias could account for ~20% of the long-term warming. It is then shown that urban blending is a major problem for the homogenized data for both countries. The IPCC's low estimate of urbanization bias in the global temperature data based on homogenized temperature records may have been biased low due to urban blending. Recommendations on how future homogenization efforts could be modified to reduce urban blending are discussed.

SIGNIFICANCE STATEMENT

Most weather station-based global land temperature datasets currently used a process called "statistical homogenization" to reduce the amount of non-climatic biases. However, using temperature data from two countries (Japan and United States), we show that the homogenization process unintentionally introduces new non-climatic biases into the data due to "urban blending" problem. Urban blending arises when the homogenization process inadvertently mixes the urbanization (warming) bias of the neighboring stations into the adjustments applied to each station record. As a result, the urbanization biases of the unhomogenized temperature records are spread throughout all of the homogenized data. The

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

net effect tends to artificially add warming to rural stations and subtract warming from urban stations until all stations have about the same amount of urbanization bias.

1. Introduction

Regional and global land surface air temperature (LSAT) trends are routinely calculated using thermometer records from weather stations (Lawrimore et al. 2011; Menne et al. 2018). However, weather station records are often affected by non-climatic biases (Karl and Williams 1987; Mitchell 1953; Pielke et al. 2007; Soon et al. 2015; Soon et al. 2018) due to, e.g., station moves (Karl and Williams 1987; Ren et al. 2015; Soon et al. 2015), changes in instrumentation (Quayle et al. 1991; Hubbard and Lin 2006) or thermometer screen (Nordli et al. 1997), changes in time of observation (Karl et al. 1986; Balling and Idso 2002; Vose et al. 2003), changes in the immediate surroundings of the weather station, i.e., "micro-climate" (Fall et al. 2011; Menne et al. 2010), and changes in the local climate that are unrepresentative of regional trends such as the growth of urban heat islands (Fukui 1957; Karl et al. 1988; Oke 1973; Stewart 2019).

The last type of non-climatic bias is often referred to as "urbanization bias" and has been particularly challenging for at least two reasons: 1) many thermometer records have experienced at least some urbanization over the last century, especially the longest records and 2) it is typically a warming bias—unlike most non-climatic biases that can be of either sign. Hence, many studies over the years have warned that at least some of the apparent long-term warming in both global and/or regional temperature estimates could be an artefact of urbanization bias (Connolly et al. 2021; Fujibe 2009; Fujibe and Ishihara 2010; Fujibe 2011, 2012; Fukui 1957; Karl et al. 1988; Oke 1973; Ren et al. 2015; Ren and Ren 2011; Scafetta 2021; Shi et al. 2019; Soon et al. 2015, 2018; Zhang et al. 2021). However, other studies have disputed this claim and argued urbanization bias is a relatively minor issue (Das et al. 2011; Efthymiadis and Jones 2010; Hansen et al. 2001, 2010; Hausfather et al. 2013; Parker 2006; Peterson et al. 1999; Wickham et al. 2013).

The Intergovernmental Panel on Climate Change (IPCC) chose the latter side of this dispute for its most recent 6th Assessment Report (AR6) and argued that it is "unlikely" that urbanization bias accounts for more than 10% of global LSAT trends (IPCC 2021), although conceding the problem might be larger for some regions, e.g., eastern China (Shi et al. 2019).

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

However, several studies disagree with this optimistic assessment (Connolly et al. 2021; Scafetta 2021; Soon et al. 2015; Soon et al. 2018; Zhang et al. 2021).

Evaluating the contribution of urban warming to a temperature record is a challenging problem – especially given the non-uniform nature in which urbanization takes place over timescales of decades to centuries. Some researchers have emphasized the point that for older urban areas (e.g., European cities) much of the growth in the magnitude of its urban heat island might have occurred earlier (Jones et al. 2008) than more modern cities (e.g., some Southeast Asian cities). Also, the urban heat island of stations located in metropolitan park areas might be reduced by the "park cool island" effect (Jones et al. 2008), although this effect is relatively modest in highly urbanized areas (Gaffin et al. 2008). Stewart (2011) has highlighted many of these challenges and problems, while Stewart and Oke (2012) recommend researchers use more nuanced "local climate zones" for evaluating urban warming trends.

At any rate, to try and reduce the problems of non-climatic biases, several groups have developed statistical "homogenization" techniques to identify and correct for non-climatic biases in the temperature records by comparing each station record to those of neighboring stations (e.g., Domonkos 2021; Easterling and Peterson, 1995; Karl and Williams 1987; Menne and Williams 2009; Mestre et al. 2013). Nowadays, most LSAT datasets used for climate research include a version that has been homogenized using one of these statistical homogenization programs. In the case of NOAA's Global Historical Climatology Network (GHCN) datasets (Lawrimore et al. 2011; Menne et al. 2018), that we will be analyzing in this study, Menne and Williams (2009)'s "Pairwise Homogenization Algorithm" (PHA) is used.

Clearly, if the homogenization process successfully removed most of the non-climatic biases, while retaining the true climatic trends, then these homogenized records would be more suitable for studying climatic trends than the unhomogenized records. As a result, most LSAT estimates explicitly and exclusively rely on homogenized records (Lenssen et al. 2019; Menne et al. 2018; Osborn et al. 2021; Sun et al. 2022; Vose et al. 2021). However, while initially this might seem reasonable, it is important to remember homogenization techniques are merely statistical attempts *to try* to improve the reliability of the data.

The hope is that a given homogenization technique correctly identifies and removes any non-climatic biases; does not overlook any remaining biases; and does not inadvertently add

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

additional biases. However, given that the motivation for homogenizing the data is that there are a large number of unidentified and unquantified biases in the original data, evaluating the reliability of a homogenization technique by merely comparing the homogenized and unhomogenized data is somewhat similar to "pulling oneself up by one's boot straps". Therefore, most assessments of the reliability of a given homogenization technique have used synthetic temperature records designed to mimic properties of actual temperature records where artificial biases have been deliberately added, (e.g., DeGaetano 2006; Domonkos 2011, 2021; Menne and Williams 2009; Pielke et al. 2007; Reeves et al. 2007; Squintu et al. 2020; Venema et al. 2012; Williams et al. 2012). Since both the "true" records and the "biases" are known in advance, the successes and failures of the homogenization technique—at correcting these synthetic records—can be directly quantified.

In terms of these synthetic benchmarking tests, most of the homogenization algorithms in use today appear to improve the quality of the data, i.e., to remove more biases than they add (Domonkos 2011, 2021; Squintu et al. 2020; Venema et al. 2012; Williams et al. 2012). This has led to a widespread belief among the homogenization community that homogenized thermometer records are automatically more reliable than the original data (Lenssen et al. 2019; Menne et al. 2018; Osborn et al. 2021; Sun et al. 2022; Vose et al. 2021). However, this belief is not necessarily correct. Recently, O'Neill et al. (2022) compared the various homogenization adjustments applied to more than 800 European station records in the GHCN datasets to the corresponding station history metadata associated with the stations. Only about 20% of the adjustments applied corresponded to documented non-climatic events.

Also, the adjustments applied to each station often changed dramatically each time the GHCN dataset was updated and re-homogenized, i.e., roughly once per day (O'Neill et al. 2022). About 80% of the breakpoint adjustments were inconsistently applied. Since the gridded averages for regional areas are typically based on multiple station records, these inconsistencies for most station records did not necessarily alter the estimated regional temperature trends. However, the inconsistencies suggested that many of the adjustments were effectively arbitrary in nature. Such serious flaws occurred despite the fact that the GHCN homogenization techniques performed relatively well in synthetic benchmarking tests (Menne and Williams 2009; Venema et al. 2012; Williams et al. 2012). This suggests synthetic benchmarking assessments might be insufficient for evaluating their real-world performance.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

An obvious limitation of synthetic benchmarking is that the evaluation depends on how representative the synthetic time series and biases are compared to real thermometer records. Indeed, DeGaetano (2006) cautioned that while many techniques were very effective at identifying non-climatic step change breakpoints when the underlying reference time series were trendless, problems due to statistical aliasing arose when some of the stations had long-term trend biases. In most of these cases, when the homogenization algorithm identified a breakpoint, about half of the trend bias was falsely included as part of the homogenization adjustment. Pielke et al. (2007) emphasized that this aliasing problem could potentially be adding artificial trends to the homogenized temperature records that were non-climatic in nature.

Menne and Williams (2009) conceded that aliasing was also an artefact of the PHA method, but seem to have concluded it was only a minor issue, and not necessarily a problem since it "would bring the adjusted target more in agreement with the background trend captured by the neighbors" (Menne and Williams 2009). Hausfather et al. (2013) acknowledged aliasing could potentially cause problems if urbanization biases were substantial, and their analysis of the United States Historical Climatology Network (USHCN; a subset of version 3 of the GHCN dataset) revealed *some* aliasing was occurring when urban neighbors were used. However, they suggested this dataset was sufficiently rural for the problem to be minor.

Connolly and Connolly also considered aliasing in a series of working papers in 2014 (Connolly and Connolly 2014a-d). In those papers, it was noted that the aliasing effect appeared to be leading to substantial "urban blending" in the homogenized records, whereby some of the urbanization bias of reference stations was aliased to the most rural records during the homogenization process (Connolly and Connolly 2014d). The flipside of this is that some of the urbanization biases of the most urban stations are also reduced via aliasing. However, the net effect was that the trends of all records converge to the average of all stations (rural and urban). These working papers noted that many attempts to accurately quantify the extent of urbanization bias in homogenized datasets by comparing rural and urban trends (Li et al. 2004; Peterson et al. 1999) appeared to have overlooked this possibility (Connolly and Connolly 2014b-d). They noted that the aliasing problem is also a concern for other non-climatic biases that introduce long-term biases into a large fraction of a reference network, e.g., siting biases (Connolly and Connolly 2014a), and furthermore, that

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

assessments of the extent of siting biases *in homogenized datasets* that compared the trends of well-sited and poorly-sited stations (Fall et al. 2011; Menne et al. 2010) similarly appeared to have overlooked this problem. At any rate, Connolly and Connolly warned that the unfortunate combination of the urban blending problem coupled with the widespread nature of the urbanization bias problem had misled the climate community into a false confidence that homogenization had largely eradicated the urbanization bias problem (Connolly and Connolly 2014a-d).

Indeed, by using manual empirical homogenizations based on known station metadata rather than statistical homogenization and by identifying four regions that collectively comprised most of the rural stations with relatively long and complete records, Soon et al. (2015), and more recently Connolly et al. (2021), generated rural northern hemisphere LSAT estimates that were markedly different from the standard estimates based on both urban and rural homogenized records.

As part of their analysis of Chinese temperature trends, Soon et al. (2018) described theoretically the statistical reasons for the urban blending problem. They also demonstrated the problem empirically using the results from an analysis of 9 stations in the Beijing, China area (He and Jia 2012). He and Jia (2012)'s results showed a strong correlation between the rate of urbanization around a station and the magnitude of the 1978-2008 linear warming trend in the unhomogenized data. After homogenization, the correlation was substantially reduced because the trends of the most rural stations were increased while those of the most urban stations were reduced. That is, all trends converged towards the *average of the station network.*

Soon et al. (2018)'s conclusions on the significance of the urban blending problem led to some debate. Li and Yang (2019) suggested that developers of these statistical homogenization techniques might have already considered this problem and overcome it somehow. In their reply, Soon et al. (2019) stressed that it is important to distinguish between two separate stages of the homogenization process:

1) Identifying when a non-climatic step change bias occurred;

2) Establishing (and adjusting for) the sign and magnitude of the bias.

In the first stage, many of the current homogenization procedures perform quite well when tested with simulated and/or synthetic biases as described above. These evaluations of

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

the first stage appear to have been what convinced Li and Yang (2019) that the homogenization process was "correct and reasonable". However, the urban blending problem arises during the second stage of the process (Soon et al. 2019).

Although Soon et al. (2018) had demonstrated the urban blending effect both theoretically and using a sample of nine stations in the Beijing, China area, this was a very small sample size that only covered a relatively small geographical area. Hence, Soon et al. (2019) explicitly called on "further research to investigate its extent".

We have attempted here to evaluate the extent of urban blending for two different countries—Japan (a heavily urbanized country) and United States (a country with a mixture of urban and rural stations).

For Japan, we base our analysis on the Japanese component of the GHCN datasets version 3 (Lawrimore et al. 2011) and version 4 (Menne et al. 2018). Japan is a highly urbanized country with a high density of weather stations. Many of the Japanese stations are affected by urbanization bias, particularly the longest and most complete records (Fujibe 2009; Fujibe and Ishihara 2010; Fujibe 2011, 2012; Fukui 1957; Matsumoto et al. 2017; Stewart 2019). In particular, the Tokyo metropolitan area has such a geographically large urban heat island that it spans multiple cities (Matsumoto et al. 2017; Yamashita 1996).

For the United States, we base our analysis on the USHCN dataset (Menne et al. 2009). The USHCN was a very high quality subset of the GHCN dataset up until version 4 that was also homogenized using PHA by NOAA, but independently from the rest of the GHCN (Lawrimore et al. 2011) and is still updated and maintained by NOAA. As mentioned above, a previous study (Hausfather et al. 2013) also looked at the aliasing problem for this dataset but reached different conclusions from us. Therefore, we will also reanalyze the Hausfather et al. (2013) data to investigate possible reasons for the apparently different conclusions.

2. Illustration of the mechanics of the blending problem

In order to understand how blending/aliasing occurs through current temperature statistical homogenization techniques (including PHA), let us consider a highly idealized thought experiment as illustrated in Fig. 1. Let us imagine a world in which no global warming or cooling was occurring. Instead, annual temperatures vary randomly from the local temperature within the range ±0.1°C. In Fig. 1, we consider six stations located in a similar part of this hypothetical world where the average regional temperature is 8.0°C.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

Station 1 remains rural. Therefore, the temperature variability shown in Fig. 1b over the arbitrarily chosen period of 1970-2000 for this station oscillates between 8.0±0.1°C for each year. In our case, the random time series we generated nominally has a linear trend of –0.01°C/century over this period, but by design this is a purely random time series.

