Vol. 91: 47–66, 2023
https://doi.org/10.3354/cr01724

CLIMATE RESEARCH
Clim Res

Published September 14

Contribution to CR Special 36 'Observed changes in climate over Asia'

Updated analysis of surface warming trends in North China based on in-depth homogenized data (1951–2020)

Jiajun He[1], Guoyu Ren[1,2,*], Panfeng Zhang[3], Xiang Zheng[1], Siqi Zhang[1,2]

[1]Department of Atmospheric Science, School of Environmental Studies, China University of Geosciences (CUG), Wuhan 430074, PR China

[2]Laboratory for Climate Studies, National Climate Center, China Meteorological Administration (CMA), Beijing 100081, PR China

[3]School of Tourism and Geographical Sciences, Jilin Normal University, Siping 136000, PR China

ABSTRACT: The reliability of climate change detection and research is significantly impacted by the inhomogeneity of surface climate observation data. However, there is an ongoing debate regarding whether comprehensive homogenization has been performed in large-scale homogenized data sets. In this study, we examined the homogeneity of the original maximum and minimum temperature (T_{max} and T_{min}) data for 662 meteorological stations in North China by using multiple methods and combining with metadata. The quantile matching method was employed to adjust the daily T_{max} and T_{min} series. In order to avoid the potential systematic bias resulting from homogenization, no reference series were introduced during the adjustment process. The adjustment results indicate that T_{min} in North China is significantly affected by non-climatic factors, particularly station relocations and environmental changes around the stations. The application of homogenization in this study led to a notable increase in the overall temperature trends of the stations, with T_{min} exhibiting a larger increase and the diurnal temperature range demonstrating a more significant downward trend. Based on the homogenized data, the annual and seasonal mean temperature trends in North China from 1951 to 2020 were re-evaluated. These temperature trends generally surpass those reported in previous research for the same period from 1961 to 2000. The higher estimate of temperature trends may be attributed to the recovered urbanization effect in the newly homogenized data. Thus, the obtained homogenization data still exhibit a significant urbanization bias that requires further assessment and adjustment.

KEY WORDS: Homogenization · Data · Maximum temperature · Minimum temperature · Warming trend · Urbanization effect · North China

1. INTRODUCTION

The global average surface temperature has undergone a rapid increase over the past century, leading to an escalating risk of climate change. To evaluate the variability and trends in global warming, several global-scale studies have been conducted (Jones et al. 2012, Cubasch et al. 2013, Sun et al. 2017). Surface air temperatures in China have also increased significantly (NARCC Compilation Committee 2015). Numerous studies have assessed the extent of ground surface air temperature warming in China, resulting in the development of several nationwide temperature data sets (Wang 1990, Song 1994, Wang et al. 1998, Ren et al. 2005, Tang & Ren 2005, Ding & Wang 2016, Soon et al. 2018). Ensuring as much homogeneity as possible is essential for long-term climate analyses, particularly in the context of surface air temperature changes (Aguilar et al. 2003). However, climate data sets often exhibit in-

*Corresponding author: guoyoo@cma.cn

Publisher: Inter-Research · www.int-res.com

homogeneities that introduce uncertainties into research conclusions. Consequently, accurately testing and adjusting the breakpoints in observation data series, known as 'homogenization,' is crucial for detecting and attributing climate change.

Homogenization research can be dated back to the end of the last century (Domonkos et al. 2022). Early development of homogenization theories and corresponding methods were pioneered by Alexandersson (1986), Easterling & Peterson (1995), and Vincent (1998). Peterson et al. (1998) comprehensively discussed the early progress of homogenization work. Aguilar (2003) evaluated the principles and characteristics of different methods comparatively, and assessed their strong and weak points. Since then, numerous homogenization methods have been developed and widely applied. Statistical theories have led to the development of techniques that can more accurately detect breakpoints for annual and monthly data (Li et al. 2004, Kuglitsch et al. 2012, Vincent et al. 2012, Wang & Feng 2013, Mamara et al. 2014). These approaches are validated with station historical data, also known as metadata, which include documentation of changes in observation practices, thermometer replacements, environmental changes, and most importantly, station relocations. However, incorporating metadata into the homogenization process of global temperature data sets has proved to be a challenging task. As a result, most global temperature data sets have not yet adopted the use of metadata for homogenization (Menne et al. 2018, Lenssen et al. 2019, Rohde & Hausfather 2020). With the advancements in homogenization theories, methods for homogenizing daily data have been supplemented and improved. Trewin (2013) developed a set of homogenized daily temperature data in Australia using the percentile matching (PM) method. Additionally, Vincent et al. (2012) employed the quantile matching (QM) method to adjust the monthly and daily temperature data in Canada. In the last decades, several methods for homogenization have also been introduced for specific purposes (Menne & Williams 2009, Domonkos 2011a, Mestre et al. 2013, Guijarro 2018, Squintu et al. 2019), some of which were adopted in this work and will be introduced later in this paper.

It is worth noting that the underlying mathematical and statistical theories of different methods vary considerably; in particular, their homogenization procedures may be quite dissimilar. Therefore, the homogenization outputs of the original data (containing non-climatic biases) could be different depending on what method is used to perform the homogenization (Venema et al. 2012). In some cases, the outputs might differ significantly from previous ones during the iteration process (O'Neill et al. 2022). In recent years, researchers have taken into account the efficiency and capability of different methods to perform benchmarking aimed at achieving more accurate and reasonable homogenization (Menne & Williams 2005, Venema et al. 2012, Domonkos 2013, Lindau & Venema 2016, Vincent et al. 2018, Domonkos et al. 2021). Moreover, recent studies benchmarked techniques by creating test series with real-world data (Gubler et al. 2017, Vincent et al. 2018, Squintu et al. 2020), given that the simulated data are not able to fully simulate the various, complex, and imperceptible climatic variations that occur in the natural environment, and are only able to test signals that are easier to recognize.

Over the past 2 decades, several national and regional homogenized temperature data sets have been established in China. One such data set is the China Homogenized Historical Temperature data set developed by Li et al. (2004) by homogenizing the daily and monthly temperature series of 671 national meteorological stations from 1951 to 2001. Similarly, Li & Yan (2009) employed the Multiple Analysis of Series for Homogenization software package to homogenize temperature series of 549 national stations in China from 1960 to 2008, resulting in a set of large-scale homogenized data. Using the penalized maximal *t*-test (PMT) and with reference to the metadata, Xu et al. (2013) homogenized the daily temperature data of 825 stations in China and further analyzed trends in extreme temperature indices. The most recent nationwide homogenization achievement is the 'China National Surface Meteorological Station Homogenization Temperature Daily Dataset (V1.0),' as developed by Cao et al. (2016) using the QM method, including both daily and monthly homogenized temperature data since 1951 for mainland China.

Given the influence of inhomogeneity factors, prevailing approaches of homogenization mainly focus on the relocation effects at a single station or within a restricted area (Yan et al. 2001, Morozova & Valente 2012, Zhang et al. 2014, Kolendowicz et al. 2019, Si et al. 2021). Distant movements of stations recorded by metadata play an essential role in identifying inhomogeneities. However, when executing large-scale homogenization projects, incomplete homogenization may occur as a result of disregarding possible impacts from environmental changes in surrounding regions and minor fluctuations in instrument positions; this is due to the significant number of samples involved and the difficulty in obtaining comprehensive metadata. He et al. (2021) suggested that significant inhomogeneity remains in existing homogenized data sets, showing greater trend changes in rural station series, which requires further examination and adjustment. In addition, differing homogenization methods utilized by various studies can lead to differences in trend estimates based on the data sets. Therefore, extensive in-depth examination and adjustment of inhomogeneity of observation data is crucial prior to regional climate change research.

The present study tests and adjusts the original temperature data in North China using a combination of several methods, and the long-term change in North China is estimated based on the newly homogenized data set. Section 2 provides details on the data sources used, as well as the specific homogenization techniques employed. Section 3 presents statistical analyses of the homogenized data, and then calculates the annual and seasonal surface warming in North China from 1951 to 2020. Section 4 discusses the uncertainties of this research, and Section 5 summarizes the findings of the study.

2. DATA AND METHODS

2.1. Data sources and research region

The 'China National Surface Meteorological Station Fundamental Meteorological Elements Daily Dataset (V3.0)' (Ren et al. 2012), developed by the National Meteorological Information Center (NMIC) of the China Meteorological Administration (CMA), was employed for the purpose of conducting homogenization. This data set encompasses the daily observation data of fundamental meteorological elements, including surface air temperature, collected from 2479 national stations across mainland China. Notably, the temperature data were not subjected to any non-climatic bias removal process or homogenization. The record series start in January 1951 at the earliest, and span a maximum of 70 yr (1951–2020). Rigorous quality control measures were implemented to ensure the overall integrity and quality of the data. In this study, the daily maximum (T_{max}) and minimum temperature (T_{min}) data were used for homogenization and analysis.

In this paper, the term 'North China' refers to the region between 108–120° E and 33–43° N, in which stations are densely distributed (Fig. 1). This region encompasses multiple provinces, regions, and cities in the central northern part of the country. Our study focuses on 662 national meteorological stations located in this area. For comparison purposes, we utilized the 'China National Surface Meteorological Station Homogenization Temperature Daily Dataset (V1.0)' developed by Cao et al. (2016), which also originates from the NMIC data set. To obtain the necessary metadata, we accessed the historical data of national surface meteorological stations from the Meteorological Data Office of the NMIC. These records document even minor changes in coordinates and altitudes resulting from station movements and environmental shifts.

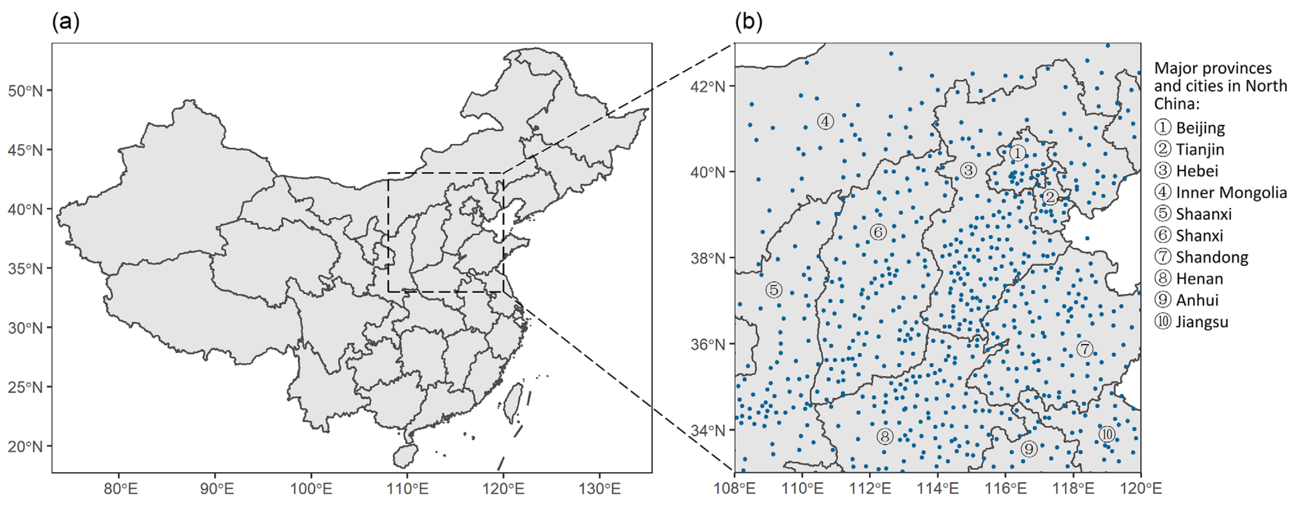

Fig. 1. Distribution of stations in North China. (a) Study area in North China (dashed square); (b) stations (n = 662) analyzed in this study (blue dots)

2.2. Homogenization methods

Inhomogeneity detection for climate data predominantly relies on statistical algorithms, and the most effective approach to test the inhomogeneity of a series is to verify and calibrate the detected breakpoints in conjunction with metadata whenever available (Aguilar et al. 2003). Unfortunately, in many cases, metadata are either missing, incomplete, or inaccurate. Metadata serve the primary purpose of confirming the significance and reliability of a statistically detected breakpoint, as well as calibrating its occurrence time. However, since metadata do not always correspond to the breakpoints individually, they cannot be relied upon as the primary measure for inhomogeneity detection. Additionally, it might not be advisable to rely solely on a single method for homogenizing candidate data owning to the inconsistency of the outputs over the iterative process of a particular method (O'Neill et al. 2022). Hence, the reliability and accuracy of statistical methods play a much more prominent role in data homogenization (Trewin et al. 2020).

