Received: 22 December 2021 | Revised: 6 May 2022 | Accepted: 13 July 2022 | Published on: 9 August 2022

DOI: 10.1002/joc.7822

RESEARCH ARTICLE

International Journal of Climatology RMetS

Assessing homogeneity of land surface air temperature observations using sparse-input reanalyses

Ian Gillespie[1] | **Leo Haimberger**[2] | **Gilbert P. Compo**[3,4] | **Peter W. Thorne**[1]

[1]ICARUS Climate Research Centre, Department of Geography, Maynooth University, Maynooth, Ireland

[2]Department of Meteorology and Geophysics, University of Vienna, Vienna, Austria

[3]CIRES, University of Colorado, Boulder, Colorado, USA

[4]Physical Sciences Laboratory, NOAA, Boulder, Colorado, USA

Correspondence
Peter W. Thorne, ICARUS, Department of Geography, Maynooth University, Maynooth, Ireland.
Email: peter@peter-thorne.net

Abstract

State-of-the-art homogenisation approaches for any test site rely upon the availability of a sufficient number of neighbouring sites with similar climatic conditions and a sufficient quantity of overlapping measurements. These conditions are not always met, particularly in poorly sampled regions and epochs. Modern sparse-input reanalysis products which are constrained by observed sea surface temperatures, sea-ice and surface pressure observations, continue to improve, offering independently produced surface temperature estimates back to the early 19th century. This study undertakes an exploratory analysis on the applicability of sparse-input reanalysis to identify breakpoints in available basic station data. Adjustments are then applied using a variety of reanalysis and neighbour-based approaches to produce four distinct estimates. The methodological independence of the approach may offer valuable insights into historical data quality issues. The resulting estimates are compared to Global Historical Climatology Network version 4 (GHCNMv4) at various aggregations. Comparisons are also made with five existing global land surface monthly time series. We find a lower rate of long-term warming which principally arises in differences in estimated behaviour prior to the early 20th century. Differences depend upon the exact pair of estimates, varying between 15 and 40% for changes from 1850–1900 to 2005–2014. Differences are much smaller for metrics starting after 1900 and negligible after 1950. Initial efforts at quantifying parametric uncertainty suggest this would be substantial and may lead to overlap between these new estimates and existing estimates. Further work would be required to use these data products in an operational context. This would include better understanding the reasons for apparent early period divergence including the impact of spatial infilling choices, quantification of parametric uncertainty, and a means to update the product post-2015 when the NOAA-CIRES-DOE 20CRv3 sparse input reanalysis product, upon which they are based, presently ceases.

KEYWORDS

homogenisation, reanalyses, temperature, trends

1 | INTRODUCTION

Homogenisation of land surface air temperature (LSAT) time series is essential prior to their use. Over multiple decades there are inevitable changes in multiple facets of a station series including local micro-environment, instrumentation, observers, methods of observation and so forth (Karl and Williams Jr, 1987; Quayle *et al.*, 1991; Peterson and Easterling, 1994; Guttman, 1998; Vincent, 1998; Bronnimann, 2015). Even if every effort is made to minimize the impacts, it is all but inevitable that nonclimatic data artefacts will be present in many station series, and that in very many cases such artefacts affect long-term trends (Caussinus and Mestre, 2004; Trewin, 2010; Lawrimore *et al.*, 2015; Domonkos and Coll, 2017; Hunziker *et al.*, 2017).

State-of-the-art homogenisation methods generally use a neighbour-based approach, typically based upon pairwise comparisons as summarized in Venema *et al.* (2012), to identify and adjust for breakpoints. All such approaches are predicated upon the availability of a sufficient pool of similar neighbour estimates and an assumption of noncoincidence of data artefacts (Caussinus and Mestre, 2004). Such assumptions cannot be guaranteed, and it is of value to explore alternative approaches which may better maintain independence between nearby stations. One such potential approach is to use sparse-input reanalyses which do not ingest or use LSAT measurements and yet provide dynamically and physically constrained estimates of LSAT (Trewin, 2010; Compo *et al.*, 2011; Compo *et al.*, 2013).

Gillespie *et al.* (2020) proposed that the most recent sparse-input reanalysis products are increasingly viable and credible candidates for use as reference series for the homogenisation of LSAT station series arising from the databank of the International Surface Temperature Initiative (ISTI) (Rennie *et al.*, 2014). They showed that the correlations and standard deviations of the monthly averaged difference series between the stations and sparse-input reanalyses products were similar to those between stations and their neighbours and highlighted particular potential benefits in data sparse regions and epochs.

This paper builds on that work by going on to apply the NOAA-CIRES-DOE Version 3 of the 20th Century Reanalysis (20CRv3) sparse-input reanalysis (Slivinski *et al.*, 2019; Slivinski *et al.*, 2021) to homogenize the ISTI databank holdings. The analysis builds upon the established methods applied to full-input reanalyses to homogenize radiosonde data records by Haimberger *et al.* (2012). These are modified for the particular circumstances of sparse-input reanalyses and land surface stations. It compares the results to NOAA NCEI's pairwise homogenisation algorithm (PHA) method (Menne and Williams, 2009) which was used to create GHCNMv4 from the same set of fundamental data holdings. It then goes on to compare globally aggregated results to the full suite of existing Global LSAT products.