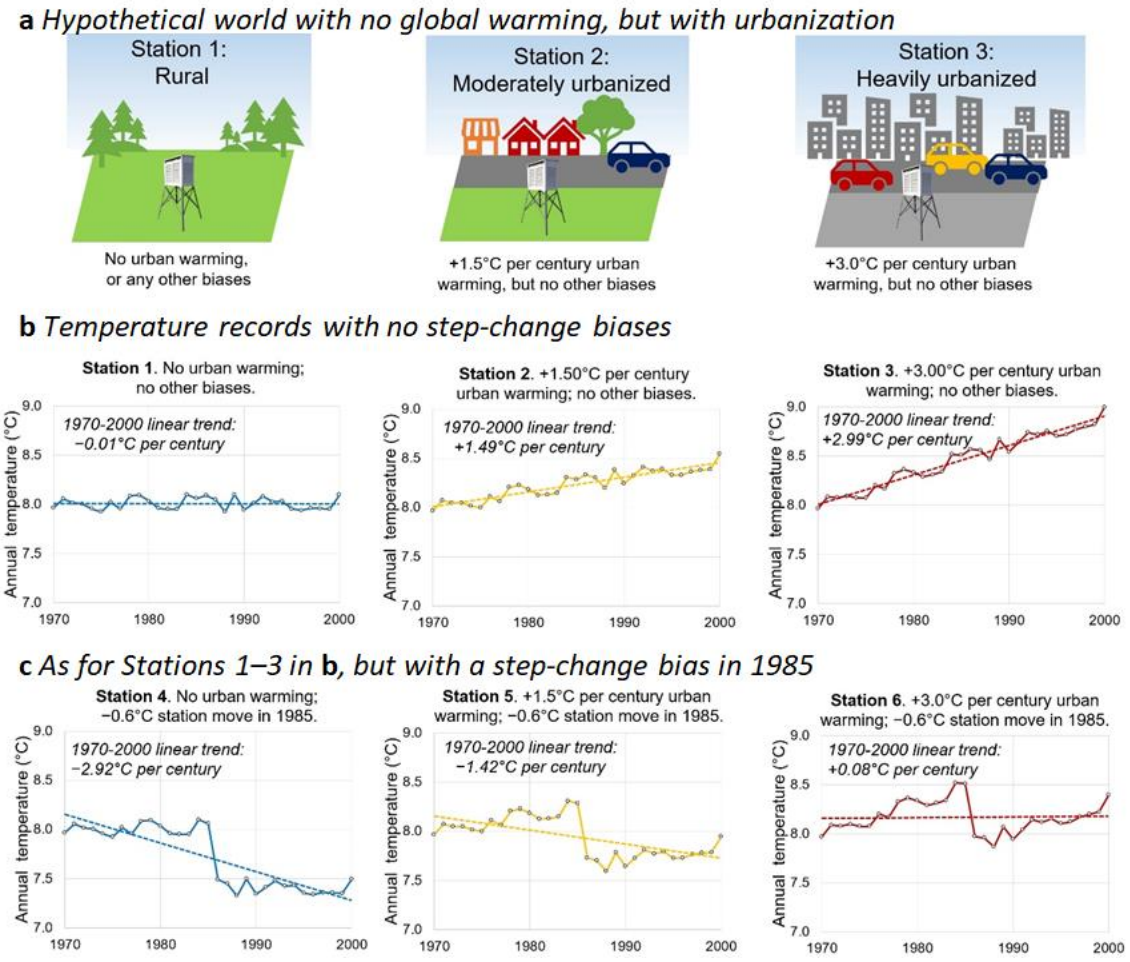


Figure 1 Synthetic temperature records (1970-2000) for six hypothetical temperature records use in our thought experiment to consider how the urban blending problem arises from current temperature homogenization techniques such as Menne and Williams (2009). Panel (a) envisages a hypothetical world where there has been no long-term global temperature trend and annual temperatures at a rural location (*Station 1*) have varied randomly between 8.0±0.1°C over the 1970-2000 period. However, two neighboring locations have become steadily more urbanized over this period and thereby each have experienced a different urban warming trend superimposed on the rural temperature variability. One location (*Station 2*) has become moderately urbanized with a linear urban warming trend of +1.5°C per century over the 1970-2000 period. The other location (*Station 3*) has become heavily urbanized with double the urban warming trend at +3.0°C per century over the period. Panel (b) plots the temperature records that would have been recorded for each station assuming that no other non-climatic biases occurred and a continuous record was kept for the entire period. Finally, Panel (c) hypotheses three equivalent stations (*Stations 4-6*) that experienced the exact same temperature variability as *Stations 1-3*, but each also

File generated with AMS Word template 2.0

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

experienced a station move to a cooler location (e.g., a higher elevation) in 1985 that introduced a one-off step-change cooling bias of −0.6°C.

For *Stations 2 and 3*, the underlying temperature variability is exactly the same. However, each station also experienced a continual "urban warming" over the 1970-2000 period that we have approximated as a linear ramp of +1.5°C per century and +3.0°C per century respectively. In our hypothetical world, all three thermometer stations were unaffected by any other non-climatic biases (unlikely in the real world) and their annual temperature records are plotted in Fig. 1b.

Meanwhile, let us suppose that each station also had an equivalent neighboring station (*Stations 4-6*) that was identical except also experiencing a station move in 1985 that introduced a one-off cooling bias of exactly –0.6°C. No other non-climatic biases occurred (unlikely in the real world) over this period. Their annual temperature records are plotted in Fig. 1c.

Now let us consider in turn what would happen if we used either *Station 1, 2 or 3* to homogenize the temperature records of the other three stations. For simplicity, let us suppose that the homogenization process only uses one neighboring station. In reality, current homogenization approaches either use the average of multiple neighbors as a "reference series" (Easterling and Peterson 1995) or apply an iterative process whereby each station is compared to multiple neighbors one-at-a-time and then the values from all of these pairwise comparisons are averaged together, e.g., PHA (Menne and Williams 2009).

Let us suppose that the homogenization technique identifies the presence and timing of any step-change biases with 100% accuracy. It therefore accurately identifies that *Stations 4-6* each experienced a non-climatic bias in 1985. How would the technique calculate the sign and magnitude of this bias? Current homogenization techniques typically do this through a statistical evaluation of the temperature difference series. Typically, the magnitude of the bias is estimated as being the average of the difference series for a period *after* the breakpoint minus the average of the difference series for a period *before* the breakpoint. In the case of the PHA, the length of these periods corresponds to the "homogeneous" series between the breakpoint and the next breakpoint in the difference series, but it must be a minimum of 24 months (2 years) (Menne & Williams, 2009). In cases where the incidence of breakpoints is

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

low, this may be 10−20 years or longer. For simplicity, let us suppose that it is exactly 10 years every time.

Figure 2 plots the resulting difference series using each of *Stations 1-3* as the neighbor. Fig. 3 plots the results of the homogenized *Station 4-6* records, depending on whether *Station 1, 2 or 3* was used as the neighbor, i.e., Fig. 3a, b or c, respectively.

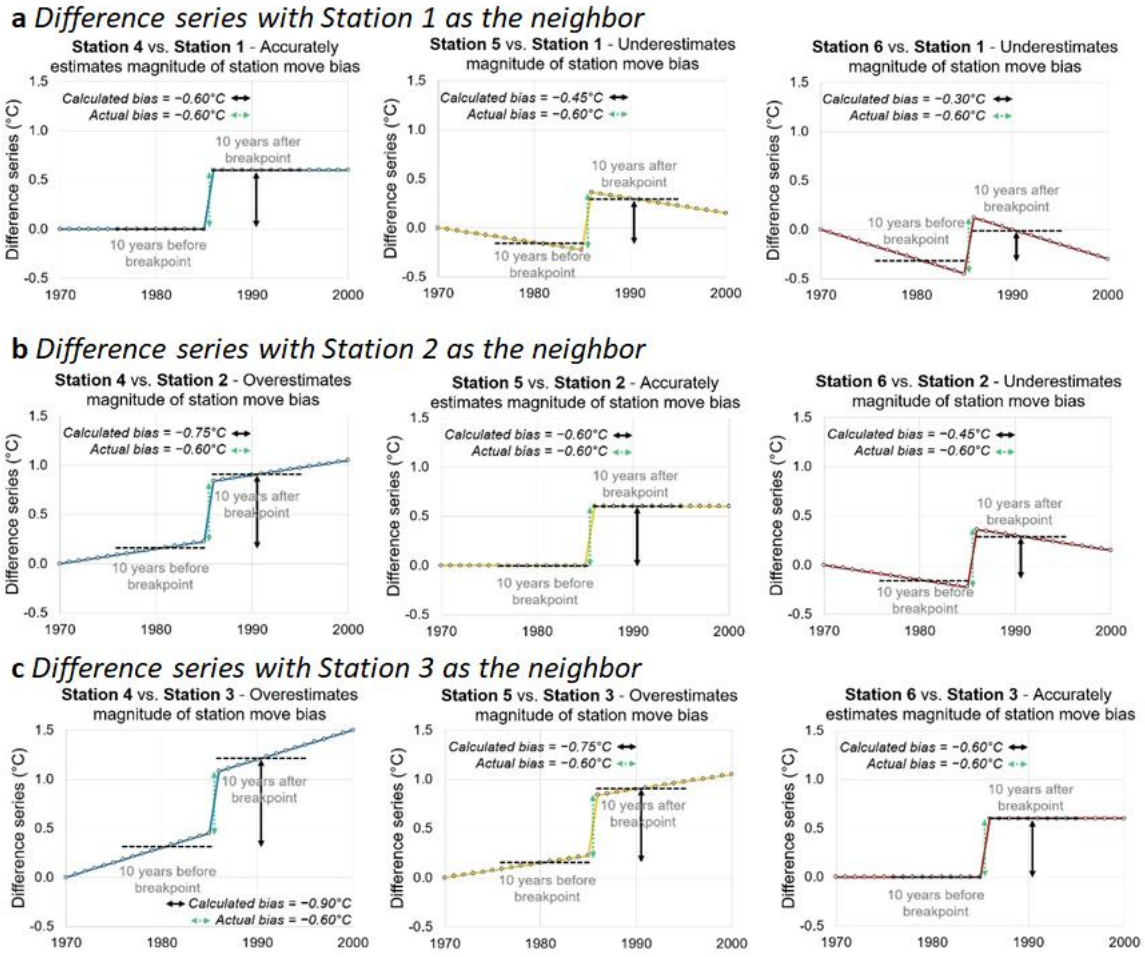


Figure 2 Difference series that would be used for estimating the timing and magnitude of any breakpoints in the temperature records for our hypothetical *Stations 4-6* from Fig. 1 when the neighbor records used are (a) *Station 1*, i.e., rural, (b) *Station 2*, i.e., moderately urbanized, and (c) *Station 3*, i.e., heavily urbanized. As explained in Fig. 1, each of *Stations 4-6* experienced a station move in 1985 that introduced a one-off step-change cooling bias of −0.6°C. However, even if the homogenization algorithm correctly identifies that a bias occurred in 1985, the calculated magnitude of the bias depends on whether the neighbor has experienced more or less urban warming than the target station. That is, the magnitude of cooling step change biases (as in this case) will be overestimated by more urbanized neighbors and underestimated by less urbanized neighbors. For warming step change biases, the opposite would occur.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

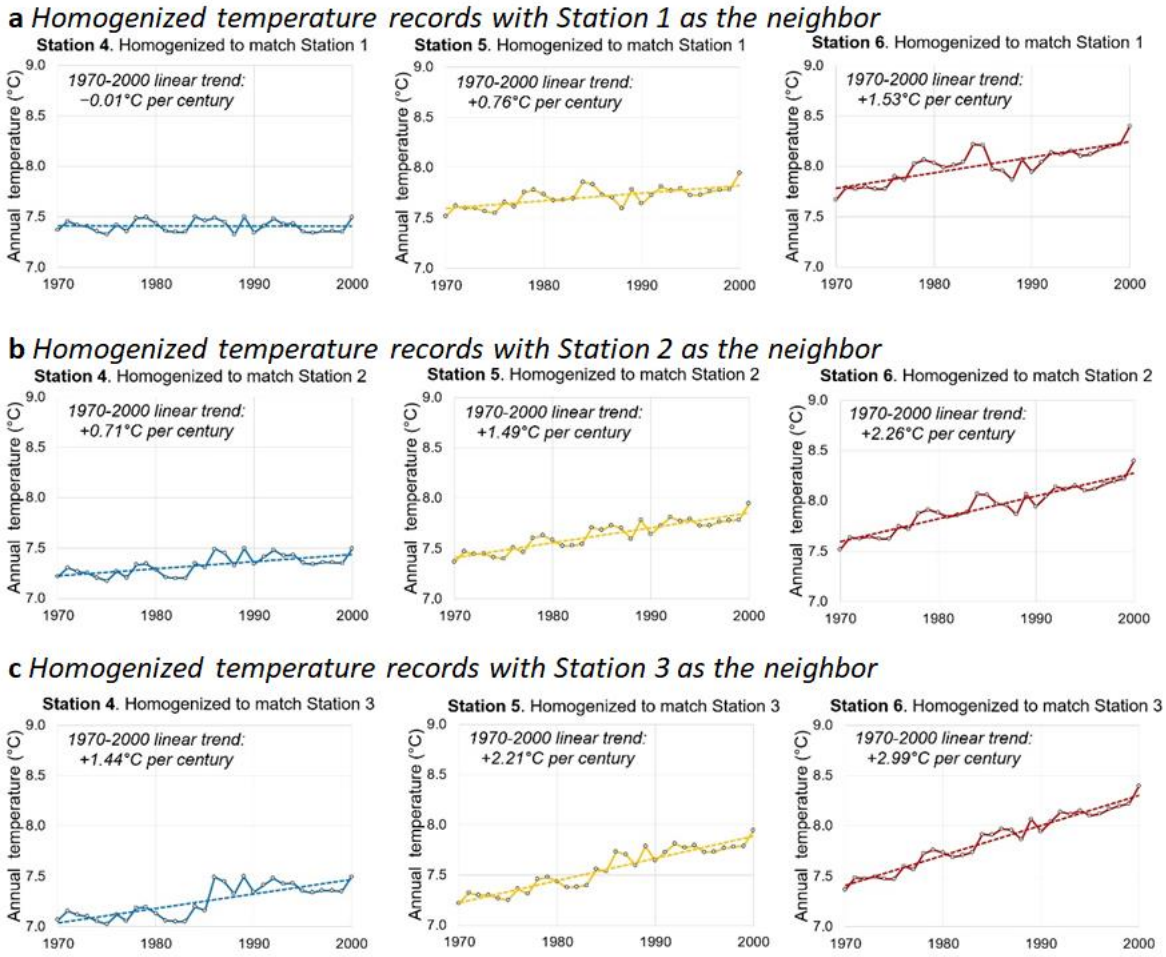


Figure 3 Illustration of how the three homogenized temperature records for the hypothetical *Stations 4-6* of Fig. 1 would vary substantially depending on how much urban warming the neighbors used for estimating the magnitude of the identified biases experienced. That is, the homogenization adjustment applied to each breakpoint varies depending on whether the target and neighbor stations have experienced different degrees of urban warming. (a)-(c) represent the different homogenized *Stations 4-6* records that would result if the neighbor records used are (a) *Station 1*, i.e., rural, (b) *Station 2*, i.e., moderately urbanized, and (c) *Station 3*, i.e., heavily urbanized. Note that if the target and neighbor records have experienced a similar urban warming over the period (e.g., *Stations 1 and 3*; *Stations 2 and 5*; or *Stations 3 and 6*), then the bias from the station move is correctly removed and the underlying urban warming remains. However, if they experienced different degrees of urban warming, then some of this difference in urban warming becomes "aliased" into the homogenized record.

Although the above thought experiment is obviously highly idealized, it allows us to see in an idealized manner how and why aliasing/blending occurs. As can be seen by comparing the results of Fig. 3 to the corresponding unhomogenized temperature records of Fig. 1c, in all cases, the homogenization process has succeeded in making the temperature record look

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

much "smoother" and more "continuous" than the unhomogenized record. However, if the neighbor experienced *more or less* urban warming than the target station, the process also involved blending. The estimated magnitude of the bias was either underestimated or overestimated depending on whether the neighbor was less or more affected by urban warming.

Specifically, in our cases, if the neighbor was more urbanized than the target station, then the process overestimated the magnitude of the cooling bias in 1985 (Fig. 2). Hence, the homogenization process added some of the extra urban warming of the neighbor into the homogenization adjustment (Fig. 3). Contrariwise, if the neighbor was more rural than the target station, the magnitude of the cooling bias was underestimated. Hence, the homogenization process failed to remove all of the actual cooling bias.

If our hypothetical station move had instead led to a warming bias, the phenomenon would be the other way around. Then, more urbanized neighbors would underestimate the magnitude of the warming bias, whereas more rural neighbors would overestimate its magnitude. This is the problem that was identified by some as "aliasing" (deGaetano 2006; Pielke et al. 2007) and others as "urban blending" (Connolly and Connolly 2014c; Soon et al. 2015, 2018, 2019; Connolly et al. 2021).