There are 3 problems that statistical methods must solve in terms of homogenization:

(1) In many cases, there may be more than one breakpoint in the time series for the test (Lindau & Venema 2013).

(2) The reference series around the target series may also contain breakpoints (Lindau & Venema 2013).

(3) 'Urban blending' indicates the systematic tendency of aliasing a portion of the trend biases of reference series onto the estimated magnitude of the breakpoints when adjusting the target series (DeGaetano 2006, Pielke et al. 2007, Soon et al. 2018).

The first 2 issues pertain to the detection of inhomogeneity. Regarding the first aspect, there are 2 approaches to address the presence of multiple breakpoints within a series. One approach involves splitting the series into 2 segments at the most significant breakpoint and then applying inhomogeneity tests to each subseries. This process is repeated until no more breakpoints can be identified in any subseries. Alternatively, multiple breakpoint detection methods can be directly employed, which have been demonstrated to effectively detect multiple breakpoints simultaneously (Domonkos 2011b, Venema et al. 2012). As for the second aspect, the following measures can be employed to obtain the reference series for inhomogeneity detection: averaging multiple reference series into a composite reference series, which helps eliminate minor breakpoints but may not effectively remove large, conspicuous jumps. In this scenario, homogeneous reference series should be used as far as possible. If working with inhomogeneous reference series becomes necessary, adjustments should be made prior to their utilization, even though this pre-homogenization process may introduce certain uncertainties.

In addition, the series can be matched in pairs for detection to avoid the impacts of reference series inhomogeneity. It is important to note that reference series may not exhibit homogeneity in a paired manner. The breakpoints in 2 series can be identified by analyzing the difference series between them. A subsequent 'attribution' step determines which breakpoints detected in the difference series belong to which individual series. To effectively address the issue of reference series inhomogeneity, a recommended approach is to perform joint detection for all series within a network simultaneously. This necessitates the use of automatic methods that homogenize large data sets in an objective computerized process.

Given that no method is capable of handling these problems concurrently in a perfect way (Venema et al. 2012), and considering the need to mitigate the uncertainties associated with different methods, a composite manner was adopted for inhomogeneity detection in our study. This approach involved combining the results obtained from multiple methods: the adapted Caussinus-Mestre algorithm for networks of temperature series (ACMANT) (Domonkos 2011a), the pairwise homogenization algorithm (PHA) (Menne & Williams 2009), and the RHtest (Wang & Feng 2013). Each of these methods possesses distinct techniques that effectively address specific types of inhomogeneity during the detection process. By subjecting the data set to tests conducted by all 3 methods, we aimed to achieve improved accuracy and reduced uncertainties in identifying potential inhomogeneities. In the following text, we will introduce how each method was employed and elucidate the rationale behind their implementation in this study.

As mentioned earlier, a data set formed as a network, like the North China temperature data analyzed in this study, could be homogenized through completely automatic methods. The perturbation of inhomogeneity within reference series can also be diminished by applying automatic software. From this point, the ACMANT software, developed during the HOME Cost Action and proven to be as reliable as manual methods (Venema et al. 2012), was considered applicable for the detection of the inhomogeneities (version 4.3 was used in this study). Based on optimal step function fitting (Hawkins 1972), ACMANT tests and adjusts the target series according to the Caussinus Lyazhri criterion (Caussinus & Lyazrhi 1997) with multiple weighted reference series. ACMANT requires at least 4 reference series for each target series, and each series in the network can serve as a reference to the others. The output of ACMANT comes from the integration of the highest correlated detection results. In our study, all of the temperature series in North China were constructed as an input network, and the parameters such as starting and ending years, data length, and other options were set to launch the inhomogeneity detection.

For the second approach of tackling inhomogeneous reference series, the PHA method matches series in pairs and calculates the difference between the target series and each reference series, then tests the difference series to determine whether the breakpoints belong to the target series. This method allows for a greater tolerance towards the inhomogeneity within the reference series. However, it remains imperative to ensure a strong correlation between the target series and the selected reference series. To achieve this, the correlation coefficients were computed after constructing first differences for both the target and reference series. Only those reference series exhibiting a correlation coefficient greater than 0.9 with the target series were considered for the detection process. In order to enhance detection efficiency, each target series was tested against 3 to 7 reference series. A breakpoint was deemed significant only if it was identified in at least 3 difference series. Throughout this procedure, the widely-applied standard normal homogeneity test method (Alexandersson 1986) was embedded to detect multiple breakpoints in the difference series. This method has demonstrated superior performance in terms of accurately identifying multiple breakpoints (DeGaetano 2006).

The last method we used, the RHtest, is a powerful software widely used to detect multiple breakpoints and homogenize various climate variables. Unlike the first 2 methods, only 1 series is referenced in the RHtest testing process. Lack of reference series is also operational, but may greatly reduce the reliability of the testing results. A good reference series should be homogeneous and highly correlated with the candidate series (Reeves et al. 2007). Therefore, in our study, high-quality reference series were matched for each target series. Briefly, we have evaluated the homogeneity of each reference series and eliminated potential undetected inhomogeneities by averaging them. Our explicit selection process is as follows: initially, the reference stations located within 300 km of the target station were used. The altitude difference between the reference station and the target station should be less than 200 m. We then calculated the correlation coefficients between each target series and its corresponding reference series, discarding the reference series with a correlation coefficient lower than 0.9. The remaining reference series underwent inhomogeneity tests using the PMFred algorithm (Wang 2008a,b) without references or metadata to optimize computing resources and save operation time. Inhomogeneous reference series were discarded, and only homogeneous ones were retained for the final detection. It is important to note that an insufficient number of reference series may fail to remove minor breakpoints when averaging, while an excessive number of reference series may result in a too smooth composite reference series, losing its original local variability. According to previous research (Peterson & Easterling 1994, Peterson et al. 1998) and our trials, we found that averaging 3 to 5 reference series is appropriate. Finally, the target series were tested with the composite reference series using the PMTred algorithm (Wang et al. 2007, Wang 2008a).

After conducting inhomogeneity tests using the 3 individual methods, 3 sets of breakpoint records were obtained, and all were included in the comprehensive evaluation. It has been suggested that certain methods, e.g. the PHA, potentially generate numerous false results that lack association with any metadata. These undocumented breakpoints reduce the reliability of the detection process (O'Neill et al. 2022). Considering that metadata are often missing or incomplete in many cases, we would expect a portion of undocumented breakpoints to be genuine in our tests, requiring a temperature shift of at least 2.0°C. To mitigate the risk of spurious breakpoints detected by any individual method, only the breakpoints identified by at least 2 methods within a 2 yr window would be considered significant and adjusted accordingly. If there was a match between the time of the breakpoint and the corresponding metadata record within a 1 yr window, the time of the breakpoint was modified to align with the metadata record. Otherwise, the time of the breakpoint was taken as the median of the 2 (or 3) results.

Lastly, breakpoints need to be adjusted according to the detection results. However, there may be 'urban blending' problems during the adjustment process of the target series. To address this, we opted not to introduce any reference series considering the potential systematic bias that resulted from the homogenization methods (Soon et al. 2018). That is, reference series were only applied during the step of breakpoint detection or identification, which was carried out with a confidence level of 99%. When adjusting the target series, the QM method (Vincent et al. 2012) was applied to the daily temperature data. Slightly different from other methods that calculate the mean shifts before and after the breakpoints, the QM method assesses the adjustment magnitude by comparing the empirical distribution of the series segments before and after the breakpoint in a quantile-based manner. The following paragraph provides a concise description of the QM method procedure.

For each breakpoint (M_c breakpoints in total), the 2 segments of the series that are separated by a breakpoint are selected to estimate the corresponding empirical cumulative frequency (ECF). The selected parts of the segments (the whole segment was selected in the present work) were sorted in ascending order and divided into M_q equally sized categories; in the present work, we set $M_q = 10$. Now let $P_b(k,l)$ denote the mean of the data in the segment for the l^{th} category before the k^{th} breakpoint, while $P_a(k,l)$ denotes the mean after the k^{th} changepoint. The difference in the l^{th} category mean between the 2 segments can then be derived as:

$$D(k,l) = P_a(k,l) - P_b(k,l) \quad (l = 1,2,\ldots,M_q) \tag{1}$$

In our adjustment, all other segments were adjusted to the latest segment of the series (i.e. the segment that contains the latest observation data, hereafter the 'base segment'). Let s_0 denote the base segment; then, for each segment $s \in [1,2,\ldots,(M_c+1)]$, the difference in the l^{th} category mean between the s^{th} and $s_0{}^{\text{th}}$ segments can be derived as:

$$A(s,l) = \begin{cases} \sum_{k=s}^{s_0-1} D(k,l) = \sum_{k=s}^{s_0-1} [P_a(k,l) - P_b(k,l)], & s < s_0 \\ 0 & , \quad s = s_0 \end{cases} \tag{2}$$

Meanwhile, a lower boundary of $A(s,0) = A(s,1)$ and an upper boundary of $A(s,M_q+1) = A(s,M_q)$ were set to restrain the mean of the QM adjustments for the respective category. Thus, including all M_q categories, we have (M_q+2) data points in total, to which a natural cubic spline was then fitted to obtain the QM adjustment for the series.

Let $\mathcal{F}_s(i)$ denote the ECF of the i^{th} data point in segment s. By referring to the inter-segment difference in the fitted spline for segment s, which corresponds to the cumulative frequency $\mathcal{F}_s(i)$, the $\mathcal{A}_s(i)$ value can be obtained, which is the difference between segments s and s_0, represented as the y-axis value. This $\mathcal{A}_s(i)$ value is the amount that will be added to the i^{th} datum in segment s of the series, which is referred to as the QM adjustment value. The specific steps of QM adjustment were described by Wang et al. (2010); refer to this research for more detailed information of the method.