The net effect is a tendency for the temperature trends of homogenized stations to converge towards those of the neighbors. Generally, this tendency will increase the more breakpoints are identified by the homogenization process. If the neighbor network comprises a mixture of stations with varying degrees of urbanization bias, the homogenized stations will converge towards the average degree of urbanization bias of the network. The most rural stations will tend to have "urban warming" added by homogenization while the most urban stations will tend to have some (but not all) of their "urban warming" removed (Soon et al. 2018).

Counterintuitively, when homogenization "successfully" removes any non-climatic step changes without aliasing, the homogenized records should ideally retain their *trend biases* (including urban warming). This is because the goal of the current homogenization techniques is to remove non-climatic *step* biases, rather than *trend* biases. Therefore, if the *step* biases are all accurately removed by a homogenization, then the homogenization process should ideally leave the *trend* biases unaffected. Trend biases can (and should) then be accounted for in a later, separate step, as suggested by Soon et al. (2018).

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

That said, we caution that, if multiple consecutive step biases of the same sign occurred over a period of time, these might initially be mistaken for a "trend bias". For instance, Menne et al. (2009) consider an example (Reno, Nevada, USA) where the unhomogenized record suggests an apparent urban warming "trend bias" beginning in the 1970s. However, they argued that in that particular case, this apparent "trend bias" was a result of "major step changes during the […] 1990s caused by station relocations" (Menne et al. 2009). Apparently, in this case, the multiple station relocations coincidentally introduced biases of the same sign.

How then can we minimize the aliasing/blending problem? We propose here two potential workarounds and consider their pros and cons:

(1) Arguably the simplest way to avoid the blending problem is to bypass the use of reference records for this second stage of the homogenization process and evaluate the value of the bias internally using statistical properties of the target record. For example, by comparing the mean temperature for a given period (perhaps 1-2 years) before and after the identified breakpoint.

(2) Another approach is to ensure only neighbors that experienced a similar degree of trend biases are used for estimating the value of the non-climatic bias. For example, if the target station is rural, then the reference neighbors should also be similarly rural, while if the target station experienced urban warming, then similarly urbanized neighbors should be used. This is shown as *Station 4* of Fig. 3a, *Station 5* of Fig. 3b, and *Station 6* of Fig. 3c in our thought experiment.

A major advantage of the first approach is that it completely avoids the blending problem since no neighbors are used in the second stage (although they could have been used in the first stage, i.e., identification of breakpoints). However, a potential disadvantage is that if the station experienced any genuine warming or cooling over the comparison period this would be lost in the process. This could be particularly problematic for station records that have many step biases.

A disadvantage of the second approach is that suitable neighbors with similar trend biases need to be identified for each target station before homogenization. This could significantly reduce the pool of potential reference series for some stations, e.g., rural stations surrounded by urban stations or vice versa—although the larger pool might be used for the first stage of identifying the breakpoints. However, it offers the advantage that the homogenization should

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

retain any regional warming or cooling that coincided with the timing of the various breakpoints.

3. Data sources and methodology

a. Analysis of Japanese temperature records

For our analysis of the Japanese temperature records, we downloaded the widely used GHCN monthly datasets from https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-monthly [last accessed on May, 2022]. In 2018, the dataset was updated from version 3 (Lawrimore et al. 2011) to version 4 (Menne et al. 2018) involving a major overhauling of the datasets, but version 3 was kept updated until late-2019. Therefore, we analyze both versions 3 and 4. Both versions have unhomogenized (henceforth, "raw") and homogenized (henceforth, "adjusted") datasets—the former has only quality control corrections applied, and the latter has been homogenized using Menne & Williams (2009)'s automated PHA techniques. For simplicity, we only study annual trends here. Therefore, we used the annual mean temperature available for 12 complete months for a given year. To analyze the temperature variations for given stations, we adopt the popular approach of converting each temperature record into an "anomaly time series" relative to a constant baseline period of 1961–1990. Following Connolly et al. (2021), we require a station to have a minimum of at least 15 complete years of data during this baseline period to be incorporated into our analysis. Statistical calculations (linear regression, correlation, and the student *t*-test) were performed using MS-Excel.

Figure 4a and b shows the location of Japanese stations extracted from the GHCN datasets. This comprises 167 stations for version 3 and 191 for version 4. Figure 4c and d compare the gridded mean LSAT time series for Japan using all available stations (regardless of urbanization) with either the raw or adjusted datasets. We note that the homogenization process has slightly increased the linear temperature trends for both GHCN versions. Also plotted are the total station numbers over time in Fig. 4e and f.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

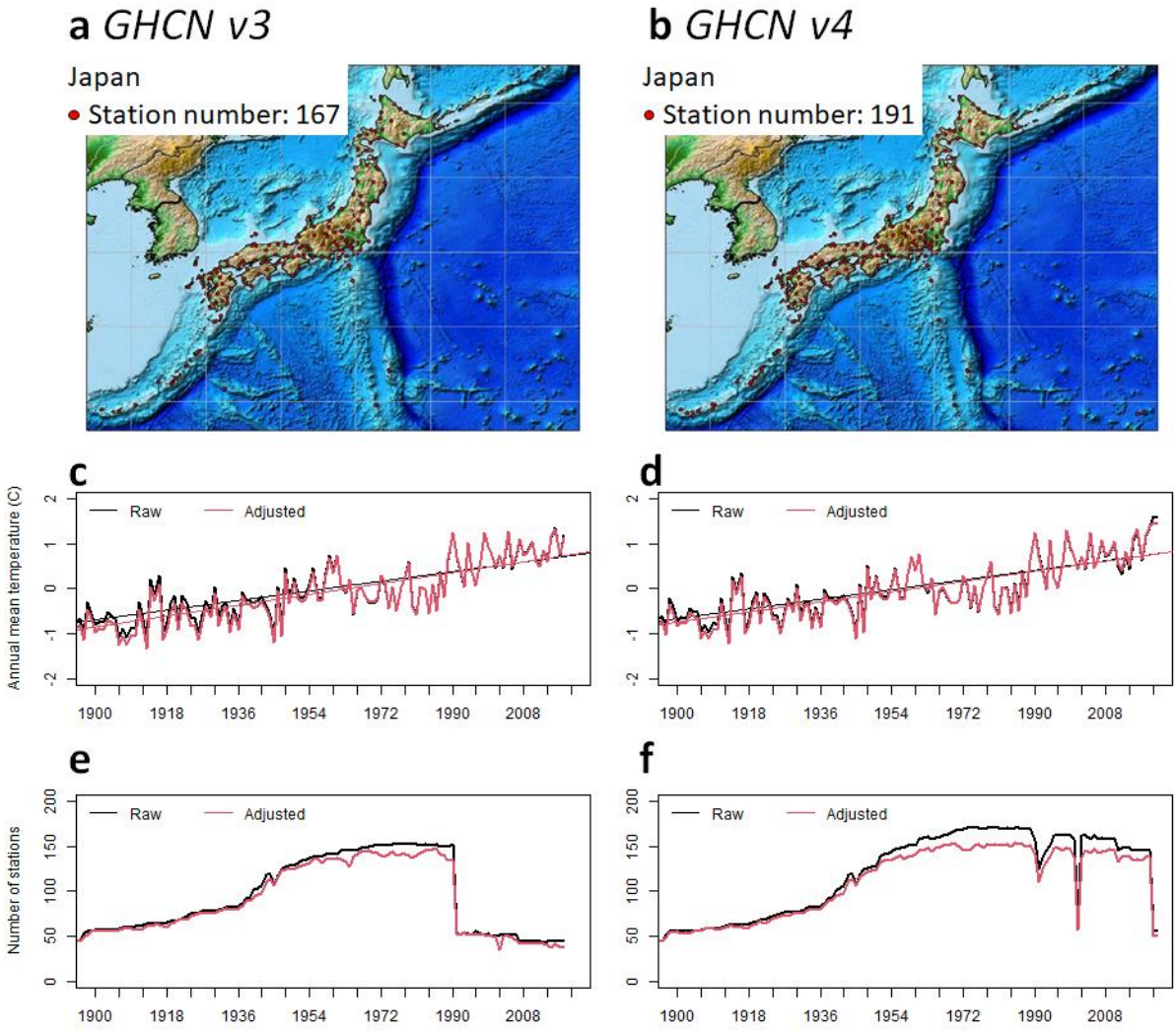


Figure 4 (a, b) The location maps of all Japanese stations regardless of record length, (c, d) annual mean temperature anomaly and (e, f) number of stations from 1900–2020 of using (a, c, e) GHCN version 3 and (b, d, f) version 4 datasets. Temperature anomalies in (c, d) are relative to a constant baseline period of 1961–1990.

Probably, the most straightforward metric for studying the urban blending phenomenon using large numbers of stations is to compare the linear temperature trends of stations before and after homogenization (He and Jia 2012; Soon et al. 2018). However, as can be seen from Fig. 4c and d, the temperature variability over the entire record is not strictly linear. Also, the timespan of each station is different. Hence, it is important to establish a suitable time period over which to calculate the linear trends.

The available station numbers in version 4 gradually increased from 1936 until the late 1950s (Fig. 4e and f). However, in version 3 there was a sharp fall in station numbers after

File generated with AMS Word template 2.0

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

1990 (Fig. 4e)—a phenomenon that has been discussed elsewhere in detail (Connolly et al. 2021; Connolly and Connolly 2014d; Lawrimore et al. 2011; Soon et al. 2015, 2018). Therefore, for our version 3 analysis, we consider the linear trends over 1955–1990, i.e., the period with maximum station coverage (*Time range 1*, as introduced below). For version 4, much of this 1990 drop-off problem appears to have been reduced, although there is a substantial drop-off in station counts in 2019. With that in mind, for version 4, we analyze three different time periods, summarized as follows:

- *Time range 1:* "maximum overlap period between versions 3 and 4 (1955–1990)";
- *Time range 2*: "maximum period for stations that are still active for version 4 (1955–2021)";
- *Time range 3*: "the longest period with a reasonable overlap for version 4 (1936–2019)".

b. Analysis of United States temperature records

The USHCN dataset was a very high quality subset of the GHCN dataset up until version 4 that was also homogenized using PHA by NOAA, but independently from the rest of the GHCN since the U.S.-based NOAA had access to additional station history information including changes in observation times (Karl et al. 1986) and other station changes (Karl and Williams 1987; Quayle et al. 1991). It was originally compiled by Karl et al. (1988) with the goal of selecting the most rural, relatively complete and climatically representative stations (or composite stations) from a much larger dataset known as the Cooperative Observer Program (COOP).

Figure 5 shows the location of the United States stations. We downloaded the datasets for minimum temperature (T_{min}), maximum temperature (T_{max}), and the average mean temperature (T_{avg}), where T_{avg} is the mean of T_{min} and T_{max}, from the link of ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/v2.5 [Last accessed in April 2023]. As can be seen from this figure, the USHCN dataset provides a very high density of stations for a country (~1200 stations) and the COOP dataset provides an even higher density (~6000 stations). When Menne et al. (2009) upgraded the USHCN to version 2.0, they switched from using Karl and Williams (1987) for their statistical homogenization to PHA (Menne and Williams 2009). They dropped the explicit empirical adjustments for changes in instrumentation (Quayle et al. 1991) and urbanization bias corrections (Karl et al. 1988) of version 1.0. They

File generated with AMS Word template 2.0

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

also decided to use the COOP dataset as the reference network for homogenization purposes (Menne et al. 2009).

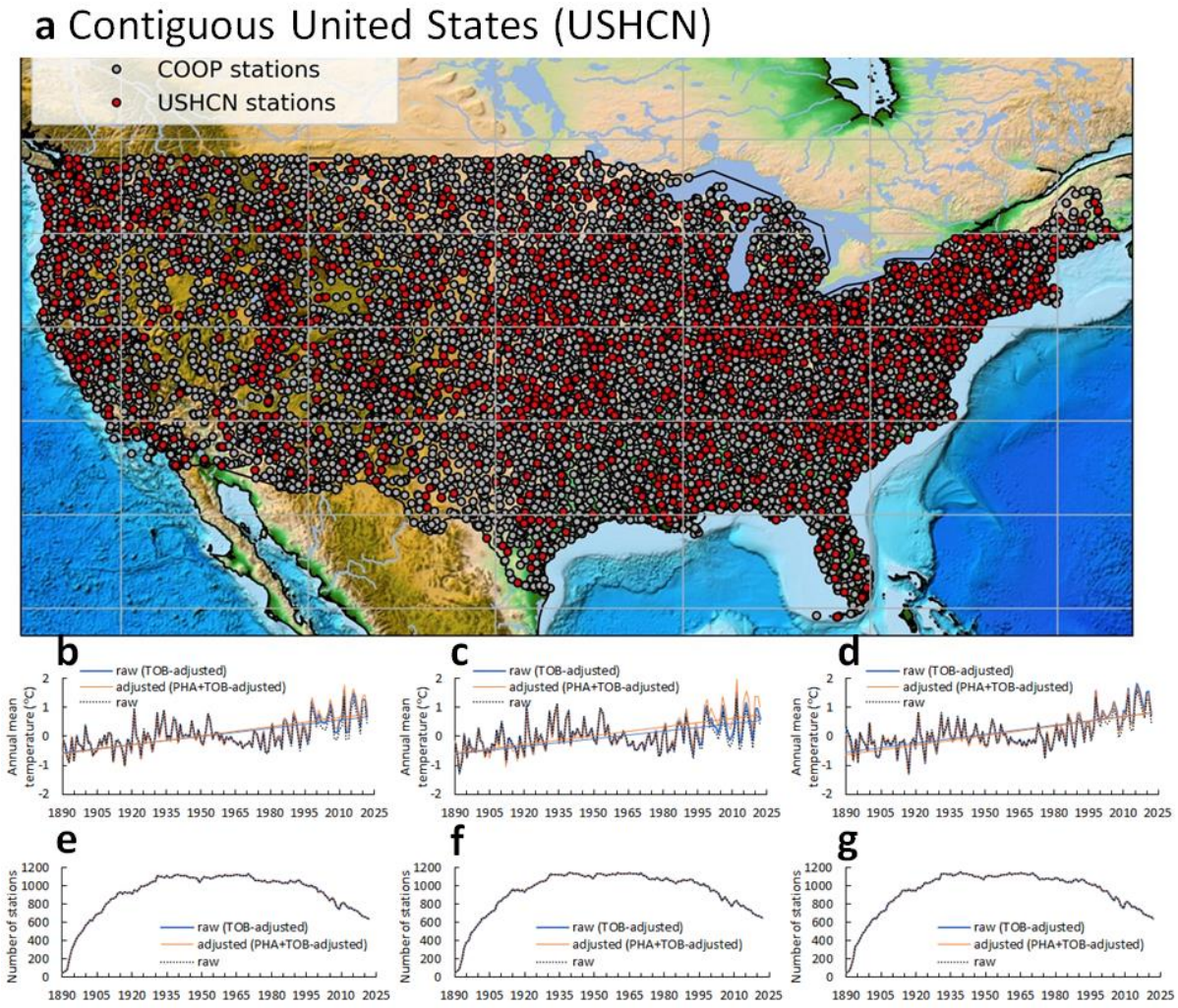


Figure 5 (a) Location of the United States stations for the USHCN dataset (red circles) as a subset of stations compiled from the larger COOP (grey circles). Our main analysis is based on the USHCN stations, although since 2009, the homogenization of the USHCN stations is carried out using the COOP stations for the reference neighbor network. This is relevant for our reanalysis of the Hausfather et al. (2013) study. Panel (b)-(d) show gridded mean annual T_{avg}, T_{max}, and T_{min} anomaly for all USHCN stations from 1895-2022, respectively. Temperature anomalies are relative to a constant baseline period of 1901–2000. Panels (e)-(g) represent the number of stations available for each year for (b)-(d), respectively.

As well as the GHCN, there is an unhomogenized ("raw") version as well as a PHA-adjusted version of the USHCN. However, *before* NOAA applies the PHA homogenization algorithm to the USHCN, they apply an independent set of time-of-observation bias (TOB) adjustments to each station to account for documented changes in observation time (Karl et

File generated with AMS Word template 2.0

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

al. 1986; Vose et al. 2003). Therefore, to study just the effects of the PHA homogenization, for our USHCN analysis we compare this partially adjusted dataset "raw (TOB-adjusted)" to the fully adjusted "adjusted (PHA+TOB-adjusted)" dataset instead of the completely unadjusted "raw" dataset.

As for Japan, we used the annual temperatures available for 12 complete months for a given year, and all temperature records (whether T_{min}, T_{max}, or T_{avg}) were first converted into an "anomaly time series". However, because the USHCN stations tend to have relatively long and complete records typically beginning in 1895 or earlier, we used a longer baseline period of 1901-2000 and required a station to have a minimum of at least 50 complete years of data during this baseline period to be incorporated into our analysis. This incorporated most of the available USHCN stations.

To calculate the gridded temperatures for each version of the USHCN, all stations were assigned into 5°× 5° horizontal grid boxes spanning the contiguous United States. For each grid box, the mean of all station anomalies available was calculated for each year (separately for T_{min}, T_{max}, and T_{avg}). The anomaly for the contiguous United States was then the area-weighted average of all grid-boxes with data for that year, where the area weighting was the cosine of the latitude of the middle of the grid-box.

In terms of time periods, as mentioned above, a large number of USHCN stations (mostly the more urbanized stations, but with many rural stations) have fairly complete records from at least 1895. Therefore, we used the 1895-2022 period for our main analysis, while we also present the results for the longest period with most stations (1917-2005) and the last century (1923-2022) in Figs. S3 and S4, respectively.

c. Estimating the degree of urbanization of the stations for both countries

Version 3 included two estimates of the degree of urbanization of each station as part of the accompanying metadata based on Peterson et al. (1999)'s ratings. These were the urbanization metrics used by (e.g., Connolly et al. 2021; Soon et al. 2015). However, no urbanization estimates were provided for version 4. Therefore, to estimate the degree of urbanization of each station, we use two metrics equivalent to those used by Soon et al. (2018), that is, 1) the average population density and 2) the average night brightness associated with the station location. The former was obtained by the Gridded Population of the World (GPW) version 4 dataset (CIESIN 2018), while the latter was derived from the

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

Global Radiance Calibrated Nighttime Lights dataset (NOAA 2015). The average values for each station location were determined from the mean of the nine pixel values centered at the station location, of the appropriate datasets for both metrics.

All stations were then ranked from most urban to most rural according to each metric. Hence, each station was initially assigned two urban rankings—one based on its local population density and the other based on its local night brightness. We then derived the degree of urbanization (*DU*) at each station, as $DU_i = 1 - R_i/n$, where *n* is the maximum station number. For Japan, *n*=167 and 191 for version 3 and version 4 of the GHCN, respectively. For the United States, *n*=1218. Scatter plots of both metrics are shown in Fig. 6.

As Soon et al. (2018) described for the Chinese stations in GHCN, both metrics provide very similar urban rankings. That is, stations with higher night brightness strongly coincide with those with higher population densities. Still, given that both metrics describe a slightly different aspect of urbanization, the exact rankings varied slightly for each metric. Therefore, for our main analysis, the overall ranking (R_i) for each station (i), was calculated as the average of the two ranks. However, we have also repeated our main analysis for Japan in Figs. S1 and S2 using each of the metrics individually, demonstrating that the results were similar to the equivalent results of Fig. 6.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

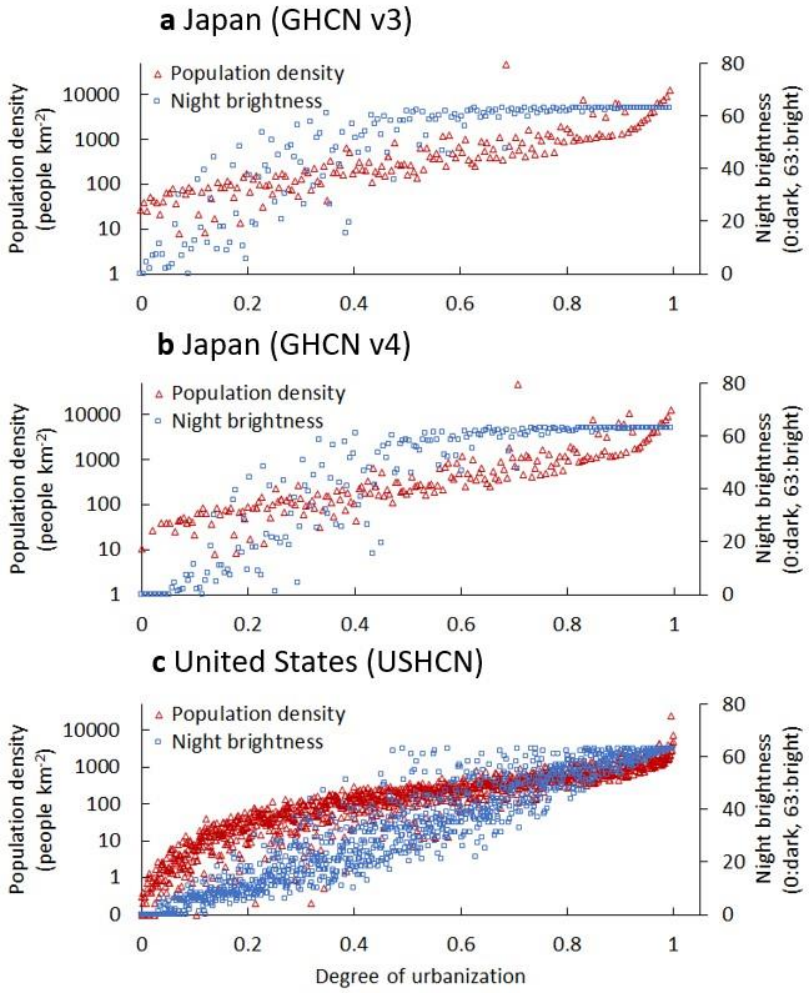


Figure 6 The population density and night brightness against the degree of urbanization (DU) for (a) all Japanese stations in GHCN version 3; (b) version 4 and (c) United States stations in USHCN. The left y-axis is shown using log-scale.

d. Reanalysis of Hausfather et al. (2013)'s results for the USHCN dataset

For our reanalysis of the Hausfather et al. (2013) study, we downloaded their supplementary information from NOAA's ftp website at: ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/papers/hausfather-etal2013-suppinfo/ [Last accessed in December 2022]. This dataset comprises different iterations of the USHCN homogenization process that were carried out using one of eight subsets of the COOP stations as reference neighbors (either "rural" or "urban" according to the thresholds identified by Hausfather et al. (2013) for one of their four urbanization metrics: ISA, population growth, night brightness or GRUMP). These iterations were carried out using an early release of the current version 2.5 PHA called "52g".

File generated with AMS Word template 2.0

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

Although NOAA's archive for Hausfather et al. (2013) also provided the results for the main part of their analysis that used all COOP stations, those results were based on the older version 2.0 PHA called "52d". Therefore, to more directly compare the Hausfather et al. (2013) rural/urban neighbor subsetting experiments to the standard approach, we need a version of USHCN that has been homogenized using version 2.5 PHA. Unfortunately, NOAA do not house a public archive of the previous daily updates of the USHCN dataset, but just the latest iteration. However, fortunately, the version we use for our main analysis in this paper is also version 2.5 PHA ("52j") and hence suitable for our reanalysis of Hausfather et al. (2013).

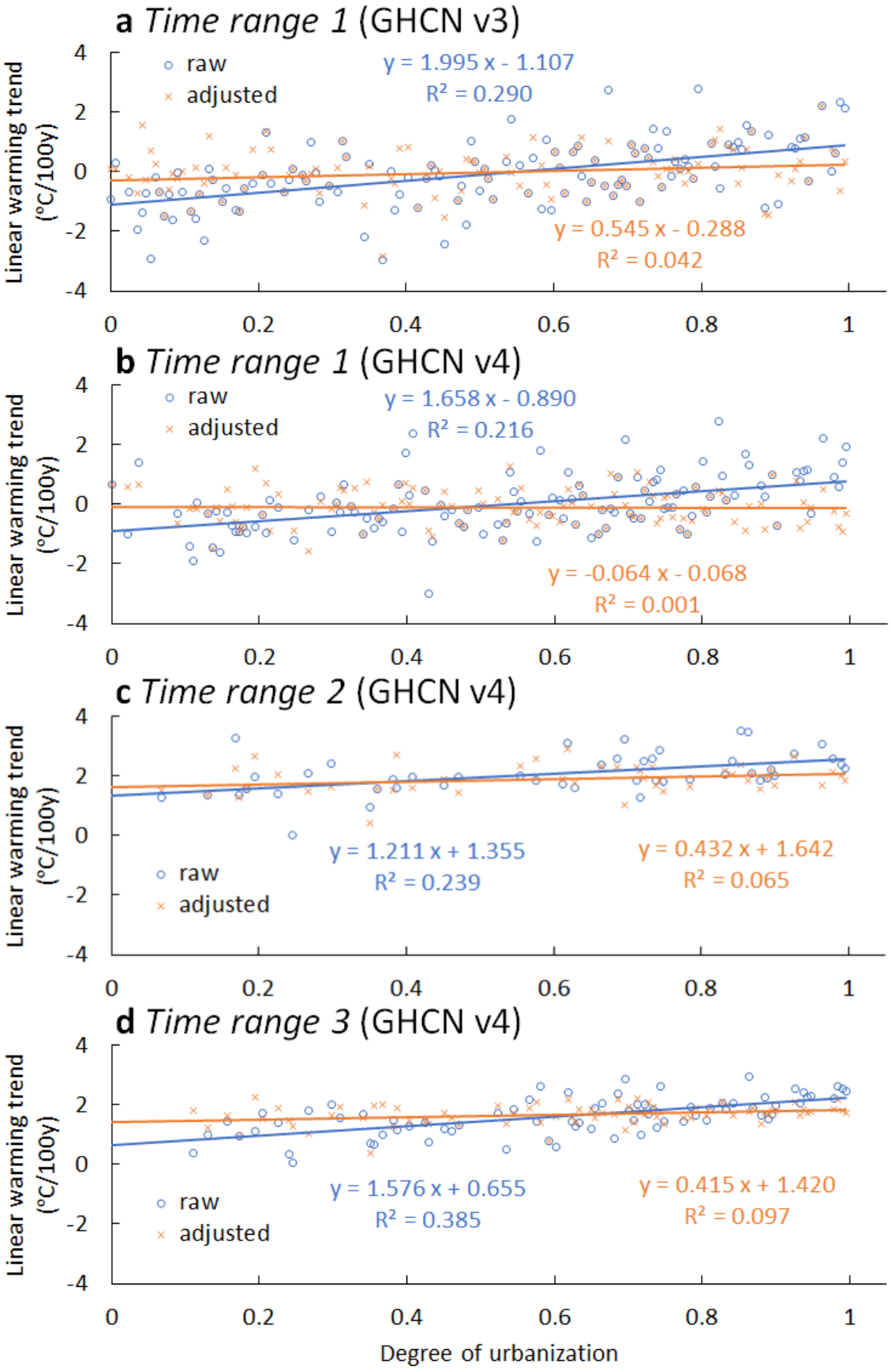
Hausfather et al. (2013) only discussed and provided the T_{min} and T_{max} results and their discussion of aliasing focused on T_{min}. Therefore, our reanalysis is based on T_{min} and T_{max}., with a particular focus on T_{min}. For our reanalysis, for direct comparison with Hausfather et al. (2013), we calculated the linear trends (using linear least squares fitting) over the 1895-2010 period in units of °C/century.

4. Results

a. Urbanization bias and urban blending in the Japanese temperature data

Figure 7 shows the relationship between the degree of urbanization and linear warming trends of GHCN versions 3 and 4 for each *Time range*. In all panels, there is an approximately linear relationship between the linear temperature trend and degree of urbanization for the raw data. The slopes and intercepts of these linear relationships are provided in each panel along with the R^2 values. Table 1a summarizes the statistics of linear temperature trend equations against the degree of urbanization (*DU*). This table provides the *p* values associated with these relationships and in all cases, *p* is much lower than 0.05 for raw data. In other words, the more urbanized the station the greater the warming trend. Therefore, the Japanese unhomogenized (raw) temperature records are strongly influenced by urbanization bias.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.



Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

Figure 7 (a-b) Linear warming trends for the Japanese stations of raw and adjusted GHCN data against the degree of urbanization (*DU*) for *Time range 1* (versions 3 and 4), (c) *Time range 2*, and (d) *Time range 3* in Table 1a.

a. Statistics of linear temperature trend equations of against *DU*

Time range number	GHCN version	Station number	Period analyzed	Trend (ºC/century)		Coefficient of determination, R^2		*p*-value for trend	
				Raw	Adjusted	Raw	Adjusted	Raw	Adjusted
1	v3	124	1955–1990	1.995	0.545	0.290	0.042	**<0.001**	**0.022**
1	v4	121	1955–1990	1.658	–0.064	0.216	0.001	**<0.001**	0.756
2	v4	47	1955–2021	1.211	0.432	0.239	0.065	**0.0011**	0.124
3	v4	76	1936–2019	1.576	0.415	0.385	0.097	**<0.001**	**0.006**

b. Temperature trends estimated at fixed *DU* using the equations based on **a**

Time range number	GHCN version	Most rural, *DU*=0 (ºC/century)		Average urban, *DU*=0.5 (ºC/century)		Most urban, *DU*=1 (ºC/century)	
		Raw	Adjusted	Raw	Adjusted	Raw	Adjusted
1	v3	–1.107	–0.288	–0.109	–0.015	0.889	0.257
1	v4	–0.890	–0.068	–0.061	–0.100	0.768	–0.132
2	v4	1.355	1.642	1.961 (+31%)	1.858 (+12%)	2.566 (+47%)	2.074 (+21%)
3	v4	0.655	1.420	1.443 (+55%)	1.627 (+13%)	2.231 (+71%)	1.835 (+23%)

Table 1 (a) Statistics of linear temperature trend equations against the degree of urbanization (*DU*), and (b) trends estimated at fixed *DU* values of 0, 0.5, and 1 for the Japanese GHCN version 3 and 4 networks for time ranges 1–3. In (a), bold numbers represent statistically significant ($p < 0.05$). In (b), the values enclosed in parentheses for time ranges 2 and 3 indicate the relative increase in the warming trend from the most rural situation (*DU*=0). For the other two periods, the most rural situation yielded cooling trends.

The linear regression equations are of the form, $y=mx+c$, where y is the linear warming trend in °C/century and x is *DU*. Therefore, they allow us to estimate the average linear warming trend we would expect for a given *DU*. Table 1b shows the temperature trends estimated at fixed *DU* values for *Time range 1–3* using linear regression equations obtained from Table 1a. For all periods analyzed, increasing the *DU* value adds warming as expected from the urban heat island effect. Indeed, in the case of the 1955–1990 periods, the "most rural" trends are cooling trends, i.e., –1.107°C/century for version 3 and –0.89 °C/century for version 4, but the "most urban" trends are always positive.