In most cases, it is necessary to use a reference series for the adjustment of a time series. However, the existence of 'urban blending' may pose problems when employing a reference series for this purpose. In the presence of a reference series, the adjustment process operates on the difference between the target and reference series, with the magnitude of the adjustment contingent upon the specific changes in the difference series. If the reference series exhibits prominent urbanization effects, the difference between it and the target series will also encompass noticeable urbanization influences, which will result in QM adjustment values inevitably reflecting urbanization trends. As a consequence, the use of a reference series for adjustment introduces 'urban blending' issues. However, if no reference series is employed for the adjustment, the estimation of QM adjustment values relies on changes inherent to the target series. Subsequently, the resulting time series will not accumulate new urbanization effects; it simply adjusts the breakpoints. This approach would effectively avoid the 'urban blending' issue. Thus, in the present work, we have adjusted the temperature data without reference to other data series, and part of the urbanization effect in the temperature data series of urban stations will not be artificially diverted to those of rural stations.

2.3. Analysis method

In the calculation of the monthly data, if there were more than 7 missing days in a month, the average value of that month was considered as missing. Subsequently, the annual temperature series was derived by averaging the monthly data. If any month in a year was missing, the average value for that particular year was also marked as missing. Taking the period 1961–1990 as the climate reference period, the anomaly series for each station was obtained by subtracting the climate mean from the original temperature series. The regional mean series was then calculated from these anomaly series. To ensure the accuracy of both means and anomalies, it was required that the station series used for calculating means must have non-missing data for at least 15 yr within the climate reference period. In total, 653 of the 662 stations fulfilled this criterion.

Stations in North China are not evenly distributed, probably making the regional mean series heavily affected by the uneven density of stations in certain areas. Hence, when calculating the regional mean series in North China, the study region was divided into 20 grids measuring 3° × 2° (longitude × latitude). Subsequently, we computed the arithmetic mean of the anomaly series for each grid and obtained the regional series of North China by averaging 20 grid series using the grid area-weighted average method (Jones & Hulme 1996). The proportion of missing values in the grid series used for calculating the regional series was below 10%. The linear trends of the temperature series were assessed by a *t*-test (significant at $p < 0.05$).

3. RESULTS

3.1. Homogenization example

The effect of our homogenization was demonstrated by examining the temperature series of 2 stations in Hebei Province: Qinhuangdao and Wu'an. Fig. 2 shows a time series before and after homoge-

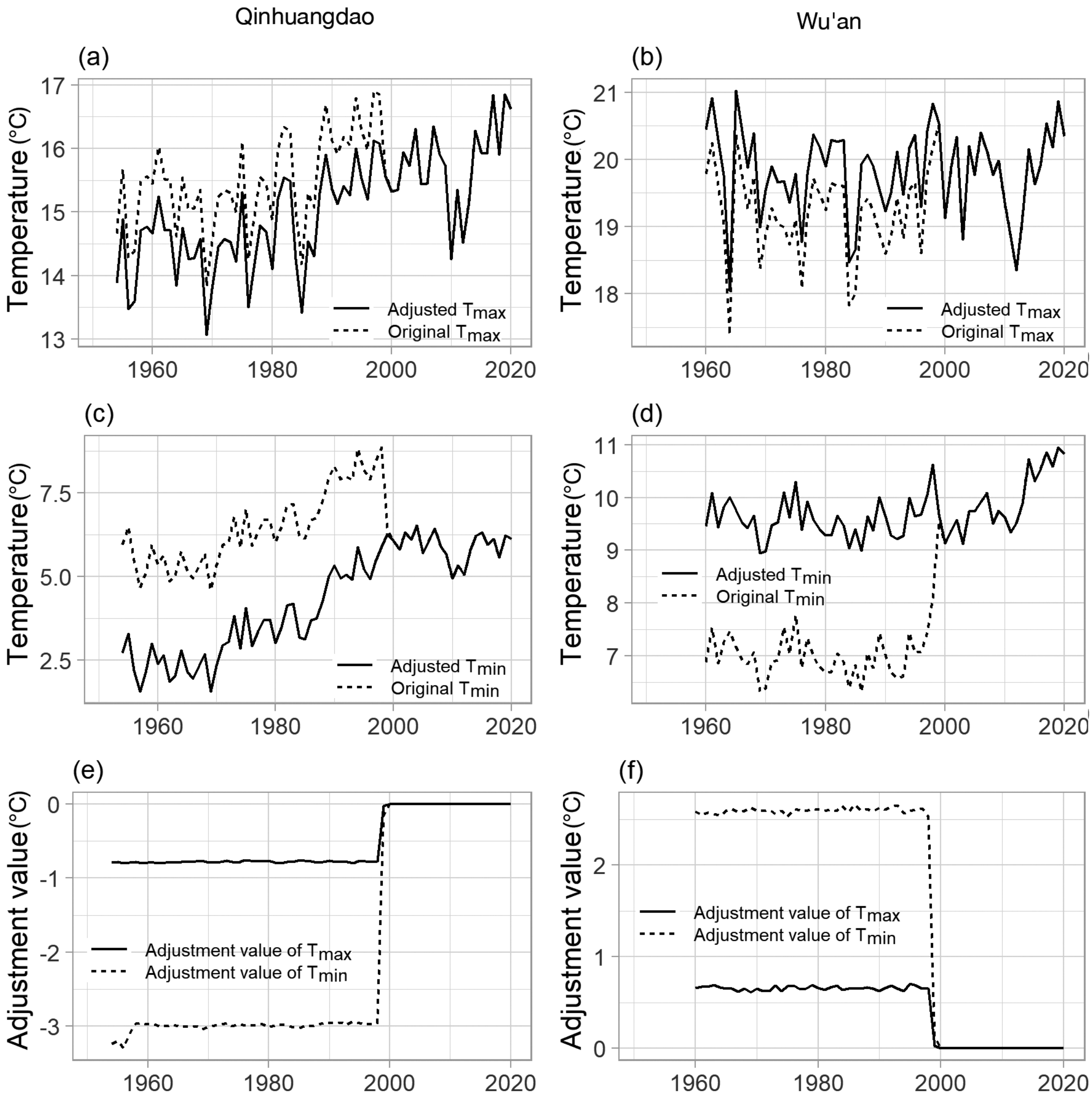

Fig. 2. Comparison between the original and adjusted series for (a,c,e) Qinhuangdao station and (b,d,f) Wu'an station. Panels (e,f) are the adjustment value curves of the corresponding series

nization, along with the adjustment values for each series. The results indicate that an average adjustment of approximately −0.8°C was applied to the T_{max} of Qinhuangdao station in January 1999. This adjustment resulted in an overall downward shift of the T_{max} series before January 1999, making the homogenized series exhibit a more pronounced trend compared to the original series. For the Qinhuangdao T_{min} series, 2 adjustments were made in June 1957 and January 1999, with corresponding adjustment values of approximately 3.3°C and −3.0°C, respectively. Homogenization had a substantial impact on the trend of the Qinhuangdao T_{min} series, which previously displayed a slight trend but exhibited a steeper incline after homogenization. For the Wu'an station, both the T_{max} and T_{min} series experienced breaks in January 1999, accompanied by adjustment values of approximately 0.7 and 2.6°C, respectively. Unlike Qinhuangdao station, the trends of Wu'an station decreased after the adjustment, with the T_{min} series undergoing the most significant change after the homogenization.

3.2. Statistics of homogenization data

Table 1 shows the detection results of the T_{max} and T_{min} series for North China from 1951 to 2020. For all 662 stations considered, the T_{max} series exhibited relatively better homogeneity compared to the T_{min} series, with approximately half indicating inhomogeneity. In contrast, the T_{min} series demonstrated higher susceptibility to inhomogeneity, with breakpoints identified for approximately 65 % of the series. Similar findings were observed by Zhou & Ren (2009) during their efforts to homogenize temperature data in North China. Regarding the number of breakpoints, the T_{max} series revealed a total of 481 instances, equating to an average of 1.43 breakpoints per series. For T_{min}, each series contained an average of 1.62 breakpoints, indicating that the T_{max} series experienced fewer abnormal breaks originating from non-climatic factors, and the man-made inhomogeneity exerted more pronounced effects on the T_{min} series.

Table 1. Numbers of inhomogeneous series and breakpoints detected from 662 stations in North China from 1951 to 2020. T_{max} (T_{min}): maximum (minimum) temperature

	T_{max}	T_{min}
Number of inhomogeneous series	337	431
Proportion of inhomogeneous series (%)	51	65
Number of breakpoints	481	699
Mean breakpoints in each series	1.43	1.62

Fig. 3a shows the temporal distribution of breakpoints. The maximum number of breakpoints was observed in 2004, corresponding to the implementation of an automatic observation system across the region. The transformation from manual to automatic observations resulted in systematic deviations in the surface air temperature. Specifically, the T_{max} values of automatic weather stations were higher than those of the manual observations, while the T_{min} for the new stations were lower than before (Wang et al. 2007). A minor peak in breakpoint occurrences was observed in 1980. Various triggers contributed to these breakpoints, including the issuance of a new observation criterion that year. Additionally, many other series suffered inhomogeneity due to station relocations. Stations such as Xiangning and Qingxu in Shanxi province, as well as Yongcheng and Lushan in Henan province, underwent relocations in 1980, which significantly impacted the homogeneity of the temperature records. Part of the breakpoints in 1980 also came from subtle instrument movements that were not documented as relocations. Occasionally, changes in longitude, latitude, or altitude occurred owing to the reconstruction of the observation sites and other factors, further affecting data homogeneity. The earlier peak period for the breakpoints falls between 1964 and 1965. Most stations in North China experienced relocations during this time, emerging as the primary factor contributing to the inhomogeneity of temperature series. Lastly, a few breakpoints were detected in 2015, which also coincided with station relocations.

Fig. 3b illustrates the classification of the breakpoint triggers in 4 aspects. Relocation emerges as the primary factor in posing inhomogeneity to the observation data, which resulted in over half of the temperature series breaks in North China. The secondary cause of the temperature inhomogeneity is attributed to alterations in the surrounding environment of the observation site or spatial displacement of the instrument. Metadata rarely document minor changes within the observation field; however, this study provides a comprehensive record of such changes. Consequently, every effort was made to determine and match the causes of the breakpoints in this regard. Additionally, changes in the observation criterion account for a fraction of the breaks. Instrument replacement is responsible for a few breakpoints, and specific metadata were unavailable to corroborate several breakpoints.