After homogenization, the apparent linear relationships between temperature trend and degree of urbanization are substantially reduced, as indicated by the very small R^2 values and higher *p* values (Table 1a). The linear fits for version 4 for 1955–1990 and 1955–2021 are not statistically significant ($p>>0.05$). Initially, one might (mistakenly) interpret this reduction as

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

meaning the homogenization process had somehow removed most of the urbanization biases from the data. However, a visual inspection of Fig. 7 shows that it is mostly due to urban blending—that is, homogenization has consistently added extra warming to the least urbanized stations and reduced the warming of the most urbanized stations until the trends of all stations have converged towards those of the network averages. This can be seen graphically by the fact that, for each panel of Fig. 7, the raw and adjusted linear fits intersect near the middle of the x-axis, i.e., around *DU*=0.5. If the homogenization process were reducing the urbanization bias relationship by truly removing urbanization bias, then the adjustments should instead act to converge all trends towards those of the most rural stations (*DU*=0). Yet, instead, the adjustments act to converge all trends towards the average of the network (*DU*=0.5). For example, let us consider the temperature trends for GHCN version 4 for *Time range 3* (1936–2019). The mean temperature trend for the entire network over this period was 1.68°C/century (not shown in figure) for the adjusted data. This is quite similar to that for the raw data as 1.63°C/century (not shown in figure). However, it is approximately 2.5 times higher than the "most rural" trend of 0.655°C/century (Table 1b). Very similar values between raw and adjusted mean temperature trends for all GHCN stations (Fig. 4c and d) can be also explained as the above average of the network (*DU*=0.5), i.e., urban blending due to homogenization.

Another way to evaluate the extent of urban blending in the Japanese data is to consider the trends of the 20% most urban stations (*DU*>0.8) and the 20% most rural stations (*DU*<0.2). In Fig. 8, we plot these trends for *Time range 3* as an example. While the temperature trends for the urban stations decreased from 2.14 to 1.80 ºC/century after homogenization was made, the trends for the rural stations increased from 0.97 to 1.56 ºC/century (Fig. 8a and b).

It might be tempting to treat the unhomogenized trends of the most rural stations as representative of "rural Japan", i.e., 0.97 ºC/century over the period 1936-2019. However, we caution that the station numbers for both of these subsets are very low (Fig. 8c and d) and especially so for the rural subset that only comprises 5 stations. Moreover, for this subset, no attempt has been made to correct for non-urban related non-climatic biases. That said, since the homogenized series is affected by urban blending, the raw series is probably more representative of "rural Japan" than the PHA-adjusted version.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

For version 4 of the GHCN, all PHA-adjustments are determined statistically without reference to any station history metadata (O'Neill et al. 2022). However, apparently some station history metadata for Japanese stations is available online at https://www.data.jma.go.jp/obd/stats/data/mdrr/chiten/meta/discnt_sfc.csv [last accessed May 2023]. Therefore, we suggest that future research into evaluating rural Japanese trends could use such information to account for other non-climatic biases in the data.

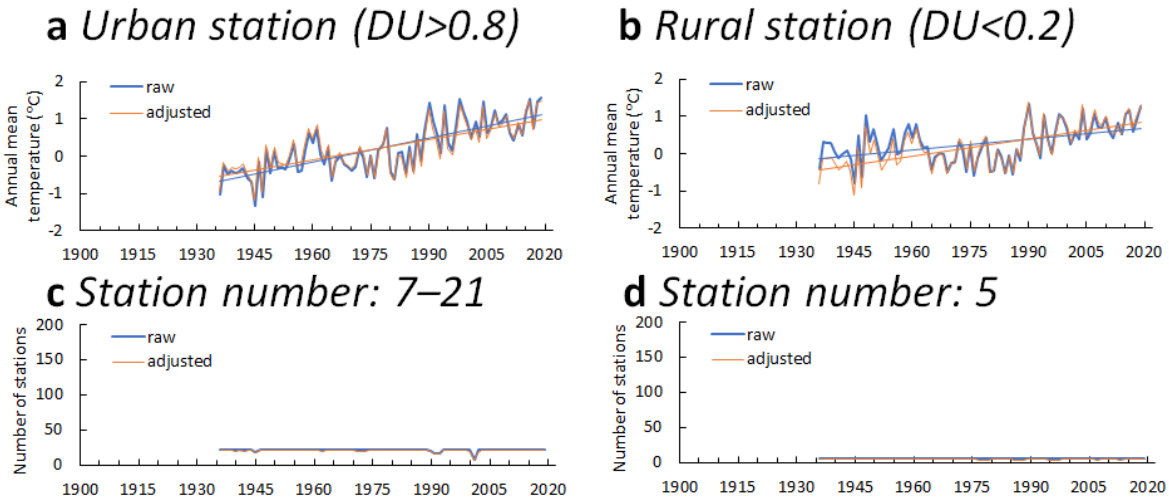


Figure 8 (a, b) Annual mean temperature anomaly and (c, d) number of urban (*DU*>0.8) and rural stations (*DU*<0.2) from 1900–2020 of using GHCN version 4 dataset for *Time range 3* in Table 1a. Temperature anomalies in (c, d) are relative to a constant baseline period of 1961–1990.

There are also challenges associated with the use of other types of data to estimate long-term rural Japanese temperature trends. Satellite-based tropospheric temperature estimates should be unaffected by urbanization, and interestingly, they suggest less warming than thermometer-based records (McKitrick and Christy 2020; Zou et al. 2023). However, they are confined to the satellite era (1978-present) and focus on atmospheric temperature trends rather than near-ground temperatures. Sea surface temperatures and marine air temperatures should be unaffected by urbanization biases, but debate is ongoing over how to best resolve the non-climatic biases associated with those datasets – especially for the earlier, data-sparse periods before the mid-20th century (Kent et al. 2017; Kent and Kennedy 2021). Japan also offers intriguing long-term records that are potentially useful for studying changes in spring-time temperatures, i.e., the flowering dates of cherry trees which have been recorded in some locations for over 1200 years (Aono and Kazui 2008; Aono and Saito 2010; Christidis et al.

File generated with AMS Word template 2.0

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

2022). However, these records are mostly associated with urbanized areas and hence significantly affected by urban warming (Aono and Kazui 2008; Christidis et al. 2022).

b. Urbanization bias and urban blending in the United States temperature data

Version 3 of the GHCN incorporated the homogenized USHCN dataset as part of the full dataset (Lawrimore et al. 2011). Although version 4 of the GHCN no longer explicitly carries out this step of separately homogenizing the USHCN and the rest of the GHCN, the USHCN dataset is still being updated and maintained by NOAA and many of the USHCN stations and COOP stations are included in the GHCN (Menne et al. 2018). Indeed, stations from the contiguous U.S. represent ~40% of the GHCN dataset (Menne et al. 2018). Moreover, as can be seen from Fig. S5, the USHCN is much less urbanized than the Japanese data. Therefore, let us now assess how prevalent urban blending is in the USHCN.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

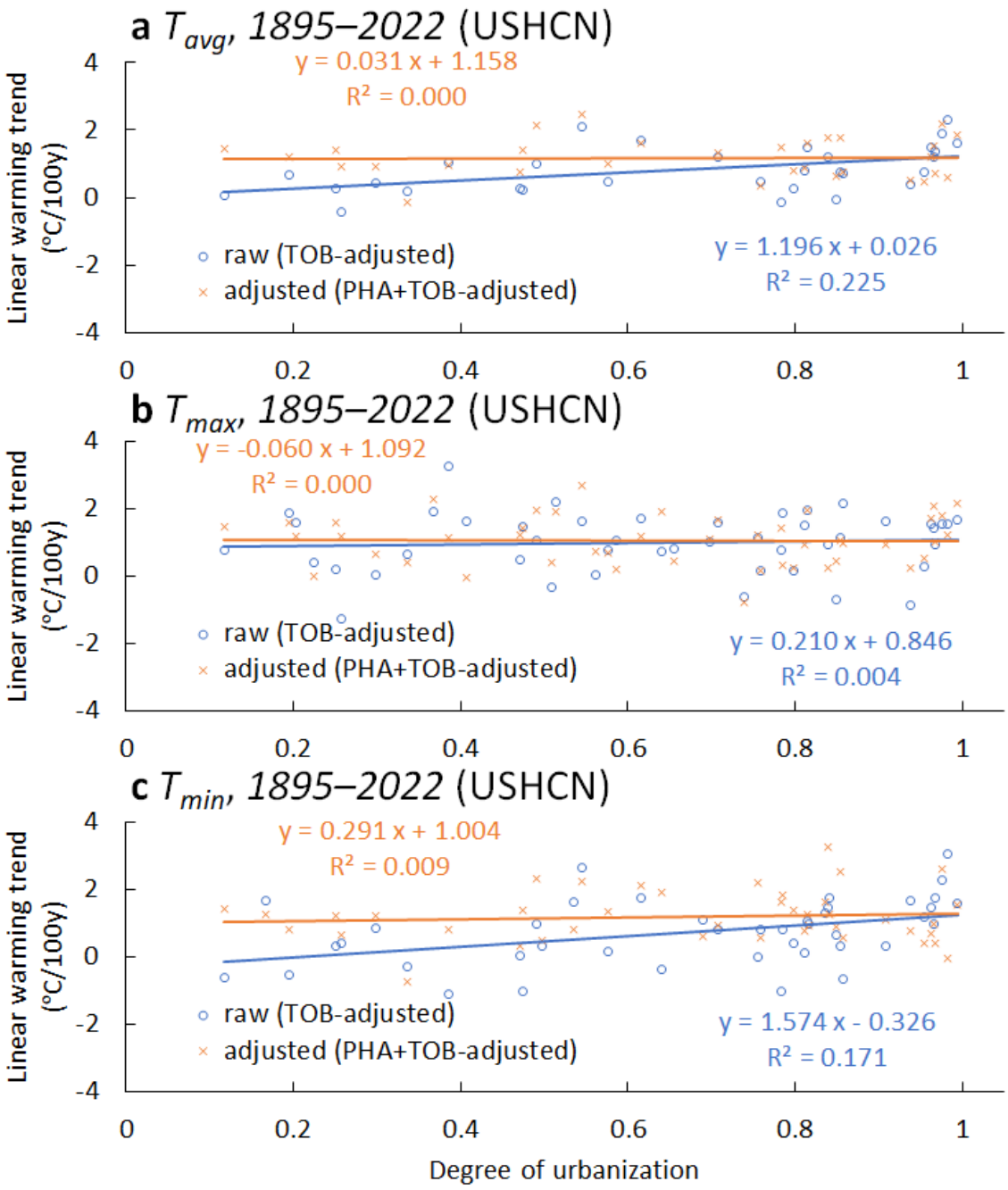


Figure 9 Linear warming trends of (a) T_{avg}, (b) T_{max}, and (c) T_{min} for the United States (USHCN) stations for the raw (TOB-adjusted) and adjusted data against the degree of urbanization (*DU*) for 1895-2022. Only stations that had at least 98% coverage for the full 1895-2022 period are plotted.

Figure 9a shows the equivalent T_{avg} results to Fig. 7 for the United States, except that the period is longer (1895-2022), while the trends for T_{max} and T_{min} are illustrated in Fig. 9b and

File generated with AMS Word template 2.0

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

c. It has already been noted for the USHCN (Hausfather et al. 2013) that the urban warming effect is mostly a phenomenon for T_{min}.

Indeed, as can be seen from Fig. 9b, there is not much difference in trends for T_{max} with increasing *DU* (either before *or* after PHA homogenization). In contrast, we can see from Fig. 9c, for the raw (TOB-adjusted) T_{min} data, the 1895-2022 linear trend increases with increasing *DU*, i.e., there is a noticeable urban warming effect. The equation of the line implies a warming trend for the most urban stations (*DU*=1) of +1.25°C/century, while there is a cooling trend for the most rural stations (*DU*=0) as −0.33°C/century over 1895-2022. However, most of this urban warming effect is apparently reduced after PHA—the difference in trends between most and least urban stations is reduced from +1.57°C/century to +0.29°C/century after homogenization. The results for T_{avg} are intermediate since T_{avg} is the average of T_{min} and T_{max}.

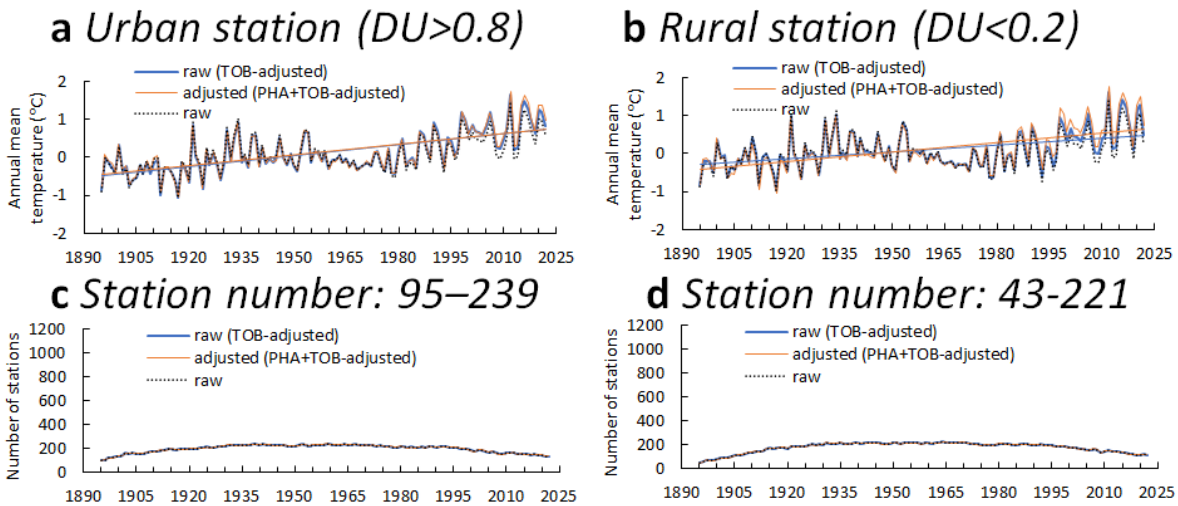


Figure 10 (a, b) Annual mean temperature anomaly and (c, d) number of urban (*DU*>0.8) and rural stations (*DU*<0.2) from 1895–2022 of USHCN stations. Temperature anomalies in (c, d) are relative to a constant baseline period of 1901–2000.

For the trend periods, we studied for Japan, T_{min} (and T_{avg}) blending occurred towards the average *DU* of the network, i.e., the point of intersection of the two lines was at *DU*≈0.5. However, here, blending appears to be towards the most urbanized stations, i.e., the point of intersection is at *DU*≈0.9. The point of intersection is slightly less urbanized at *DU*≈0.8 for Fig. S3 (1917-2005), but similar for Fig. S4 (1923-2022). This is probably at least partially a consequence of the fact that—unlike the GHCN dataset – the reference neighbor network used for homogenizing the USHCN dataset is a different dataset, i.e., the larger COOP

File generated with AMS Word template 2.0

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

dataset (Fig. 5a). Hence, for the USHCN, urban blending should converge towards the average urbanization of this larger COOP dataset rather than that of the USHCN. When the USHCN stations were originally been selected from the COOP dataset, one of the selection criteria was to identify relatively rural stations (Karl et al. 1988).

In Fig. 10, we compare the different T_{avg} trends for United States estimated using either (a) and (c) the 20% most urban stations of the USHCN (DU>0.8) or (b) and (d) the 20% most rural stations (DU<0.2). At the visual resolution plotted here, the differences might appear subtle. However, they are quite substantial, as can be seen from Table 2. Equivalent results for different periods (1917-2005 and 1923-2022) and for the "raw" dataset can be found in Table S1 and S2, along with the gridded time series.

	raw (TOB-adjusted)		adjusted (TOB+PHA-adjusted)	
Subset	Trend (°C/century)	Coefficient of determination (R^2) and p-value for trend	Trend (°C/century)	Coefficient of determination (R^2) and p-value for trend
a. T_{min} for 1895-2022				
All stations	1.020	0.43 (**p<0.001**)	0.973	0.42 (**p<0.001**)
Most rural (DU<0.2)	0.765	0.28 (**p<0.001**)	0.841	0.34 (**p<0.001**)
Most urban (DU>0.8)	1.299	0.55 (**p<0.001**)	0.983	0.43 (**p<0.001**)
Urban bias in "all stations"	0.255 (25.0%)		0.132 (13.6%)	
Urban bias in "most urban" subset	0.534 (41.1%)		0.142 (14.4%)	
b. T_{max} for 1895-2022				
All stations	0.514	0.16 (**p<0.001**)	0.925	0.35 (**p<0.001**)
Most rural (DU<0.2)	0.490	0.13 (**p<0.001**)	0.845	0.28 (**p<0.001**)
Most urban (DU>0.8)	0.590	0.20 (**p<0.001**)	0.932	0.36 (**p<0.001**)
Urban bias in "all stations"	0.024 (4.7%)		0.080 (8.6%)	
Urban bias in "most urban" subset	0.100 (16.9%)		0.087 (9.3%)	
c. T_{avg} for 1895-2022				
All stations	0.766	0.32 (**p<0.001**)	0.949	0.41 (**p<0.001**)
Most rural (DU<0.2)	0.616	0.22 (**p<0.001**)	0.839	0.33 (**p<0.001**)
Most urban (DU>0.8)	0.948	0.43 (**p<0.001**)	0.955	0.42 (**p<0.001**)
Urban bias in "all stations"	0.150 (19.6%)		0.110 (11.6%)	
Urban bias in "most urban" subset	0.332 (35.0%)		0.116 (12.1%)	

Table 2 Statistics of linear temperature trend equations of (a) T_{min}, (b) T_{max}, and (c) T_{avg} against the degree of urbanization (DU) for 1895-2022 period for all USHCN stations and "most rural" (DU<0.2) and "most urban" stations (DU>0.8). The difference in trends (defined as "urban bias") between "most rural" stations and "all stations" or "most urban" stations are also shown in the table. Bold numbers represent statistically significant (p < 0.05).

It might be tempting to treat the raw T_{avg} trends of the most rural stations (DU<0.2) as representative of "rural United States", i.e., +0.616 ºC/century over the period 1895-2022.

File generated with AMS Word template 2.0

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

Indeed, this data has already been adjusted by NOAA to account for the documented changes in TOB that collectively introduce a long-term net cooling bias for both T_{min} and T_{max} and hence T_{avg} (Karl et al. 1986; Balling and Idso 2002; Vose et al. 2003). However, there are also other known non-climatic biases associated with the USHCN data:

- The network-wide transition from analogue to digital thermometers in the 1980s-1990s is associated with a cooling bias of ~0.4 ºC for T_{max}, a warming bias of ~0.3 ºC for T_{min} and a net cooling bias of ~0.1 ºC for T_{avg} (Quayle et al. 1991; Hubbard and Lin 2006; Menne et al. 2009).
- There seems to have been a network-wide reduction in the average quality of the station exposure leading to siting biases that are collectively associated with a long-term warming bias for T_{min} (Fall et al. 2011).

We encourage further research to account for these known biases without introducing urban blending. Meanwhile, since the homogenized series is affected by urban blending—and possibly blending of siting biases (Connolly and Connolly 2014a; Soon et al. 2018)—the raw (TOB-adjusted) series derived from the 20% most rural stations is probably more representative of "rural United States" than the PHA-adjusted version.

c. Reanalysis of Hausfather et al. (2013)'s USHCN aliasing experiments

The findings described above for the USHCN initially appear to contradict one of the conclusions of Hausfather et al. (2013), an important study that attempted to quantify "the effect of urbanization on U.S. Historical Climatology Network temperature records".

As part of their analysis, Hausfather et al. (2013), briefly considered the possibility that aliasing of urbanization bias might be a concern for the USHCN. To test this, they repeated the PHA procedure using subsets of the COOP stations that had been divided into "rural" or "urban" according to one of four urbanization metrics (Fig. 11) for details of the station breakdown.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

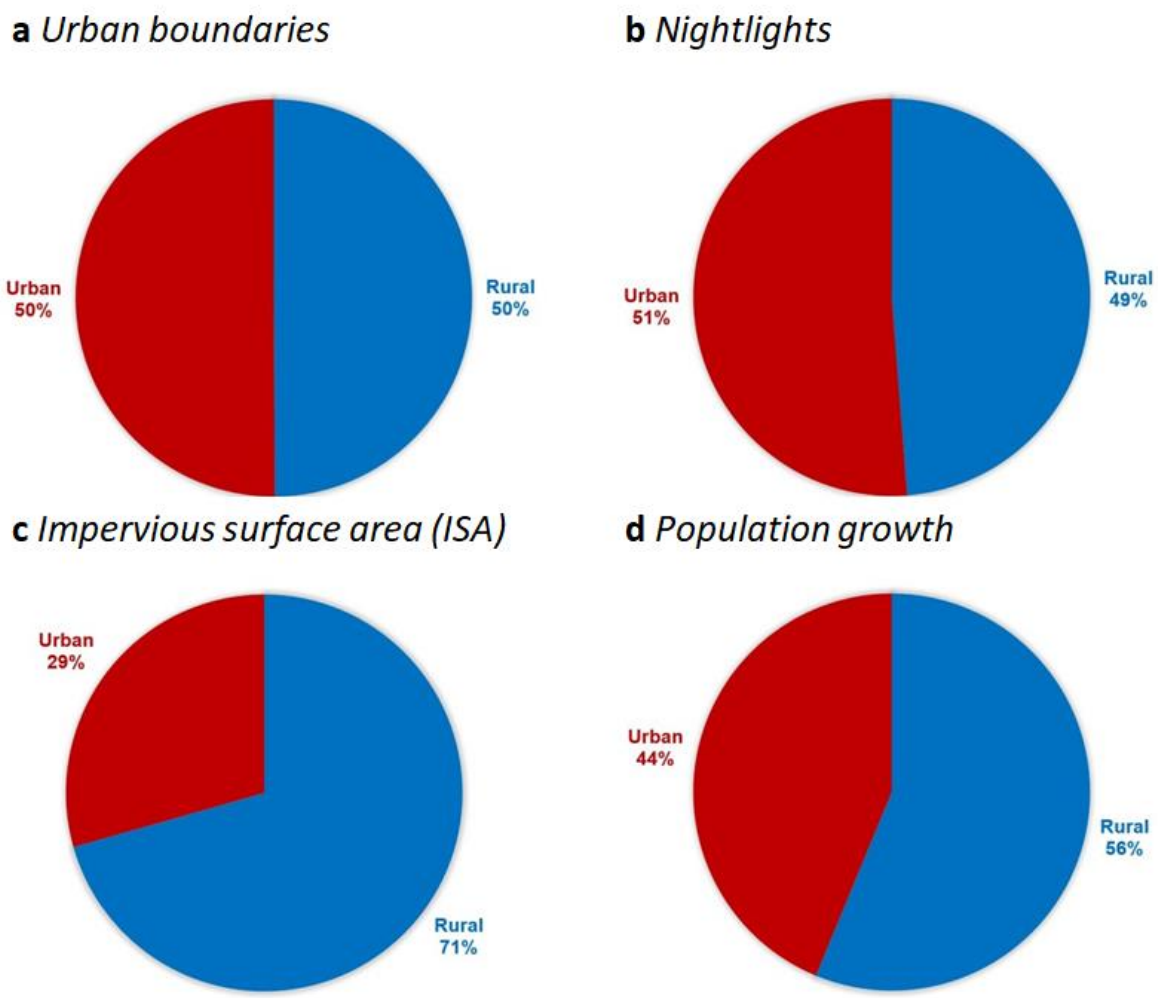


Figure 11 Breakdown of how many USHCN stations were classified as either "rural" or "urban" by each of Hausfather et al. (2013)'s four urbanity proxies.

Since aliasing was not the primary focus of their study, their discussion of aliasing mostly focused on the results for the impervious surface area classification which identified the 29% most urban stations as "urban" and the remaining 71% as "rural" (Fig. 11c). However, they noted that the "(r)esults using the other three station classification approaches are similar" (Hausfather et al. 2013).

Their discussion focused on T_{min} trends since they had already established the urban warming effects were most obvious for this metric. They firstly confirmed that when urban-only neighbors were used, significant aliasing of urban warming occurred for T_{min}. However, they noted that the homogenization adjustments when using either rural-only neighbors or all neighbors were "nearly identical" and concluded that "the Coop neighbors that surround USHCN stations are sufficiently 'rural' to prevent a transfer of undetected urban bias from the neighbors to the USHCN station series during the homogenization procedure."

File generated with AMS Word template 2.0

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

(Hausfather et al. 2013). This assessment appears to have led them to conclude that aliasing was not a particular concern for the USHCN dataset at least. Indeed, they suggested that using all stations had the advantage of more dense station network.

Therefore, we have reanalyzed the subsetting results from Hausfather et al. (2013) to investigate the reasons for the apparent contradiction with our findings. In Fig. 12, we have plotted the relevant linear trends for the longest period considered by Hausfather et al. (2013), i.e., 1895-2010. The annual adjustments are also plotted in Fig. S6. After analysis, we noted that the archived data provided by NOAA for "rural using urban boundaries" and "rural using population growth" are identical copies.

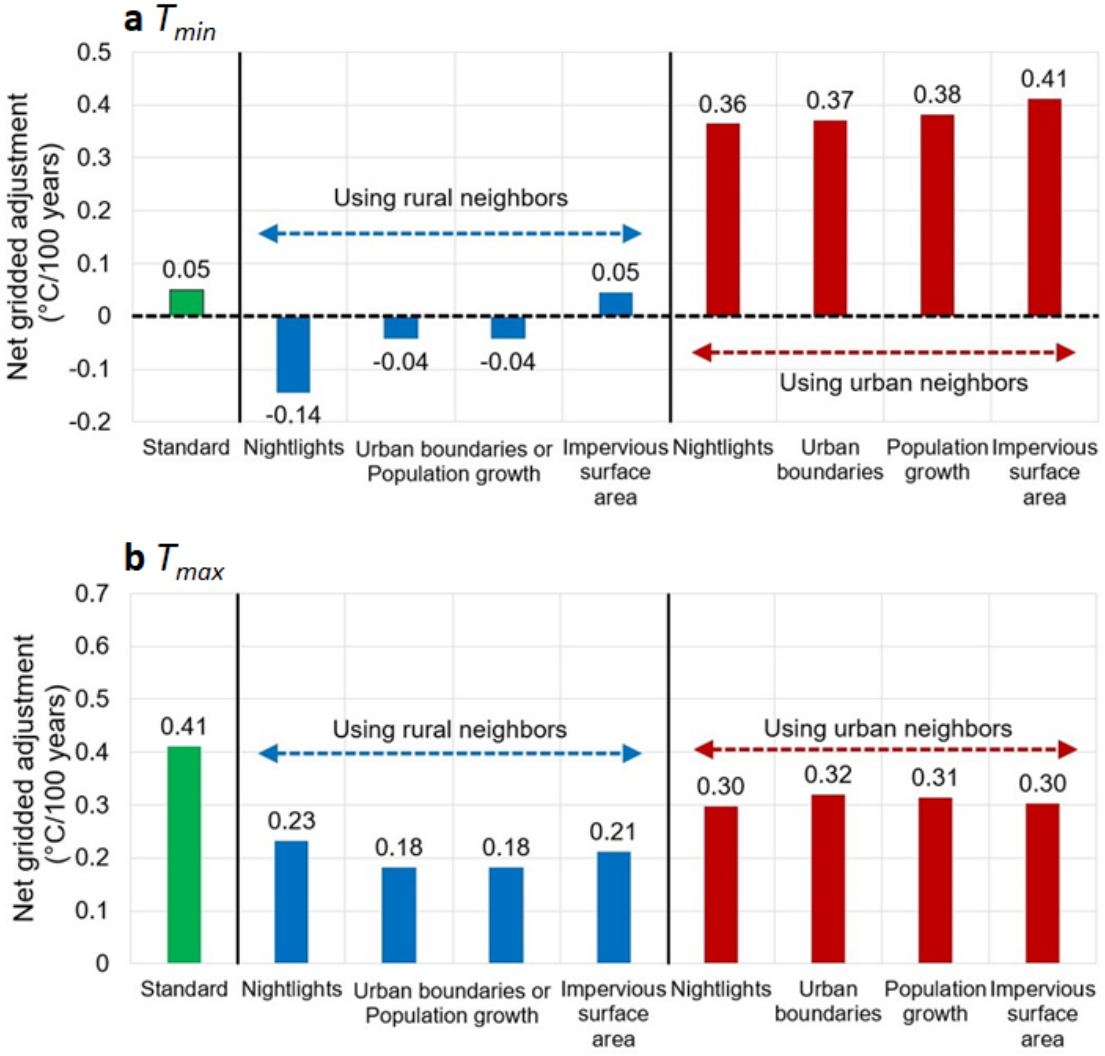


Figure 12 The 1895-2010 linear trends from the net gridded homogenization adjustments applied to (a) the T_{min} data and (b) T_{max} for the USHCN stations depending on whether they were homogenized using all COOP stations, i.e., the standard approach; only COOP stations identified by "rural" according to each of Hausfather et al. (2013)'s four urbanity proxies;

File generated with AMS Word template 2.0

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

only COOP stations identified by "urban" according to each of Hausfather et al. (2013)'s four urbanity proxies. Note that in NOAA's ftp archive of the Hausfather et al. (2013) results, the data for either "rural using urban boundaries" or "rural using population growth" appears to have been inadvertently duplicated and replaced the other, i.e., these two datasets are identical copies and therefore are plotted here with identical trends.

We can confirm that the trends of the adjustments for T_{min} are indeed "nearly identical" when using the impervious surface area "rural-only" stations or all stations, i.e., both add an extra +0.05°C/century to the homogenized trends (Fig. 12a). However, as can be seen from Fig. 11, this particular urbanization metric was the least restrictive of the four "rural" thresholds, only excluding 29% of the stations. When any of the other metrics were used, the 1895-2010 T_{min} trends of the adjustments were negative, i.e., homogenization cooled the USHCN temperature records. Indeed, for the most restrictive of the four "rural" thresholds (nightlights), the net adjustments led to a substantial cooling of –0.14°C/century.

Our reanalysis also confirms Hausfather et al. (2013)'s other finding that using urban-only neighbors leads to urban blending for T_{min} (Fig. 12a). We note that the homogenization adjustments for T_{max} are also different when using either rural-only or urban-only compared to all stations (Fig. 12b).