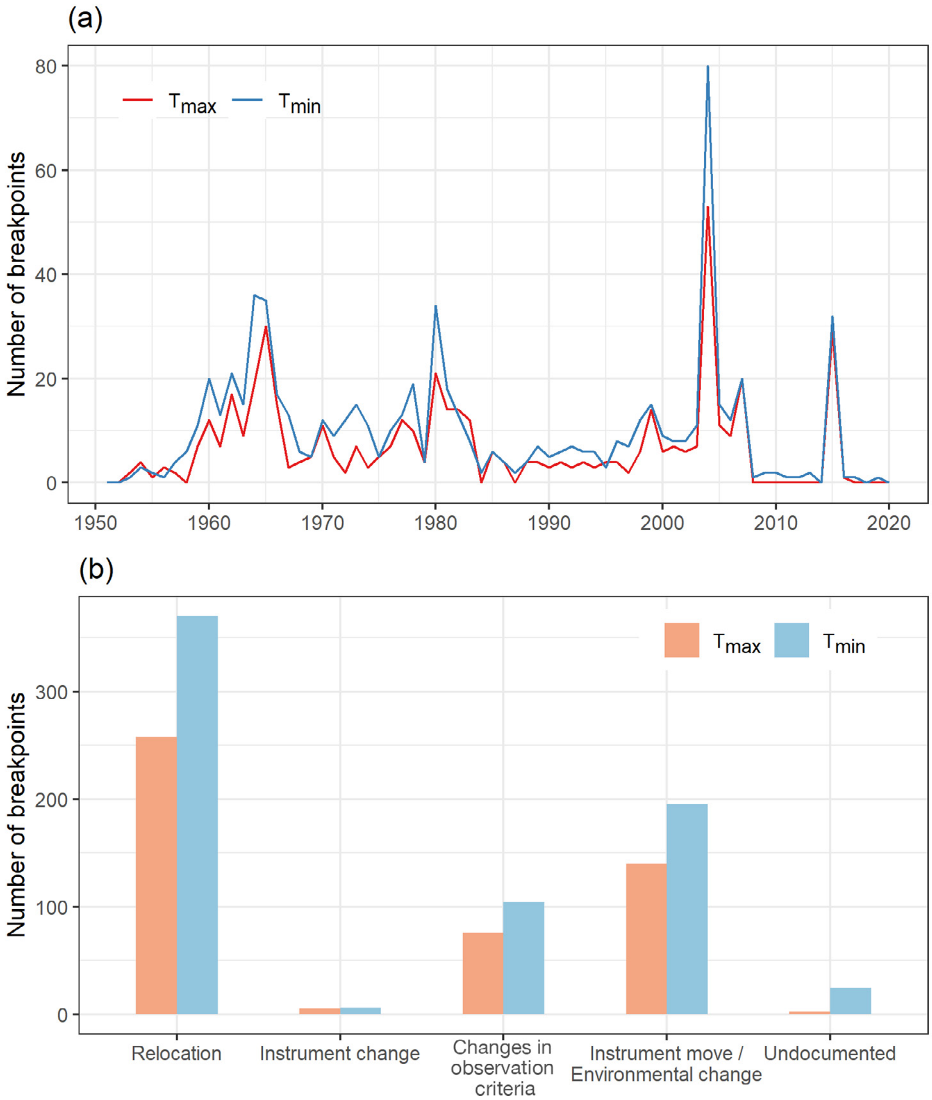

Fig. 3. (a) Total number of breakpoints for 662 stations in North China from 1951 to 2020, and (b) number of breakpoints for different causes

The distribution of adjusted values for all the breakpoints of T_{max} and T_{min} is presented in Fig. 4. Regarding T_{max}, the majority of break adjustments are concentrated within the range of −1.5 to 1.5°C, with a significant number of breakpoints adjusted around the value of −0.4°C. The adjustments for T_{min} are primarily distributed between −2.5 and 2.5°C, and the highest number of breakpoints occurs at the adjustment value of approximately −0.6°C. In general, there is a prevalence of negative adjustments for both T_{max} and T_{min} breakpoints. The average adjustment value for T_{max} breakpoints is −0.21°C, while that of T_{min} breakpoints is −0.33°C.

In terms of the effect of break adjustment, the adjustment magnitude associated with each break indicates how much the series segment before the breakpoint shifts in relation to the segment after. A negative adjustment results in a downward shift of the former segment split by the break, consequently causing the adjusted series to exhibit a larger trend than before the adjustment, as shown at Qinhuangdao station. Therefore, overall, the positive trends in the T_{min} series in North China increased after homogenization compared to their previous values. Similar changes happened to the T_{max} series, although they were slightly smaller compared to the T_{min} series.

In the spatial distribution of trend changes (Fig. 5), only a few stations in Hebei, Shandong, and Jiangsu provinces exhibit noticeable differences in the

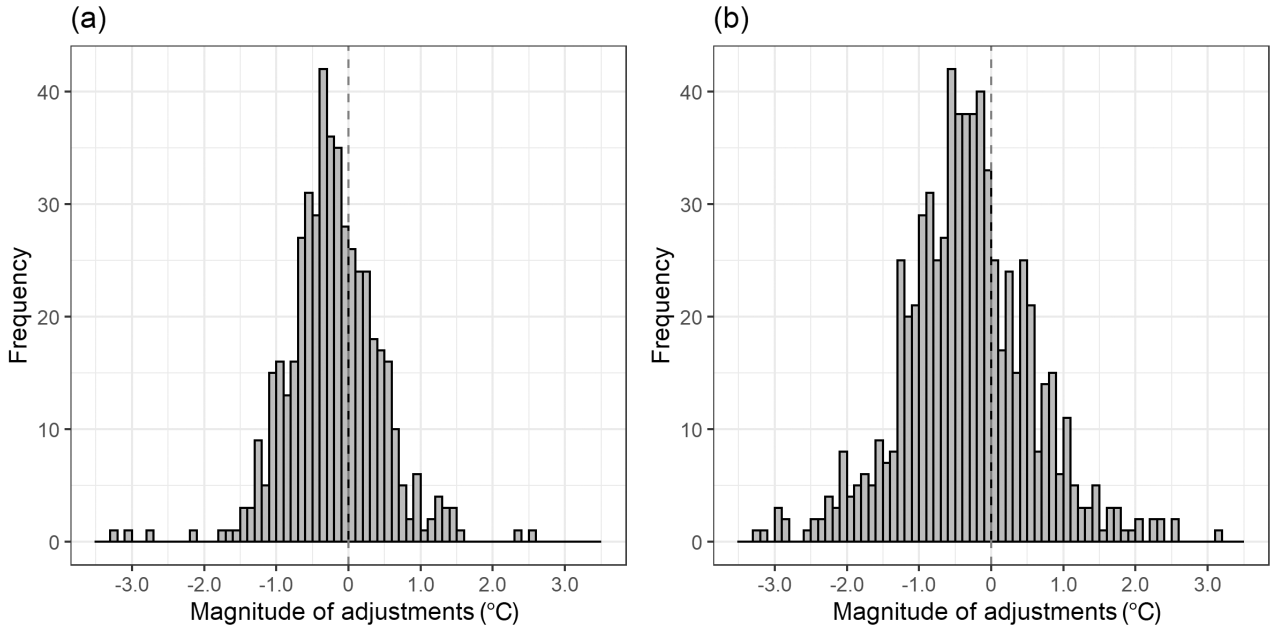

Fig. 4. Frequency of breakpoint adjustments for (a) maximum temperature (T_{max}) and (b) minimum temperature (T_{min}). The vertical dashed line represents zero value

trends before and after the homogenization process for T_{max}. Specifically, these stations experienced a more significant rise in temperature from 0.2–0.4°C per decade to 0.4–0.6°C per decade. In contrast, the trend changes observed in T_{min} after the homogenization are more pronounced. Among them, there is a significant increase in the number of stations exhibiting a warming trend above 0.8°C per decade. Overall, the warming of the homogenized T_{max} series increased to a small extent, which was not statistically significant, whereas the warming of T_{min} was more evident after homogenization.

3.3. Comparison with existing homogenized data

To compare and evaluate the homogenization of 2 groups of data, the difference series was calculated by subtracting the original series from the homogenized series adjusted by both the method used in this study and the one introduced by Cao et al. (2016). As demonstrated in Fig. 6, there is minimal discrepancy in the T_{max} series of the 3 data sets after 1965. However, the homogenized series in this study exhibit significantly smaller anomalies than the original series during the early decade. On the other hand, the series presented by Cao et al. (2016) show a slightly lower value than the original series in 1952 and 1953. By comparing their difference series with the original series, we can also observe the divergence between the approach employed in this study and that of Cao et al. (2016). The difference series calculated from our study display low values before 1965, resulting in a more pronounced increasing trend. This indicates that the homogenized T_{max} values in our study exhibit a higher increasing trend compared to the original data. Conversely, the trend observed by Cao et al. (2016) demonstrates little deviation from the original data.

The T_{min} series before 1965 exhibits similar characteristics. The homogenized T_{min} series in this study shows a significant decrease compared to the original series before 1965, which aligns with the findings of Cao et al. (2016), but only evident in 1952 and 1953. After 2007, both groups of the homogenized data demonstrate significantly greater anomalies than the original data. However, the results of this study indicate higher values than those of Cao et al. (2016). The rising trend of the T_{min} difference series, calculated from the current research and original data, is more noticeable than that of the T_{max} difference. This indicates that the homogenized T_{min} in this study exhibits a more pronounced warming trend than the original data. Although to a lesser extent, the homogenization performed by Cao et al. (2016) also increased the T_{min} trend compared to the original data.

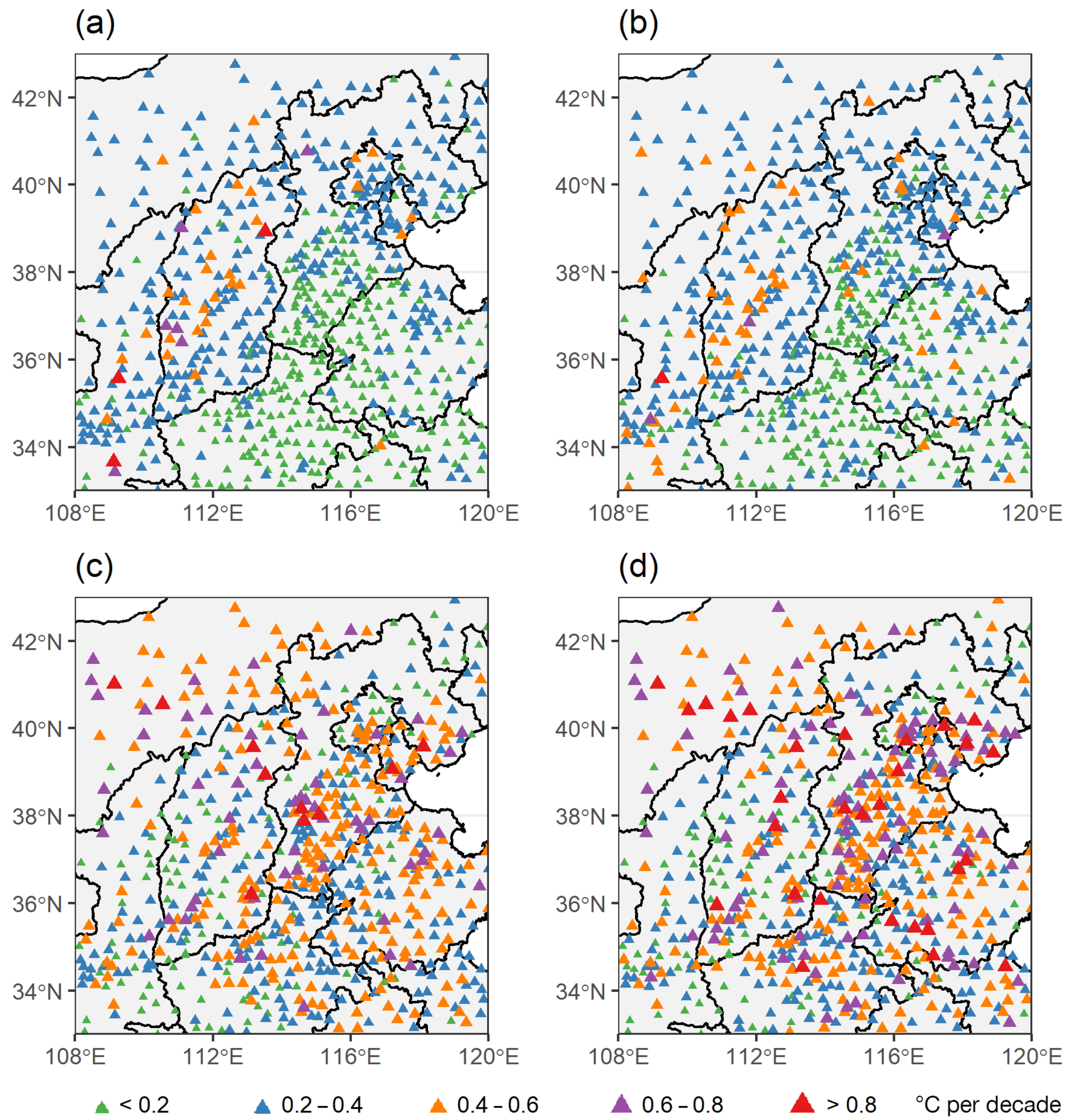

Fig. 5. Comparison between the (a,c) original and (b,d) adjusted temperature trend for 653 stations in North China from 1951 to 2020 for maximum temperature (T_{max}) (a,b) and minimum temperature (T_{min}) (c,d)

3.4. Warming trend in North China

Based on the new homogenized data, the long-term trends of T_{max} and T_{min} in the North China region from 1951 to 2020 were calculated. In addition, we obtained the mean temperature (T_{mean}) by averaging T_{max} and T_{min}, and the diurnal temperature range (DTR) by subtracting T_{min} from T_{max}. To analyze the seasonal variations, the year was divided into 4 seasons: spring (March, April, and May), summer (June, July, and August), autumn (September, October, and November), and winter (December, January, and February).