Therefore, while Hausfather et al. (2013)'s discussion of aliasing was only a minor aspect of their study and their qualitative assessment that the adjustments for one of their "rural-only" subsets were indeed "nearly identical", we believe that their data confirm our findings that urban blending is a significant concern even for relatively rural regions such as the contiguous U.S.

5. Discussion and conclusions

In this paper, we demonstrated the problem of urban blending associated with the statistical homogenization of temperature records using two different countries as case studies – Japan and United States.

Urban blending is a subtle, but insidious, unintended consequence of using a network of both urbanized and non-urbanized stations as reference stations for statistical homogenization (Section 2 for an overview). It reduces the apparent differences in temperature trends between the most urban and most rural stations by reducing the warming of the most urban stations and adding warming to the most rural stations (Figs. 7 and 9).

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

The net effect of urban blending is that the trends of all *homogenized* stations converge towards the average trends of the dataset. This is a problem because the converging of the trends is towards the *average of the station network (i.e., a mix of urban and rural stations)* rather than towards those of the least urbanized. Therefore, if a substantial amount of urbanization bias is associated with the unhomogenized (raw) temperature data, then urban blending will be a significant concern for the homogenized (adjusted) dataset. We emphasize that many attempts to evaluate the extent of urbanization biases by comparing the differences between *homogenized* rural and urban trends (e.g., Li et al. 2004; Peterson et al. 1999) do not appear to have considered this urban blending problem.

Our analysis firstly reveals that the unhomogenized temperature records for Japan are heavily contaminated by urbanization bias. For example, the average temperature trends for *Time range 3* of 1936-2019 (the longest period with a reasonable overlap for stations) of GHCN version 4 are +0.655°C/century for the most rural stations but +2.231°C/century for the most urban stations, i.e., 71% of the warming of the most urban stations could be due to urbanization (Table 1b). This range from 0.655 to 2.231°C/century encompasses 1.65°C/century, i.e., the trend estimated for *Time range 3* using the data from the Japan Meteorological Agency (JMA) for a selection of 15 stations that are considered not to have been highly influenced by urbanization and have continuous records from 1898 onwards (JMA 2022). However, past studies have suggested that even in this sample of 15 stations, there might still be some urbanization bias, since several of the sites are moderately urbanized (Fujibe and Ishihara 2010).

For the United States, the network is not as heavily urbanized as Japan and there is a large number of rural stations with relatively long records (often covering more than a century). However, even here urbanization bias has noticeably affected the data. For example, for the unhomogenized records that have been empirically adjusted for documented changes in time-of-observation, the T_{avg} trends for the longest period with a reasonable overlap for stations (1895-2022) are +0.616°C/century for the 20% most rural stations, in contrast to +0.948°C/century for the 20% most urban stations and +0.766°C/century for all stations (rural and urban). That is, ~20% of the warming for the full network and ~35% for the most urban stations could be due to urbanization.

Secondly, our analysis reveals urban blending is a serious problem for the homogenized records for both Japan (Fig. 7) and the United States (Fig. 9). Although a previous study

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

(Hausfather et al. 2013) had included an analysis that suggested urban blending was not a major problem for the USHCN, our reanalysis of the Hausfather et al. (2013) data reveals that urban blending is indeed a problem for the USHCN dataset.

Some previous studies have cautioned against the urban blending problem, sometimes called "statistical aliasing of trends" (Connolly and Connolly 2014d; DeGaetano 2006; Pielke et al. 2007; Soon et al. 2015; Soon et al. 2018, 2019). Others have assessed the problem to be minor or negligible (Hausfather et al. 2013; Menne et al. 2009) or even beneficial (Menne and Williams 2009). Most studies appear to have overlooked the problem until now. However, the results of this study show that the problem should be a major concern for users of current homogenized temperature datasets.

The goal of homogenizing temperature datasets is an admirable one—to reduce the non-climatic biases in the underlying data—thereby hopefully allowing users of the homogenized datasets to assume any trends are genuinely climatic in nature. For this reason, current global LSAT estimates typically explicitly rely on homogenized records (Lenssen et al. 2019; Menne et al. 2018; Osborn et al. 2021; Sun et al. 2022; Vose et al. 2021). However, most current approaches to statistically homogenizing these temperature records do not appear to have explicitly considered the urban blending problem. Therefore, the homogenized temperature datasets currently being used for evaluating LSAT trends are contaminated by urban blending.

In terms of global temperature datasets, the latest IPCC Working Group 1 report concluded that urbanization bias was unlikely to have contributed more than 10% to global land temperature trends, although they conceded that "larger signals have been identified in some specific regions, especially rapidly urbanizing areas such as eastern China" (IPCC 2021). However, several studies disagree with that particular claim of the IPCC and suggest that urbanization bias might account for greater than 10% of the global land temperature trends (Soon et al. 2015; Connolly et al. 2021; Scafetta 2021; Zhang et al. 2021). Zhang et al. (2021) calculate that urbanization bias accounts for 12.7% of the 1951-2018 global land temperature trends. Soon et al. (2015)'s analysis implies that urbanization bias accounted for 32.7% of the 1881-2014 trends for the Northern Hemisphere, while Connolly et al. (2021)'s update implies that urbanization bias accounted for 38.2% of 1850-2018 trends. Meanwhile Scafetta (2021) estimated that urbanization bias could account for up to 25-45% of the global warming of the last 40-80 years.

File generated with AMS Word template 2.0

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

Another concern is that much of the justification for the IPCC's low estimate of urbanization bias in the global temperature data appears to be based on *homogenized* temperature records. Therefore, their estimates may have been biased low due to urban blending.

Therefore, given the wide range of estimates for the degree of urbanization bias in the global temperature datasets, it is still unclear exactly how widespread and how large the urban blending problem is. However, it is probably not insignificant and more research into this challenging problem should be encouraged.

We note that this has implications for many attempts to attribute LSAT trends between natural and anthropogenic factors since most such attempts appear to implicitly assume the homogenized temperature is relatively unaffected by urbanization bias (e.g., Gillett et al. 2021; Masson-Delmotte et al. 2021), other than a few exceptions (Connolly et al. 2021; Soon et al. 2015; Sun et al. 2016, 2019).

In terms of solutions to the urban blending problem, one potential approach would be to avoid using reference stations for calculating the value of non-climatic biases associated with breakpoints, although they could still be used for identifying the breakpoints. Alternatively, Soon et al. (2018) offered another approach to reduce the problem. They suggested the urban blending problem should be substantially reduced if the reference stations used for homogenizing the data have a similar degree of urbanization to the target stations. They noted that studies that had explicitly developed a rural reference network before homogenizing may have indirectly reduced urban blending problems (e.g., Karl et al. 1988; Ren et al. 2015; Ren and Ren 2011; Shi et al. 2015). However, given that the longest and most-complete station records tend to be relatively urbanized, researchers may see some value in homogenizing urban records using similarly urbanized reference neighbors. The homogenization process could then reduce the non-urban-related biases with minimal urban blending. Correcting for urbanization bias could then be carried out *after* the homogenization process.

Soon et al. (2018) cautioned that the blending problem could also occur for other non-climatic biases that have similarly affected a large number of stations in a network, e.g., the siting biases identified by Fall et al. (2011) for the U.S. temperature data. Therefore, efforts to homogenize temperature datasets should ideally also consider the blending problems of other non-climatic biases. This could include expanding the site inspections of Fall et al. (2011) to establish if other regions, including Japan, are similarly affected.

File generated with AMS Word template 2.0

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

Other efforts could include the collection and digitization of station history metadata from the station observers documenting any known non-climatic changes, e.g., as O'Neill et al. (2022) has been doing for Europe. Aside from generally helping to better identify potential non-climatic breakpoints, it could also reveal cases where systemic non-climatic biases occurred simultaneously or near-simultaneously across large regions, e.g., the documented shifts in observation times over the 20th century associated with the USHCN and COOP stations (Karl et al. 1986). We note that some station history metadata for Japanese stations is available online at https://www.data.jma.go.jp/obd/stats/data/mdrr/chiten/meta/discnt_sfc.csv [last accessed May 2023].

In the meantime, we advise users of temperature datasets to be wary of assuming that homogenized temperature records are automatically more reliable. It is true that unhomogenized temperature records are often plagued by non-climatic biases and that homogenization can often reduce these biases. However, due to urban blending, the homogenization process also inadvertently introduces many fresh non-climatic biases of its own. Users of both unhomogenized *and* homogenized temperature records should be very cautious about the problems of non-climatic biases.

Finally, with regards to the implications of this analysis for Japan, Japan is a highly urbanized country with a population density of 347 people/km^2 (~7 times the world average of 52 people/km^2 and ~10 times the U.S. average of 36 people/km^2) and 91.8% of the population is urban in 2020 (https://www.worldometers.info/world-population/japan-population/, last accessed on 1 January, 2023). Therefore, one could argue that using urbanized stations to describe the climate of Japan is not a "bias", since most of the population experiences an urban climate. However, the "urbanization bias" occurs when this localized urban warming of Japan is mistakenly assumed to be part of *global* warming. In the case of Japan, urban warming appears to have dominated temperature trends over the last century. Therefore, efforts to reduce future warming *in Japan* probably should prioritize urban heat island mitigation (Enteria et al. 2021), rather than focusing almost exclusively on reducing greenhouse gas emissions as currently appears to be the case (Sugiyama et al. 2019).

Acknowledgments.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

We thank Michael Connolly, Willie Soon, and Taishi Sugiyama for helpful comments and suggestions. We would like to thank the three anonymous reviewers for their feedback that substantially improved our manuscript.

Data Availability Statement.

The GHCN version 3 and 4 datasets used for our analysis were provided by National Oceanic and Atmospheric Administration (NOAA) and are available at https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-monthly. The USHCN dataset was also provided by NOAA and available from their ftp website at: ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/v2.5. They also provide the data we used for our reanalysis of the Hausfather et al. (2013) results on aliasing at: ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/papers/hausfather-etal2013-suppinfo/.

REFERENCES

Aono, Y., and K. Kazui, 2008: Phenological data series of cherry tree flowering in Kyoto, Japan, and its application to reconstruction of springtime temperatures since the 9th century. *International Journal of Climatology*, **28**, 905–914, https://doi.org/10.1002/joc.1594.

——, and S. Saito, 2010: Clarifying springtime temperature reconstructions of the medieval period by gap-filling the cherry blossom phenological data series at Kyoto, Japan. *Int J Biometeorol*, **54**, 211–219, https://doi.org/10.1007/s00484-009-0272-x.

Balling, R. C., and C. D. Idso, 2002: Analysis of adjustments to the United States Historical Climatology Network (USHCN) temperature database. *Geophysical Research Letters*, **29**, 25-1-25–3, https://doi.org/10.1029/2002GL014825.

Christidis, N., Y. Aono, and P. A. Stott, 2022: Human influence increases the likelihood of extremely early cherry tree flowering in Kyoto. *Environ. Res. Lett.*, **17**, 054051, https://doi.org/10.1088/1748-9326/ac6bb4.

CIESIN, (Center for International Earth Science Information Network, Columbia University), 2018: Gridded Population of the World, Version 4 (GPWv4): Population Density Adjusted to Match 2015 Revision UN WPP Country Totals, Revision 11. Palisades, New York: NASA Socioeconomic Data and Applications Center (SEDAC); gpw_v4_population_density_adjusted_to_2015_unwpp_country_totals_rev11_2000_30_sec.tif.

File generated with AMS Word template 2.0

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

Connolly, R., and M. Connolly, 2014a: Has poor station quality biased U.S. temperature trend estimates? *Open Peer Review Journal*, http://oprj.net/articles/climate-science/11.

——, and ——, 2014b: Urbanization bias I. Is it a negligible problem for global temperature estimates? *Open Peer Review Journal*, http://oprj.net/articles/climate-science/28.

——, and ——, 2014c: Urbanization bias II. An assessment of the NASA GISS urbanization adjustment method. *Open Peer Review Journal*, http://oprj.net/articles/climate-science/31.

——, and ——, 2014d: Urbanization bias III. Estimating the extent of bias in the Historical Climatology Network datasets. *Open Peer Review Journal*, http://oprj.net/articles/climate-science/34.

——, and Coauthors, 2021: How much has the Sun influenced Northern Hemisphere temperature trends? An ongoing debate. *Res. Astron. Astrophys.*, **21**, 131, https://doi.org/10.1088/1674-4527/21/6/131.

Das, L., J. D. Annan, J. C. Hargreaves, and S. Emori, 2011: Centennial scale warming over Japan: are the rural stations really rural? *Atmospheric Science Letters*, **12**, 362–367, https://doi.org/10.1002/asl.350.

DeGaetano, A. T., 2006: Attributes of Several Methods for Detecting Discontinuities in Mean Temperature Series. *Journal of Climate*, **19**, 838–853, https://doi.org/10.1175/JCLI3662.1.

Domonkos, P., 2011: Efficiency evaluation for detecting inhomogeneities by objective homogenisation methods. *Theor Appl Climatol*, **105**, 455–467, https://doi.org/10.1007/s00704-011-0399-7.

——, 2021: Combination of Using Pairwise Comparisons and Composite Reference Series: A New Approach in the Homogenization of Climatic Time Series with ACMANT. *Atmosphere*, **12**, 1134, https://doi.org/10.3390/atmos12091134.

Easterling, D. R., and T. C. Peterson, 1995: A new method for detecting undocumented discontinuities in climatological time series. *International Journal of Climatology*, **15**, 369–377, https://doi.org/10.1002/joc.3370150403.

Efthymiadis, D. A., and P. D. Jones, 2010: Assessment of Maximum Possible Urbanization Influences on Land Temperature Data by Comparison of Land and Marine Data around Coasts. *Atmosphere*, **1**, 51–61, https://doi.org/10.3390/atmos1010051.

Enteria, N., M. Santamouris, and U. Eicker, eds., 2021: *Urban Heat Island (UHI) Mitigation*. Springer Singapore, 307 pp.

Fall, S., A. Watts, J. Nielsen-Gammon, E. Jones, D. Niyogi, J. R. Christy, and R. A. Pielke, 2011: Analysis of the impacts of station exposure on the U.S. Historical Climatology Network temperatures and temperature trends. *Journal of Geophysical Research: Atmospheres*, **116**, https://doi.org/10.1029/2010JD015146.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

Fujibe, F., 2009: Detection of urban warming in recent temperature trends in Japan. *International Journal of Climatology*, **29**, 1811–1822, https://doi.org/10.1002/joc.1822.

——, 2011: Urban warming in Japanese cities and its relation to climate change monitoring. *International Journal of Climatology*, **31**, 162–173, https://doi.org/10.1002/joc.2142.

——, 2012: Evaluation of background and urban warming trends based on centennial temperature data in Japan. *Pap. Met. Geophys.*, **63**, 43–56, https://doi.org/10.2467/mripapers.63.43.

——, and K. Ishihara, 2010: Possible Urban Bias in Gridded Climate Temperature Data over the Japan Area. *Sola*, **6**, 61–64, https://doi.org/10.2151/sola.2010-016.

Fukui, E., 1957: Increasing Temperature due to the Expansion of Urban Areas in Japan. *Journal of the Meteorological Society of Japan. Ser. II*, **35A**, 336–341, https://doi.org/10.2151/jmsj1923.35A.0_336.

Gaffin, S. R., and Coauthors, 2008: Variations in New York city's urban heat island strength over time and space. *Theor Appl Climatol*, **94**, 1–11, https://doi.org/10.1007/s00704-007-0368-3.

Gillett, N. P., and Coauthors, 2021: Constraining human contributions to observed warming since the pre-industrial period. *Nat. Clim. Chang.*, **11**, 207–212, https://doi.org/10.1038/s41558-020-00965-9.