Fig. 7 depicts the spatial distribution of the trends of T_{max}, T_{min}, T_{mean}, and DTR across North China from 1951 to 2020. It can be seen that the spatial variation in T_{max} trends is relatively limited, with prominent temperature increases observed in southern central Inner Mongolia, most of Shanxi, northern Shaanxi, and southeastern Tianjin. The warming rates of T_{max} in other regions are comparably slower, particularly in Henan, where the warming effect is least discernible, and some areas experience nearly negligible warming trends. In stark contrast to T_{max}, T_{min} demonstrates a significant increase over a broader region, displaying notewor-

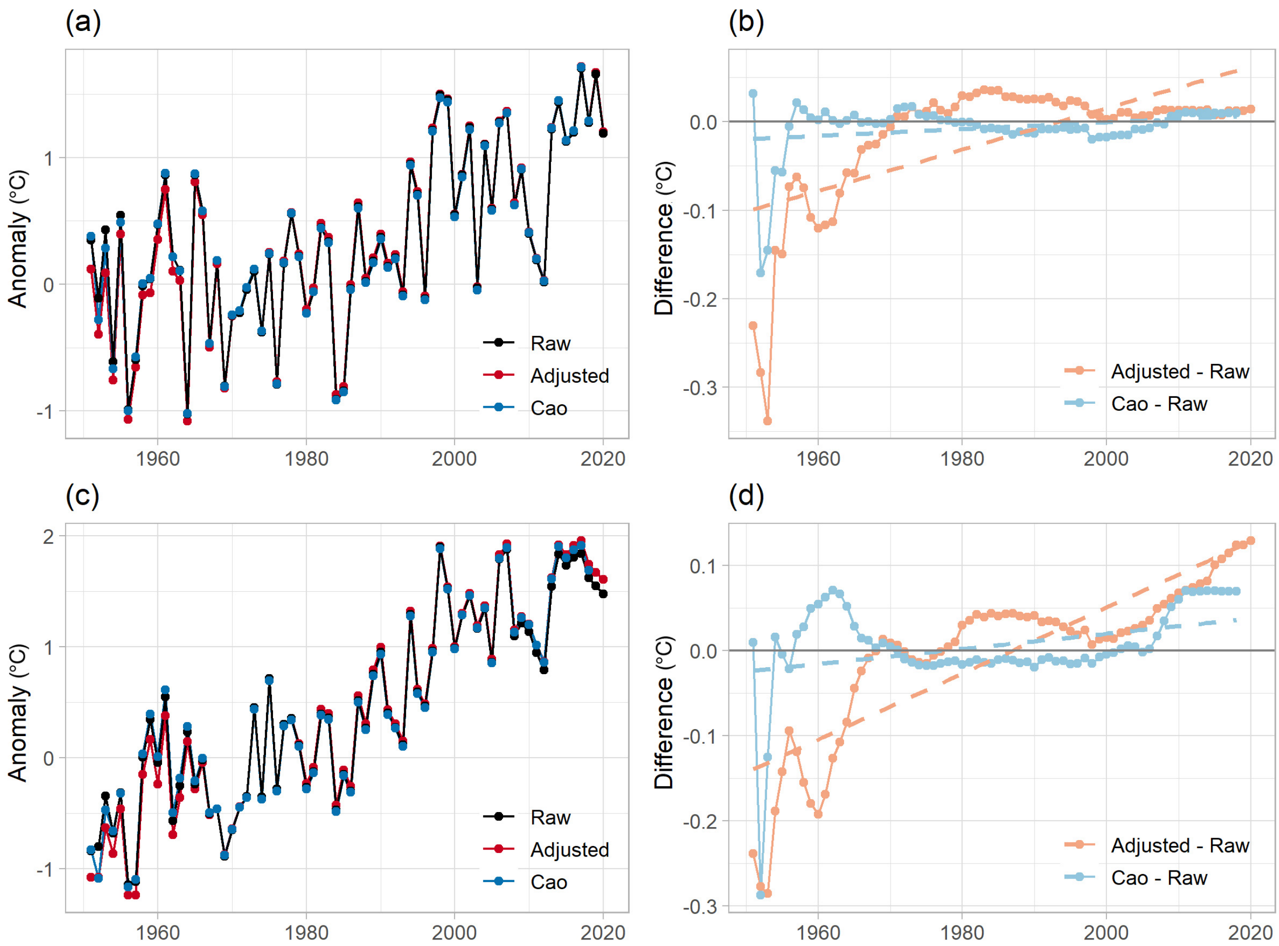

Fig. 6. Comparison between series calculated from (a,c) the different data sets and (b,d) the corresponding difference series from 1951 to 2020 for maximum temperature (T_{max}) (a,b) and minimum temperature (T_{min}) (c,d). Dashed lines: fitted linear trend

thy spatial disparities. The most pronounced warming trend is concentrated in central Inner Mongolia. The warming areas in the middle and southeast of North China take on an island-shaped pattern, with the largest T_{min} increase corresponding to the urban concentrated areas in various regions. Meanwhile, the warming trends in Shaanxi and the northeast corner of North China exhibit relatively minor changes. The spatial characteristics of T_{ave} roughly align with those of T_{min}, albeit with weaker and smaller warming rates and warming ranges. Central Inner Mongolia and the eastern part of the Beijing–Tianjin–Hebei region emerge as the major warming zones, while the remaining regions of North China exhibit scattered warming trends in a smaller range. DTR generally exhibits negative trends throughout North China, mirroring the spatial variations observed in T_{min}. Notably, there is a distinct positive value in the mid-western region of Shanxi, which corresponds to a significant warming center of T_{max}.

Overall, the spatial variations of T_{min} and DTR trends roughly align with the urban distribution in North China. The concentrated urban areas exhibit the most notable long-term trends, which could be attributed to the urbanization effect in this region (Ren et al. 2008). In contrast, the spatial variations of T_{max} trends do not show apparent differences. The trends observed in T_{mean} lie between those of T_{max} and T_{min}, probably indicating a relatively weak urbanization effect in the data series compared to that in T_{min}.

Fig. 8 illustrates the annual and seasonal time series, depicting temperature trends in North China. In general, T_{min} exhibits the most significant trends among both annual and seasonal series, while T_{max} experiences relatively smaller trends. Concerning the seasonal temperatures, the warming trends are more pronounced in spring and winter compared to

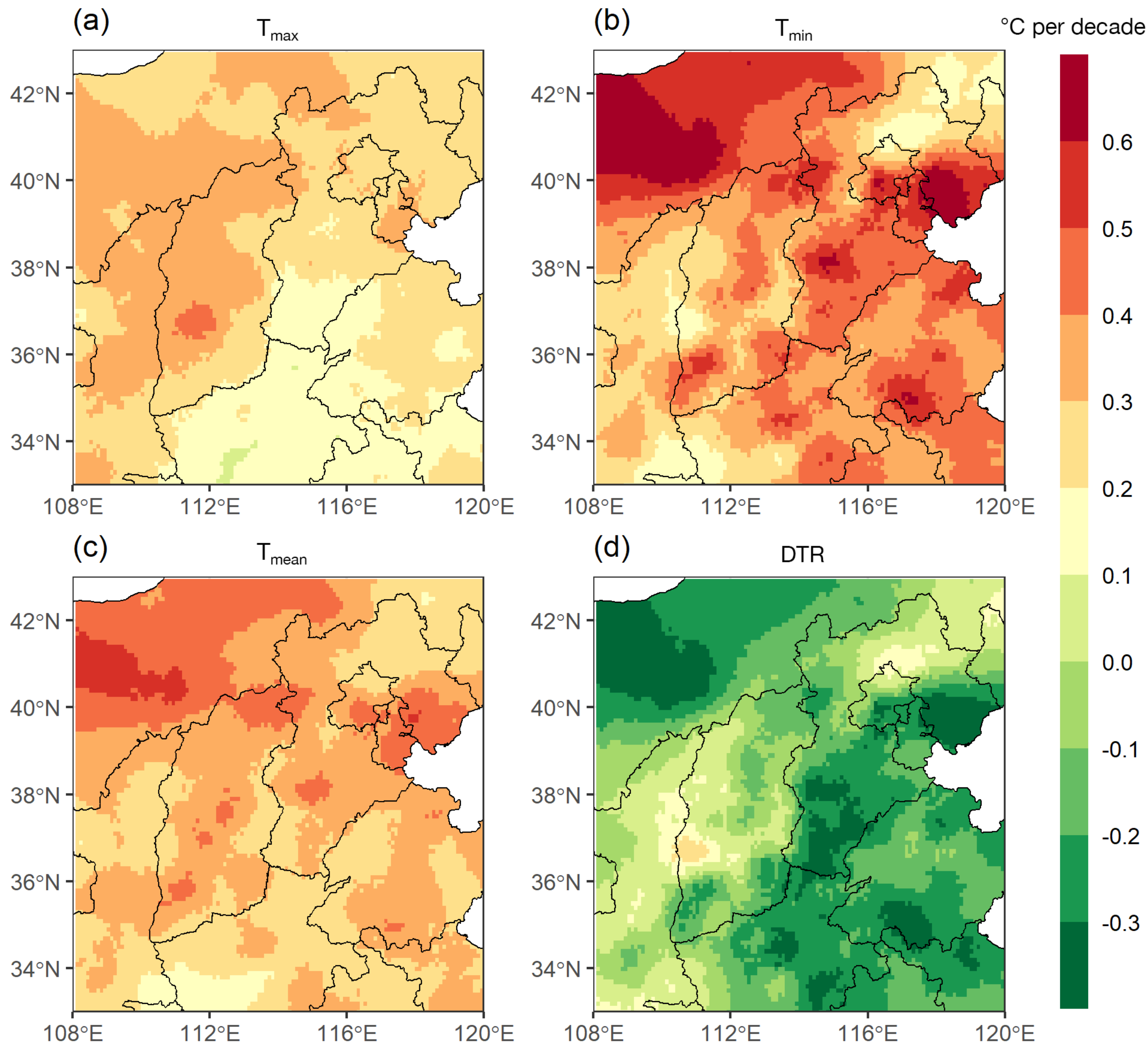

Fig. 7. Distribution of trends of (a) maximum temperature (T_{max}), (b) minimum temperature (T_{min}), (c) mean temperature (T_{mean}), and (d) diurnal temperature range (DTR) in North China from 1951 to 2020

autumn and summer. Before 1990, the warming trend in spring was relatively small, with no apparent changes observed in T_{max}. However, after 1990, North China experienced rapid warming in spring, except for a 'cold spring' year in 2010. Winter temperatures exhibit large variations before 1980 and after 1998, with relatively gentle trends during both periods. Notably, the coldest winter in the last 70 yr occurred in 1967, while apparent winter cooling took place between 1998 and 2012, possibly linked to the regional warming hiatus during the same period (Sun et al. 2018). Winter warming in North China mainly occurred from 1984 to 1998. In comparison, temperature changes in summer and autumn were moderate, with relatively small trends in both seasons. Before 1993, summer mean temperatures showed little variation, and the temperature trends remained relatively stable during this period. Autumn warming in North China primarily happened during 1981–1998, with no apparent temperature trends before or after this period, similar to the winter changes. Similarly, autumn temperatures in North China also experienced a hiatus after 1998, with T_{max} even exhibiting a weak downward trend during this hiatus period.

From the perspective of the annual mean temperatures, the temperature trends prior to 1984 were relatively small. However, after 1984, North China

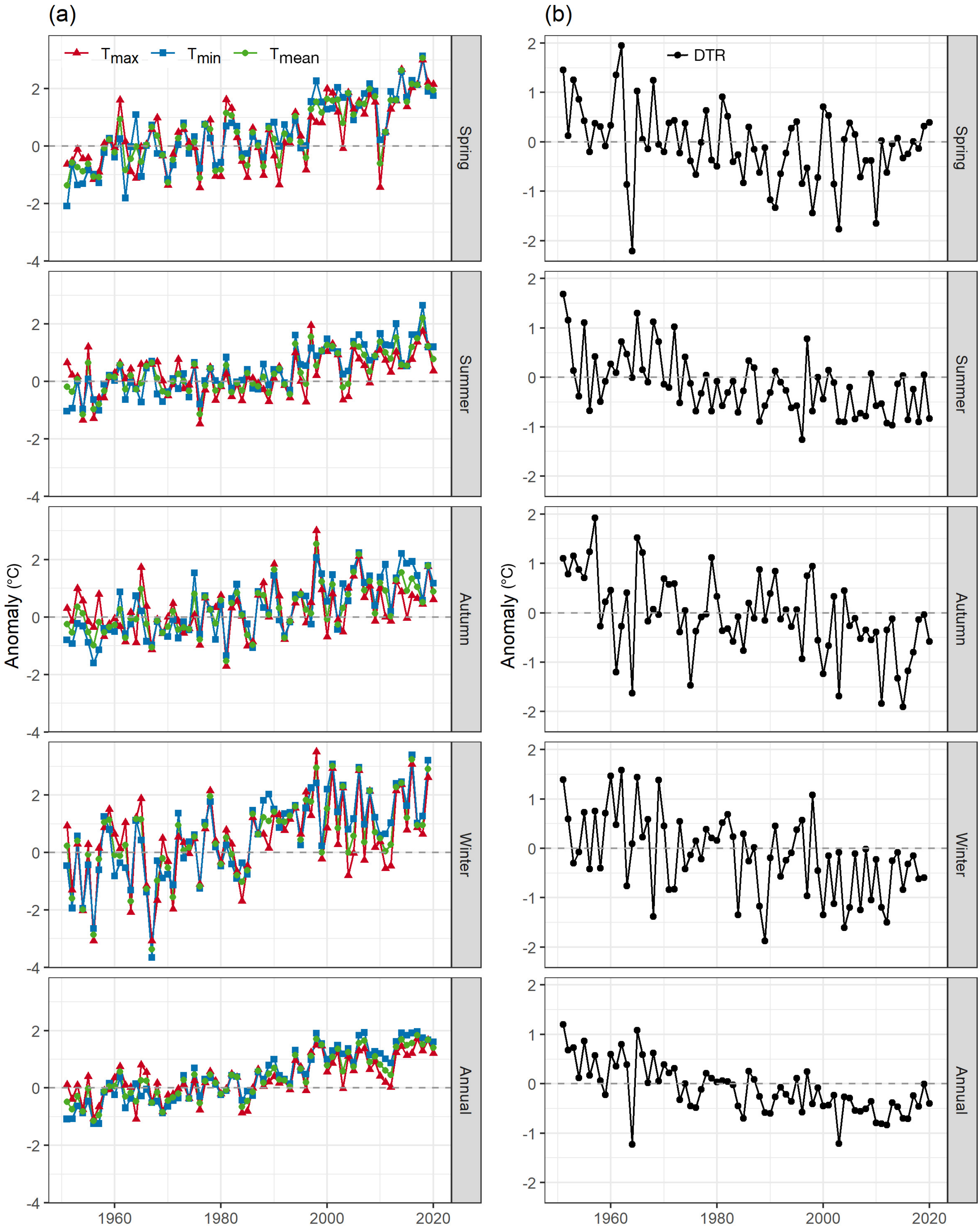

Fig. 8. Annual and seasonal mean time series for (a) maximum, minimum, and mean temperature (T_{max}, T_{min}, T_{mean}, respectively) and (b) diurnal temperature range (DTR) from 1951 to 2020

experienced a rapid temperature increase, reaching its peak in 2017. A warming hiatus occurred from 1998 to 2012 in North China, during which T_{max}, T_{min}, and T_{mean} all exhibited a decrease. Over the past 2 decades, temperatures showed no significant changes. For DTR, apparent downward trends can be observed in both annual and seasonal mean series throughout the study period. In particular, the DTR decreased most rapidly in the initial 30 yr of the study period. Except for autumn, the decline in DTR has slowed down in the last 40 yr.

Table 2 presents the trends of T_{max}, T_{min}, T_{mean}, and DTR in North China during the period 1951–2020. To compare our findings with the study conducted by Zhou & Ren (2009), which covers the years 1961–2000, the trend differences were calculated by subtracting their results from those of our own. These differences, indicated in parentheses in the table, demonstrate the variations between the 2 studies. In terms of the trends in annual mean temperature, our results generally exhibit slightly higher values compared to those of Zhou & Ren (2009). However, substantial disparities can be observed when analyzing the seasonal trends. Specifically, our study indicates significantly greater summer temperature trends compared to those of Zhou & Ren (2009). Notably, the trend of summer mean T_{min} exhibits a notable difference of 0.10°C per decade during same period. Conversely, the winter trends presented in our paper are generally lower than the findings of Zhou & Ren (2009). For instance, the trend of mean T_{min} in winter is 0.07°C per decade less than that reported by Zhou & Ren (2009) within the same period. Relatively small differences exist between the 2 studies for spring and autumn, with our results consistently showing slightly higher trends than those of Zhou & Ren (2009). Furthermore, differences between the 2 sets of results are also evident when considering the trends of DTR across seasons. Our paper demonstrates significantly larger DTR trends in spring and winter compared to those of Zhou & Ren (2009). Conversely, DTR trends in summer and autumn are slightly smaller in our study comparison to those of Zhou & Ren (2009). However, no significant differences were found in the trends of annual mean DTR between the 2 studies.

Table 2. Trends in annual and seasonal mean series for maximum temperature (T_{max}), minimum temperature (T_{min}), mean temperature (T_{mean}), and diurnal temperature range (DTR) of North China from 1951 to 2020. Values in parentheses are trend differences in the same period relative to the research of Zhou & Ren (2009) for 1961–2000 (unit:°C per decade). All trends in North China from 1951 to 2020 shown in the table are significant at the 0.01 level

	T_{max}	T_{min}	T_{mean}	DTR
Spring	0.34 (0.01)	0.46 (−0.03)	0.40 (0.01)	−0.13 (0.04)
Summer	0.14 (0.04)	0.31 (0.10)	0.23 (0.07)	−0.17 (−0.05)
Autumn	0.14 (0.02)	0.35 (0.05)	0.25 (0.04)	−0.20 (−0.02)
Winter	0.30 (−0.02)	0.49 (−0.07)	0.40 (−0.03)	−0.19 (0.06)
Annual	0.23 (0.02)	0.40 (0.01)	0.32 (0.02)	−0.17 (0.01)

4. DISCUSSION

A new homogenization was conducted for surface air temperature data in North China, using the ACMANT, PHA, and RHtest V4 software. The QM method was then applied to adjust the daily T_{max} and T_{min} data series. This adjustment process involved no reference series, thus eliminating the potential introduction of 'urban blending' (Soon et al. 2018) into the homogenized series. According to Soon et al. (2018), when homogenization is applied to an urbanized network, it tends to blend (or smooth or homogenize) the non-climatic biases among all stations, resulting in similar trends for most stations. Specifically, the trends of the highly urbanized stations are partially reduced, while those of predominantly rural stations are increased after homogenization. As our adjustment was made by applying the QM method and did not rely on any reference data series, it can be inferred that the homogenized data series in our study remains unaffected by the 'urban blending' effect.

We evaluated the potential 'urban blending' effect in the adjusted data by categorizing the series into groups. Based on Tysa et al. (2019), the temperature series undergoing adjustments were classified into 6 urbanization levels, ranging from L1 (the most 'rural' stations) to L6 (the most 'urban' stations), as shown in Table 3. The averaged adjustments demonstrated a general decrease from L1 to L5 for T_{max} and from L1 to L4 for T_{min}. This indicates that there were upward step changes in average trend differences as urbanization increased (negative adjustments before the breakpoints resulted in trend increases). Stations with urbanization levels ranging from L2 to L5, which likely experienced significant urbanization, showed larger linear trends after homogenization.

Table 3. Averaged net adjustments and trends of surface air temperature for different groups of stations at various urbanization levels. Urbanization increases from L1 (the most 'rural' stations) to L6 (the most 'urban' stations). For comparison of the difference before and after adjustment for inhomogeneous series, only the adjusted series are listed in this table. The difference values are approximated after calculation. T_{max} (T_{min}): maximum (minimum) temperature

Data	Urbanization level	Station count	Average net adjustment per station (°C)	Raw trend per station (°C per decade)	Homogenized trend per station (°C per decade)	Trend differences (°C per decade)
T_{max}	L1	19	−0.22	+0.39	+0.29	−0.09
	L2	49	−0.20	+0.23	+0.26	+0.03
	L3	120	−0.30	+0.22	+0.25	+0.03
	L4	129	−0.35	+0.20	+0.23	+0.03
	L5	14	−0.43	+0.18	+0.26	+0.08
	L6	6	+0.07	+0.30	+0.26	−0.04
T_{min}	L1	31	−0.24	+0.24	+0.22	−0.02
	L2	57	−0.51	+0.31	+0.36	+0.04
	L3	148	−0.63	+0.36	+0.44	+0.08
	L4	171	−0.56	+0.40	+0.46	+0.06
	L5	16	−0.27	+0.44	+0.47	+0.03
	L6	8	−0.20	+0.52	+0.50	−0.02

This should be interpreted as the recovery of the urbanization effect when stations relocated from more urban areas to more rural ones (Zhang et al. 2014, Ren et al. 2015). However, for both T_{max} and T_{min} of the L1 stations, which are usually situated in remote environments and rarely experience relocation events, the trends were somewhat reduced after homogenization. This reduction may represent the true rural temperature changes. Likewise, the homogenized linear trends of L6 stations, representing the most urbanized areas, were also reduced for both T_{max} and T_{min}. This may have to do with the smaller sample size of the stations in these groups. It is also possible that urban developments started much earlier at L6 stations compared to others and has stagnated in recent decades, resulting in a much lower recovered urbanization effect (represented by the averaged adjustments) than other station groups. Therefore, after adjustment, stations with higher levels of urbanization did not exhibit lower trends than before, while stations with lower levels of urbanization did not display significantly higher trends. This indicates that the adjusted data series do not contain the 'urban blending' effect.