Hansen, J., R. Ruedy, M. Sato, M. Imhoff, W. Lawrence, D. Easterling, T. Peterson, and T. Karl, 2001: A closer look at United States and global surface temperature change. *Journal of Geophysical Research: Atmospheres*, **106**, 23947–23963, https://doi.org/10.1029/2001JD000354.

——, ——, ——, and K. Lo, 2010: Global Surface Temperature Change. *Reviews of Geophysics*, **48**, https://doi.org/10.1029/2010RG000345.

Hausfather, Z., M. J. Menne, C. N. Williams, T. Masters, R. Broberg, and D. Jones, 2013: Quantifying the effect of urbanization on U.S. Historical Climatology Network temperature records. *Journal of Geophysical Research: Atmospheres*, **118**, 481–494, https://doi.org/10.1029/2012JD018509.

He, Y.-T., and G.-S. Jia, 2012: A Dynamic Method for Quantifying Natural Warming in Urban Areas. *Atmospheric and Oceanic Science Letters*, **5**, 408–413, https://doi.org/10.1080/16742834.2012.11447029.

Hubbard, K. G., and X. Lin, 2006: Reexamination of instrument change effects in the U.S. Historical Climatology Network. *Geophysical Research Letters*, **33**, https://doi.org/10.1029/2006GL027069.

IPCC, 2021: *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press,.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

JMA, (Japan Meteorological Agency), 2022: Annual average temperature anomaly in Japan over time (1898-2021). https://www.data.jma.go.jp/cpdinfo/temp/an_jpn.html (Accessed May 25, 2023).

Jones, P. D., D. H. Lister, and Q. Li, 2008: Urbanization effects in large-scale temperature records, with an emphasis on China. *Journal of Geophysical Research: Atmospheres*, **113**, https://doi.org/10.1029/2008JD009916.

Karl, T. R., and C. N. Williams, 1987: An Approach to Adjusting Climatological Time Series for Discontinuous Inhomogeneities. *Journal of Applied Meteorology and Climatology*, **26**, 1744–1763, https://doi.org/10.1175/1520-0450(1987)026<1744:AATACT>2.0.CO;2.

——, ——, P. J. Young, and W. M. Wendland, 1986: A Model to Estimate the Time of Observation Bias Associated with Monthly Mean Maximum, Minimum and Mean Temperatures for the United States. *J. Climate Appl. Meteor.*, **25**, 145–160, https://doi.org/10.1175/1520-0450(1986)025<0145:AMTETT>2.0.CO;2.

——, H. F. Diaz, and G. Kukla, 1988: Urbanization: Its Detection and Effect in the United States Climate Record. *Journal of Climate*, **1**, 1099–1123, https://doi.org/10.1175/1520-0442(1988)001<1099:UIDAEI>2.0.CO;2.

Kent, E. C., and J. J. Kennedy, 2021: Historical Estimates of Surface Marine Temperatures. *Annual Review of Marine Science*, **13**, 283–311, https://doi.org/10.1146/annurev-marine-042120-111807.

——, and Coauthors, 2017: A Call for New Approaches to Quantifying Biases in Observations of Sea Surface Temperature. *Bulletin of the American Meteorological Society*, **98**, 1601–1616, https://doi.org/10.1175/BAMS-D-15-00251.1.

Lawrimore, J. H., M. J. Menne, B. E. Gleason, C. N. Williams, D. B. Wuertz, R. S. Vose, and J. Rennie, 2011: An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3. *Journal of Geophysical Research: Atmospheres*, **116**, https://doi.org/10.1029/2011JD016187.

Lenssen, N. J. L., G. A. Schmidt, J. E. Hansen, M. J. Menne, A. Persin, R. Ruedy, and D. Zyss, 2019: Improvements in the GISTEMP Uncertainty Model. *Journal of Geophysical Research: Atmospheres*, **124**, 6307–6326, https://doi.org/10.1029/2018JD029522.

Li, Q., and Y. Yang, 2019: Comments on "Comparing the current and early 20th century warm periods in China" by Soon W., R. Connolly, M. Connolly et al. *Earth-Science Reviews*, **198**, 102886, https://doi.org/10.1016/j.earscirev.2019.102886.

Li, Q., H. Zhang, X. Liu, and J. Huang, 2004: Urban heat island effect on annual mean temperature during the last 50 years in China. *Theor Appl Climatol*, **79**, 165–174, https://doi.org/10.1007/s00704-004-0065-4.

Matsumoto, J., F. Fujibe, and H. Takahashi, 2017: Urban climate in the Tokyo metropolitan area in Japan. *Journal of Environmental Sciences*, **59**, 54–62, https://doi.org/10.1016/j.jes.2017.04.012.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

McKitrick, R., and J. Christy, 2020: Pervasive Warming Bias in CMIP6 Tropospheric Layers. *Earth and Space Science*, **7**, e2020EA001281, https://doi.org/10.1029/2020EA001281.

Menne, M. J., and C. N. Williams, 2009: Homogenization of Temperature Series via Pairwise Comparisons. *J. Climate*, **22**, 1700–1717, https://doi.org/10.1175/2008JCLI2263.1.

——, ——, and R. S. Vose, 2009: The U.S. Historical Climatology Network Monthly Temperature Data, Version 2. *Bulletin of the American Meteorological Society*, **90**, 993–1008, https://doi.org/10.1175/2008BAMS2613.1.

——, ——, and M. A. Palecki, 2010: On the reliability of the U.S. surface temperature record. *Journal of Geophysical Research: Atmospheres*, **115**, https://doi.org/10.1029/2009JD013094.

——, ——, B. E. Gleason, J. J. Rennie, and J. H. Lawrimore, 2018: The Global Historical Climatology Network Monthly Temperature Dataset, Version 4. *J. Climate*, **31**, 9835–9854, https://doi.org/10.1175/JCLI-D-18-0094.1.

Mestre, O., and Coauthors, 2013: HOMER: A homogenization software - methods and applications. *Idojaras*, **117**.

Mitchell, J. M., Jr., 1953: On the Causes of Instrumentally Observed Secular Temperature Trends. *Journal of Atmospheric Sciences*, **10**, 244–261, https://doi.org/10.1175/1520-0469(1953)010<0244:OTCOIO>2.0.CO;2.

NOAA, (National Oceanic and Atmospheric Administration), 2015: Image and Data processing by NOAA's National Geophysical Data Center. DMSP data collected by the US Air Force Weather Agency. https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html, F182013.v4c_web.stable_lights.avg_vis.tif from https://ngdc.noaa.gov/eog/data/web_data/v4composites/F182013.v4.tar. https://ngdc.noaa.gov/eog/data/web_data/v4composites/F182013.v4.tar (Accessed August 16, 2017).

Nordli, P. Ø., H. Alexandersson, P. Frich, E. J. Førland, R. Heino, T. Jónsson, H. Tuomenvirta, and O. E. Tveito, 1997: The effect of radiation screens on Nordic time series of mean temperature. *International Journal of Climatology*, **17**, 1667–1681, https://doi.org/10.1002/(SICI)1097-0088(199712)17:15<1667::AID-JOC221>3.0.CO;2-D.

Oke, T. R., 1973: City size and the urban heat island. *Atmospheric Environment (1967)*, **7**, 769–779, https://doi.org/10.1016/0004-6981(73)90140-6.

O'Neill, P., and Coauthors, 2022: Evaluation of the Homogenization Adjustments Applied to European Temperature Records in the Global Historical Climatology Network Dataset. *Atmosphere*, **13**, 285, https://doi.org/10.3390/atmos13020285.

Osborn, T. J., P. D. Jones, D. H. Lister, C. P. Morice, I. R. Simpson, J. P. Winn, E. Hogan, and I. C. Harris, 2021: Land Surface Air Temperature Variations Across the Globe

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

Updated to 2019: The CRUTEM5 Data Set. *Journal of Geophysical Research: Atmospheres*, **126**, e2019JD032352, https://doi.org/10.1029/2019JD032352.

Parker, D. E., 2006: A Demonstration That Large-Scale Warming Is Not Urban. *J. Climate*, **19**, 2882–2895, https://doi.org/10.1175/JCLI3730.1.

Peterson, T. C., K. P. Gallo, J. Lawrimore, T. W. Owen, A. Huang, and D. A. McKittrick, 1999: Global rural temperature trends. *Geophysical Research Letters*, **26**, 329–332, https://doi.org/10.1029/1998GL900322.

Pielke, R., and Coauthors, 2007: Documentation of Uncertainties and Biases Associated with Surface Temperature Measurement Sites for Climate Change Assessment. *Bulletin of the American Meteorological Society*, **88**, 913–928, https://doi.org/10.1175/BAMS-88-6-913.

Quayle, R. G., D. R. Easterline, T. R. Karl, and P. Y. Hughes, 1991: Effects of Recent Thermometer Changes in the Cooperative Station Network. *Bulletin of the American Meteorological Society*, **72**, 1718–1724, https://doi.org/10.1175/1520-0477(1991)072<1718:EORTCI>2.0.CO;2.

Reeves, J., J. Chen, X. L. Wang, R. Lund, and Q. Q. Lu, 2007: A Review and Comparison of Changepoint Detection Techniques for Climate Data. *Journal of Applied Meteorology and Climatology*, **46**, 900–915, https://doi.org/10.1175/JAM2493.1.

Ren, G., and Coauthors, 2015: An Integrated Procedure to Determine a Reference Station Network for Evaluating and Adjusting Urban Bias in Surface Air Temperature Data. *Journal of Applied Meteorology and Climatology*, **54**, 1248–1266, https://doi.org/10.1175/JAMC-D-14-0295.1.

Ren, Y., and G. Ren, 2011: A Remote-Sensing Method of Selecting Reference Stations for Evaluating Urbanization Effect on Surface Air Temperature Trends. *J. Climate*, **24**, 3179–3189, https://doi.org/10.1175/2010JCLI3658.1.

Scafetta, N., 2021: Detection of non- climatic biases in land surface temperature records by comparing climatic data and their model simulations. *Clim Dyn*, **56**, 2959–2982, https://doi.org/10.1007/s00382-021-05626-x.

Shi, T., Y. Huang, H. Wang, C.-E. Shi, and Y.-J. Yang, 2015: Influence of urbanization on the thermal environment of meteorological station: Satellite-observed evidence. *Advances in Climate Change Research*, **6**, 7–15, https://doi.org/10.1016/j.accre.2015.07.001.

Shi, Z., G. Jia, Y. Hu, and Y. Zhou, 2019: The contribution of intensified urbanization effects on surface warming trends in China. *Theor Appl Climatol*, **138**, 1125–1137, https://doi.org/10.1007/s00704-019-02892-y.

Soon, W., R. Connolly, and M. Connolly, 2015: Re-evaluating the role of solar variability on Northern Hemisphere temperature trends since the 19th century. *Earth-Science Reviews*, **150**, 409–452, https://doi.org/10.1016/j.earscirev.2015.08.010.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

Soon, W. W.-H., R. Connolly, M. Connolly, P. O'Neill, J. Zheng, Q. Ge, Z. Hao, and H. Yan, 2018: Comparing the current and early 20th century warm periods in China. *Earth-Science Reviews*, **185**, 80–101, https://doi.org/10.1016/j.earscirev.2018.05.013.

——, ——, ——, ——, ——, ——, ——, and ——, 2019: Reply to Li & Yang's comments on "Comparing the current and early 20th century warm periods in China." *Earth-Science Reviews*, **198**, 102950, https://doi.org/10.1016/j.earscirev.2019.102950.

Squintu, A. A., G. van der Schrier, P. Štěpánek, P. Zahradníček, and A. K. Tank, 2020: Comparison of homogenization methods for daily temperature series against an observation-based benchmark dataset. *Theor Appl Climatol*, **140**, 285–301, https://doi.org/10.1007/s00704-019-03018-0.

Stewart, I. D., 2011: A systematic review and scientific critique of methodology in modern urban heat island literature. *International Journal of Climatology*, **31**, 200–217, https://doi.org/10.1002/joc.2141.

Stewart, I. D., 2019: Why should urban heat island researchers study history? *Urban Climate*, **30**, 100484, https://doi.org/10.1016/j.uclim.2019.100484.

Stewart, I. D., and T. R. Oke, 2012: Local Climate Zones for Urban Temperature Studies. *Bulletin of the American Meteorological Society*, **93**, 1879–1900, https://doi.org/10.1175/BAMS-D-11-00019.1.

Sugiyama, M., and Coauthors, 2019: Japan's long-term climate mitigation policy: Multi-model assessment and sectoral challenges. *Energy*, **167**, 1120–1131, https://doi.org/10.1016/j.energy.2018.10.091.

Sun, W., Y. Yang, L. Chao, W. Dong, B. Huang, P. Jones, and Q. Li, 2022: Description of the China global Merged Surface Temperature version 2.0. *Earth Syst. Sci. Data*, **14**, 1677–1693, https://doi.org/10.5194/essd-14-1677-2022.

Sun, Y., X. Zhang, G. Ren, F. W. Zwiers, and T. Hu, 2016: Contribution of urbanization to warming in China. *Nature Climate Change*, **6**, 706–709, https://doi.org/10.1038/nclimate2956.

——, T. Hu, X. Zhang, C. Li, C. Lu, G. Ren, and Z. Jiang, 2019: Contribution of Global warming and Urbanization to Changes in Temperature Extremes in Eastern China. *Geophysical Research Letters*, **46**, 11426–11434, https://doi.org/10.1029/2019GL084281.

Venema, V. K. C., and Coauthors, 2012: Benchmarking homogenization algorithms for monthly data. *Climate of the Past*, **8**, 89–115, https://doi.org/10.5194/cp-8-89-2012.

Vose, R. S., C. N. Williams, T. C. Peterson, T. R. Karl, and D. R. Easterling, 2003: An evaluation of the time of observation bias adjustment in the U.S. Historical Climatology Network. *Geophysical Research Letters*, **30**, https://doi.org/10.1029/2003GL018111.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.

Vose, R. S., and Coauthors, 2021: Implementing Full Spatial Coverage in NOAA's Global Temperature Analysis. *Geophysical Research Letters*, **48**, e2020GL090873, https://doi.org/10.1029/2020GL090873.

Wickham, C., and Coauthors, 2013: Influence of Urban Heating on the Global Temperature Land Average using Rural Sites Identified from MODIS Classifications. *Geoinformatics & Geostatistics: An Overview*, **2013**, https://doi.org/10.4172/2327-4581.1000104.

Williams, C. N., M. J. Menne, and P. W. Thorne, 2012: Benchmarking the performance of pairwise homogenization of surface temperatures in the United States. *Journal of Geophysical Research: Atmospheres*, **117**, https://doi.org/10.1029/2011JD016761.

Yamashita, S., 1996: Detailed structure of heat island phenomena from moving observations from electric tram-cars in Metropolitan Tokyo. *Atmospheric Environment*, **30**, 429–435, https://doi.org/10.1016/1352-2310(95)00010-0.

Zhang, P., and Coauthors, 2021: Urbanization Effects on Estimates of Global Trends in Mean and Extreme Air Temperature. *Journal of Climate*, **34**, 1923–1945, https://doi.org/10.1175/JCLI-D-20-0389.1.

Zou, C.-Z., H. Xu, X. Hao, and Q. Liu, 2023: Mid-Tropospheric Layer Temperature Record Derived From Satellite Microwave Sounder Observations With Backward Merging Approach. *Journal of Geophysical Research: Atmospheres*, **128**, e2022JD037472, https://doi.org/10.1029/2022JD037472.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-22-0122.1.