To avoid false adjustments to potential climate variations, a 2.0°C threshold was set for identifying undocumented breakpoints. We also counted the number of discarded undocumented breakpoints using this criterion. In total, 1716 breakpoints for T_{max} series and 2339 breakpoints for T_{min} series were detected. This indicates that approximately 70% of the detected breakpoints did not meet the criteria and were discarded. In our study, a 1 yr window was used to match the breakpoints with the metadata. O'Neill et al. (2022) reported that about 80% of the breakpoints in the European GHCN data, identified by the PHA method, could not be matched with any metadata within a 1 yr window. Given that a portion of the undocumented breakpoints were conditionally adjusted in our study, we would expect that the proportion of the unadjusted breakpoints in our study to be slightly lower than the proportion of undocumented breakpoints in the study by O'Neill et al. (2022). Moreover, multiple methods were applied in our detection process, and some breakpoints identified by any single method exclusively were not deemed significant. Consequently, fewer undocumented breakpoints were taken into account. This suggests that employing combined methods in our detection process reduces uncertainty to some extent.

The statistical analysis in this study reveals that the number of breakpoints adjusted in T_{min} is greater than that in T_{max}, and the average number of breakpoints per series is also higher in T_{min}. Regarding the spatial movement of stations, distant station relocations can have a considerable impact on both T_{max} and T_{min} data, leading to noticeable discontinuities in the temperature series. However, minor relocations or slight movements, especially those occurring within close proximity, tend to only affect the T_{min} series, while having no apparent influence on T_{max}. Previous studies (Zhou & Ren 2009, Xu et al. 2013, Cao et al. 2016) have emphasized that T_{min} data are more susceptible to spatial movements of the observation devices, primarily related to the significant urbanization effect (particularly the urban heat island effect) on T_{min} measurement (Ren & Zhou 2014). When a station relocates to a remote rural or suburban location –an approach commonly practiced in developing countries and regions– T_{min} may exhibit a significant decrease due to the absence or weakening of the urban heat island effect, resulting in an apparent break in the series. Occasionally, stations might be relocated from the suburbs to urban areas owing to station management or other factors. In such cases, T_{min} could considerably increase from a lower temperature to a higher temperature, generating noticeable breaks in the series.

In terms of the adjustment magnitude of the breakpoints, T_{min} exhibits a greater adjustment than T_{max}. This indicates that T_{min} is more significantly influenced by inhomogeneous factors, whereas T_{max} is relatively less susceptible to such non-climatic factors. When an artificial discontinuity occurs, it induces a relatively minor inhomogeneity in the T_{max} series. Meteorological stations typically record T_{max} at 14:00 h local time, which represents the daytime temperature at the station. Ordinarily, urban and rural areas experience similar daytime temperatures. Significant effects on T_{max} occur only when there is a long-distance or substantial vertical relocation, for example, between urban center and remote rural areas or between valleys and mountain tops. However, historical records indicate infrequent occurrences of such station relocations. As a result, fewer breakpoints were detected in T_{max}, and their magnitudes were comparatively low.

Our research shows overall larger trends in annual and seasonal temperature changes in North China compared to other data sets, including the original data and the existing homogenized data (Table 2, Fig. 6). This difference may be mainly related to the homogenization methods employed. The observed negative differences in winter between our study and that of Zhou & Ren (2009) may be owing to the difference in data type and the regional winter warming hiatus in eastern China since 1998 (Sun et al. 2017), which has been included in the updated analysis. Additionally, the dissimilarity in the number of stations used by the 2 studies could also serve as a contributing factor. Zhou & Ren (2009) estimated the temperature change in North China using 95 national basic meteorological stations and national reference climate stations (i.e. the national stations), whereas our study encompassed 653 stations in North China, including ordinary meteorological stations alongside the national stations.

In China, most observation stations are located within or near cities and towns of varying sizes, so they may have recorded urbanization effects in their historical temperature data series. Zhang et al. (2010) analyzed the urbanization effect among 614 national stations in China from 1961 to 2004, revealing that at least 27% of the temperature trend results from urbanization. Ren et al. (2008) demonstrated that the urbanization effect contributes more than 38% of the temperature trend increases at the national stations of North China from 1961 to 2000, which surpasses the national average. Notably, although the urbanization contribution in the national stations remains relatively small in winter, the absolute warming induced by urbanization, i.e. the urbanization effect, is often the most prominent of the year (Ren et al. 2008, Zhou & Ren 2009). The present analysis did not take into consideration the urbanization effects, thus leading to an evident overestimation of regional warming trends in the study area. In addtion, data homogenization could have caused a recovery of urbanization effects in the station temperature series due to the fact that the majority of the station relocations in China occurred from urban to rural areas (Zhang et al. 2014, Ren et al. 2015). This perhaps explains why the current homogenization amplifies the temperature trends in North China to a certain extent. The recovered urbanization effect at urban stations also accounts for the somewhat inconsistent spatial patterns of homogenized temperature trends depicted in Figs. 5 & 7. The homogenized temperature data may exhibit scattered climatic gradients due to the non-uniform urbanization effect over China, including North China, whereas adjusting the urbanization bias in surface air temperature data would yield a more consistent spatial pattern in the climatic background warming trends (Wen et al. 2019). Therefore, the present obtained homogenized temperature data in North China still retain significant urbanization bias. Thus a homogenization of temperature data, especially in-depth homogenization like that conducted in our work, necessitates further evaluations and adjustments of urbanization bias more than before.

Another reason for the overestimated trends in annual and seasonal mean temperatures in North China could be attributed to the statistical method employed to calculate the daily mean (monthly and annual mean) temperature. Traditionally, the daily, monthly, and annual mean temperatures are obtained by calculating the arithmetical average of T_{max} and T_{min}. However, compared to the standard procedure or equal-interval averaging method (4, 8, 24 records a day), the arithmetic mean of T_{max} and T_{min} tends to overestimate the overall average and the linear trends, as pointed out by Liu et al. (2019). It is worth noting that although the bias resulting from the averaging method is generally smaller in North China compared to the Qinghai-Tibet Plateau and Northwest China, it still holds statistical significance when considering the data series for spring, summer, autumn, and the entire year. Therefore, it is imperative to take this factor into account in future studies.

5. CONCLUSIONS

In this study, we present an updated analysis of surface warming in North China spanning the period 1951–2020. This analysis is based on newly homogenized surface air temperature data derived from a high-density observation network. The following are the main conclusions drawn from our study.

(1) Using the ACMANT, PHA, and RHtest software, the maximum and minimum temperature series in North China were jointly tested for inhomogeneity with reference to the metadata of each station. During the testing, a reference series was strictly constructed for each target series. A set of homogenized daily maximum and minimum temperature data of 662 meteorological stations in North China were then obtained by using the QM method for adjustment. This adjustment process, without relying on any reference data series, should avoid the 'urban blending' bias resulting from the homogenization method.

(2) The inhomogeneity of the temperature series in North China is primarily attributed to station relocation, which accounts for over half of the detected breakpoints in the temperature data. Furthermore, environmental changes at the stations also have noticeable effects on the homogeneity of the temperature series. The periods of extensive station relocation primarily occurred during the 1960s and 1980s, resulting in numerous breakpoints. Additionally, the introduction of automatic observation instruments in the early 2000s caused significant discontinuities in the data.

(3) The inhomogeneity of minimum temperature was more prevalent compared to the maximum temperature, which was less disturbed by non-climatic factors. The homogenization conducted in this study resulted in an overall increase in the temperature trend in North China. Although the maximum temperature exhibited a relatively small rise range, the adjustment significantly increased the rise rate of the minimum temperature. Furthermore, little difference was noted in the spatial distribution of the linear trend between the original and homogenized maximum temperature data, while apparent rising trends were observed in most provinces and cities for the homogenized minimum temperature.

(4) The annual T_{max} trend in North China was generally small, displaying negligible spatial difference. It tended to be higher in the northwest and lower in the southeast. On the other hand, the trends for T_{min} and DTR were more pronounced and exhibited distinct spatial patterns. Each high- or low-value center aligned with specific urban areas, suggesting that urbanization has had a notable impact on the trends of T_{min} and DTR in North China. Similarly, annual and seasonal mean T_{mean} also exhibited significant changes, although these changes were typically less pronounced compared to T_{min}.

(5) The annual mean temperature trend in North China has undergone slight changes in recent years. The warming in this region primarily began in the 1980s, followed by a regional slowdown in warming from 1998 to 2012. Notably, the most significant increases in temperature occurred during spring and winter, while the trends for summer and autumn were relatively modest. Additionally, both the annual and seasonal mean DTR exhibited noticeable downward trends.

(6) Larger trends in the annual and seasonal mean temperatures in North China were observed compared to previous research. However, it is worth noting that the temperature trend during winter was significantly weaker than that reported in previous analyses. These significant changes in temperature may be attributed primarily to the recovery of the urbanization effect in the temperature data series of national stations after undergoing thorough homogenization. Thus, the homogenized temperature data obtained in this study still retain apparent urbanization bias, which necessitates further evaluation and adjustment in future research endeavors.

Acknowledgements. This study was supported by the National Key Research and Development Project (Grant No. 2018YFA0605603). We thank Dr. Ronan Connolly and 2 anonymous reviewers for their valuable comments and suggestions on our manuscript, which significantly improved the robustness of our study.

LITERATURE CITED

Aguilar E, Auer I, Brunet M, Peterson TC, Wieringa J (2003) Guidance on metadata and homogenization. Wmo Td, 1186, WMO, Geneva

Alexandersson H (1986) A homogeneity test applied to precipitation data. J Climatol 6:661–675

Cao L, Zhu Y, Tang G, Yuan F, Yan Z (2016) Climatic warming in China according to a homogenized data set from 2419 stations. Int J Climatol 36:4384–4392

Caussinus H, Lyazrhi F (1997) Choosing a linear model with a random number of change-points and outliers. Ann Inst Stat Math 49:761–775

Cubasch UD, Wuebbles D, Chen D, Facchini MC, Winther JG (2013) Introduction to Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge

DeGaetano AT (2006) Attributes of several methods for detecting discontinuities in mean temperature series. J Clim 19:838–853

Ding YH, Wang HJ (2016) Newly acquired knowledge on the scientific issues related to climate change over the recent 100 years in China. Chin Sci Bull 61:1029–1041

Domonkos P (2011a) Adapted Caussinus-Mestre algorithm for networks of temperature series (ACMANT). Int J Geosci 2:293–309

Domonkos P (2011b) Efficiency evaluation for detecting inhomogeneities by objective homogenisation methods. Theor Appl Climatol 105:455–467

Domonkos P (2013) Efficiencies of inhomogeneity-detection algorithms: comparison of different detection methods and efficiency measures. J Climatol 2013:390945

Domonkos P, Guijarro JA, Venema V, Brunet M, Sigró J (2021) Efficiency of time series homogenization: method comparison with 12 monthly temperature test datasets. J Clim 34:2877–2891

Domonkos P, Tóth R, Nyitrai L (2022) Climate observations: data quality control and time series homogenization. Elsevier, Amsterdam

Easterling DR, Peterson TC (1995) A new method for detecting undocumented discontinuities in climatological time series. Int J Climatol 15:369–377

Gubler S, Hunziker S, Begert M, Croci-Maspoli M and others (2017) The influence of station density on climate data homogenization. Int J Climatol 37:4670–4683

Guijarro JA (2018) Homogenization of climatic series with Climatol. Tech Rep. State Meteorological Agency (AEMET), Balearic Islands Office

Hawkins DM (1972) On the choice of segments in piecewise approximation. IMA J Appl Math 9:250–256

He J, Ren G, Zhang P (2021) Effects of data homogenization on the estimates of temperature trend and urbanization bias: taking Beijing area as an example. Adv Clim Change Res 17:503–513 (in Chinese with English abstract)

Jones PD, Hulme M (1996) Calculating regional climatic time series for temperature and precipitation: methods and illustrations. Int J Climatol 16:361–377

Jones PD, Lister DH, Osborn TJ, Harpham C, Salmon M, Morice CP (2012) Hemispheric and large-scale land-surface air temperature variations: an extensive revision and an update to 2010. J Geophys Res 16:D05127

Kolendowicz L, Czernecki B, Półrolniczak M, Taszarek M, Tomczyk AM, Szyga-Pluga K (2019) Homogenization of air temperature and its long-term trends in Poznań (Poland) for the period 1848–2016. Theor Appl Climatol 136:1357–1370

Kuglitsch FG, Auchmann R, Bleisch R, Brönnimann S, Martius O, Stewart M (2012) Break detection of annual Swiss temperature series. J Geophys Res D Atmos 117:D13105

Lenssen NJ, Schmidt GA, Hansen JE, Menne MJ, Persin A, Ruedy R, Zyss D (2019) Improvements in the GISTEMP uncertainty model. J Geophys Res D Atmos 124: 6307–6326

Li Q, Liu X, Zhang H, Peterson TC, Easterling DR (2004) Detecting and adjusting temporal inhomogeneity in Chinese mean surface air temperature data. Adv Atmos Sci 21:260–268

Li Z, Yan Z (2009) Homogenized daily mean/maximum/minimum temperature series for China from 1960–2008. Atmos Ocean Sci Lett 2:237–243

Lindau R, Venema V (2013) On the multiple breakpoint problem and the number of significant breaks in homogenization of climate records. Idojaras Q J Hung Meteorol Serv 117:1–34

Lindau R, Venema V (2016) The uncertainty of break positions detected by homogenization algorithms in climate records. Int J Climatol 36:576–589

Liu Y, Ren G, Kang H, Sun X (2019) A significant bias of T_{max} and T_{min} average temperature and its trend. J Appl Meteorol Climatol 58:2235–2246

Mamara A, Argiriou AA, Anadranistakis M (2014) Detection and correction of inhomogeneities in Greek climate temperature series. Int J Climatol 34:3024–3043

Menne MJ, Williams CN (2005) Detection of undocumented changepoints using multiple test statistics and composite reference series. J Clim 18:4271–4286

Menne MJ, Williams CN (2009) Homogenization of temperature series via pairwise comparisons. J Clim 22:1700–1717

Menne MJ, Williams CN, Gleason BE, Rennie JJ, Lawrimore JH (2018) The global historical climatology network monthly temperature dataset, version 4. J Clim 31: 9835–9854

Mestre O, Domonkos P, Picard F, Auer I and others (2013) HOMER: a homogenization software — methods and applications. Idojaras Q J Hung Meteorol Serv 117:47–67

Morozova AL, Valente MA (2012) Homogenization of Portuguese long-term temperature data series: Lisbon, Coimbra and Porto. Earth Syst Sci Data 4:187–213

NARCC Compilation Committee (2015) China's Third National Assessment Report on Climate Change. Science Press, Beijing (in Chinese)

O'Neill P, Connolly R, Connolly M, Soon W and others (2022) Evaluation of the homogenization adjustments applied to European temperature records in the global historical climatology network dataset. Atmosphere 13:285

Peterson TC, Easterling DR (1994) Creation of homogeneous composite climatological reference series. Int J Climatol 14:671–679

Peterson TC, Easterling DR, Karl TR, Groisman P and others (1998) Homogeneity adjustments of *in situ* atmospheric climate data: a review. Int J Climatol 18:1493–1517

Pielke R Sr, Nielsen-Gammon J, Davey C, Angel J and others (2007) Documentation of uncertainties and biases associated with surface temperature measurement sites for climate change assessment. Bull Am Meteorol Soc 88: 913–928

Reeves J, Chen J, Wang XL, Lund R, Lu QQ (2007) A review and comparison of changepoint detection techniques for climate data. J Appl Meteorol Climatol 46:900–915

Ren G, Zhou Y (2014) Urbanization effect on trends of extreme temperature indices of national stations over Mainland China, 1961–2008. J Clim 27:2340–2360

Ren G, Xu M, Chu Z, Guo J, Li Q, Liu X, Wang Y (2005) Changes of surface air temperature in China during 1951–2004. Clim Environ Res 10:717–727

Ren G, Zhou Y, Chu Z, Zhou J, Zhang A, Guo J, Liu X (2008) Urbanization effects on observed surface air temperature trends in North China. J Clim 21:1333–1348

Ren G, Li J, Ren Y, Chu Z and others (2015) An integrated procedure to determine a reference station network for evaluating and adjusting urban bias in surface air temperature data. J Appl Meteorol Climatol 54:1248–1266

Ren Z, Yu Y, Zou F, Xu Y (2012) Quality detection of surface historical basic meteorological data. Yingyong Qixiang Xuebao 23:739–747 (in Chinese)

Rohde RA, Hausfather Z (2020) The Berkeley Earth land/ocean temperature record. Earth Syst Sci Data 12: 3469–3479

Si P, Li Q, Jones P (2021) Construction of homogenized daily surface air temperature for the city of Tianjin during 1887–2019. Earth Syst Sci Data 13:2211–2226

Song L (1994) Characteristics of temperature in spatial and temporal variation in China during recent 40 years. Q J Appl Meteorol 5:119–124

Soon WWH, Connolly R, Connolly M, O'Neill P and others (2018) Comparing the current and early 20th century warm periods in China. Earth Sci Rev 185:80–101

Squintu AA, van der Schrier G, Brugnara Y, Klein Tank A (2019) Homogenization of daily temperature series in the European Climate Assessment & Dataset. Int J Climatol 39:1243–1261

Squintu AA, van der Schrier G, Štěpánek P, Zahradníček P, Tank AK (2020) Comparison of homogenization methods for daily temperature series against an observation-based benchmark dataset. Theor Appl Climatol 140: 285–301

Sun X, Ren G, Xu W, Li Q, Ren Y (2017) Global land-surface air temperature change based on the new CMA GLSAT data set. Sci Bull (Beijing) 62:236–238

Sun X, Ren G, Ren Y, Fang Y, Liu Y, Xue X, Zhang P (2018) A remarkable climate warming hiatus over Northeast China since 1998. Theor Appl Climatol 133:579–594

Tang GL, Ren GY (2005) Reanalysis of surface air temperature change of the last 100 years over China. Clim Environ Res 10:791–798

Trewin B (2013) A daily homogenized temperature data set for Australia. Int J Climatol 33:1510–1529

Trewin B, Braganza K, Fawcett R, Grainger S and others (2020) An updated long-term homogenized daily temperature data set for Australia. Geosci Data J 7:149–169

Tysa SK, Ren G, Qin Y, Zhang P, Ren Y, Jia W, Wen K (2019) Urbanization effect in regional temperature series based on a remote sensing classification scheme of stations. J Geophys Res D Atmos 124:10646–10661

Venema VKC, Mestre O, Aguilar E, Auer I and others (2012) Benchmarking homogenization algorithms for monthly data. Clim Past 8:89–115

Vincent LA (1998) A technique for the identification of inhomogeneities in Canadian temperature series. J Clim 11: 1094–1104

Vincent LA, Wang XL, Milewska EJ, Wan H, Yang F, Swail V (2012) A second generation of homogenized Canadian monthly surface air temperature for climate trend analysis. J Geophys Res D Atmos 117:D18110

Vincent LA, Milewska EJ, Wang XL, Hartwell MM (2018) Uncertainty in homogenized daily temperatures and derived indices of extremes illustrated using parallel observations in Canada. Int J Climatol 38:692–707

Wang SW (1990) Variations of temperature in China for the 100 year period in comparison with global temperatures. Meteorol Monthly 16:11–15 (in Chinese)

Wang SW, Ye JL, Gong DY, Zhu JH, Yao TD (1998) Construction of mean annual temperature series for the last one hundred years in China. Q J Appl Meteorol 9:392–401

Wang XL (2008a) Accounting for autocorrelation in detecting mean shifts in climate data series using the penalized maximal *t* or *F* test. J Appl Meteorol Climatol 47: 2423–2444

Wang XL (2008b) Penalized maximal *F* test for detecting undocumented mean shift without trend change. J Atmos Ocean Technol 25:368–384

Wang XL, Feng Y (2013) RHtestsV4 user manual. http://etccdi.pacificclimate.org/RHtest/RHtestsV4_UserManual_10Dec2014.pdf

Wang XL, Wen QH, Wu Y (2007) Penalized maximal t test for detecting undocumented mean change in climate data series. J Appl Meteorol Climatol 46:916–931

Wang XL, Chen H, Wu Y, Feng Y, Pu Q (2010) New techniques for the detection and adjustment of shifts in daily precipitation data series. J Appl Meteorol Climatol 49: 2416–2436

Wang Y, Liu X, Ju X (2007) Differences between automatic and manual observation. Yingyong Qixiang Xuebao 18: 849–855 (in Chinese)

Wen K, Ren G, Li J, Ren Y, Sun X, Zhou Y, Zhang A (2019) Adjustment of urbanization bias in surface air temperature over the mainland of China. Prog Geogr 38:600–611 (in Chinese with English abstract)

Xu W, Li Q, Wang XL, Yang S, Cao L, Feng Y (2013) Homogenization of Chinese daily surface air temperatures and analysis of trends in the extreme temperature indices. J Geophys Res D Atmos 118:9708–9720

Yan Z, Yang C, Jones P (2001) Influence of inhomogeneity on the estimation of mean and extreme temperature trends in Beijing and Shanghai. Adv Atmos Sci 18: 309–322

Zhang A, Ren G, Zhou X, Chu Z, Ren Y, Tang G (2010) On the urbanization effect on surface air temperature trends over China. Acta Meteorol Sin 68:957–966 (in Chinese)

Zhang L, Ren GY, Ren YY, Zhang AY, Chu ZY, Zhou YQ (2014) Effect of data homogenization on estimate of temperature trend: a case of Huairou station in Beijing Municipality. Theor Appl Climatol 115:365–373

Zhou Y, Ren G (2009) The effect of urbanization on maximum, minimum temperatures and daily temperature range in North China. Plateau Meteorol 28:1158–1166 (in Chinese)

Editorial responsibility: Oliver Frauenfeld, College Station, Texas, USA
Reviewed by: R. Connolly and 2 anonymous referees

Submitted: November 24, 2022
Accepted: August 9, 2023
Proofs received from author(s): September 8, 2023