The remainder of the paper is structured as follows. In section 2 the quality control and breakpoint detection steps are outlined. Section 3 summarizes and assesses the overall application of adjustments. The approach yields four estimates of the required adjustments which all use the same method to detect breakpoints but differ in how adjustments are calculated. Section 4 assesses the efficacy of the resulting homogenisation techniques by comparing adjustment behaviour, station series and gridbox anomalies and trends. In section 5 regional, hemispherical and global analyses are conducted. Section 6 compares the results at the global mean aggregation with estimates from other published LSAT products variously used in global monitoring and assessment activities. Section 7 includes a discussion of limitations, outstanding questions and potential next steps. Section 8 concludes.

2 | QUALITY CONTROL AND BREAKPOINT IDENTIFICATION

2.1 | Removal of gross outliers

In this analysis, quality control is not the intended focus. Nevertheless, it is necessary to remove gross outliers. QC is applied to the difference series between the station series anomalies from the ISTI databank and the matched 20CRv3 anomalies, both normalized to their common period-of-record as described in Gillespie *et al.* (2020). All points where the absolute differences are greater than three times the inter-quartile range were removed. This resulted in 15,537 stations that had one or more observations removed (Figure 1). The vast majority of stations have <1% of values removed although 11% have between 1 and 5% removed, with less than 0.6% having more than 5% removed.

2.2 | Breakpoint detection

The present analysis makes use of the same variant as Haimberger *et al.* (2012) of the standard normal homogeneity test (SNHT) (Alexandersson and Moberg, 1996) which forms the breakpoint detection component of many homogenisation algorithms in common use today including the PHA algorithm used in GHCNMv4 (Menne *et al.*, 2018) and the radiosonde work of Haimberger *et al.* (2012). The SNHT was applied to 20CRv3-station

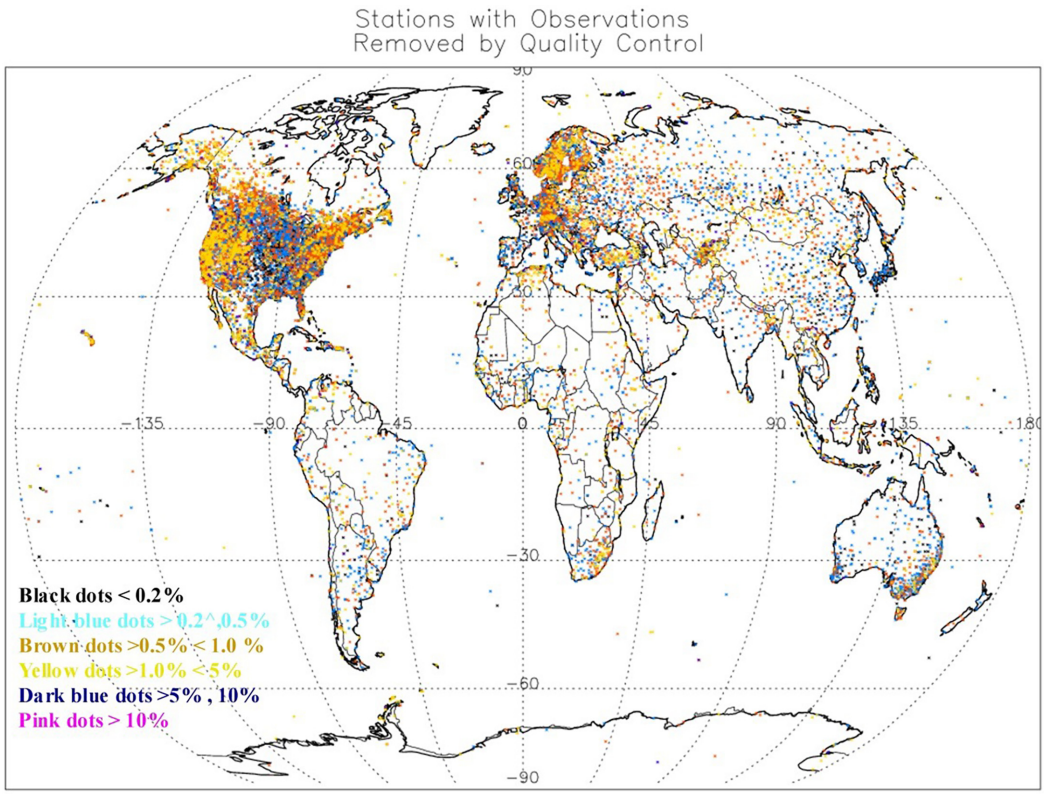

FIGURE 1 Map showing the station network used in the analysis and QC results. Higher QC failure rates overplot stations with lower failure rates in densely sampled regions [Colour figure can be viewed at wileyonlinelibrary.com]

difference series following Gillespie *et al.* (2020) and can be described mathematically as follows:

$$T_i = \frac{N}{S_i^2} * \left(\overline{X} - \overline{Z}\right)^2 + \left(\overline{Y} - \overline{Z}\right)^2, \quad (1)$$

where T_i is the SNHT score, N is the total number of observations, S_i is the pooled standard deviation of valid points before and after the potential break, $\overline{X}$ is the mean of valid points before the break, $\overline{Y}$ is the mean of valid points after the break and $\overline{Z}$ is the pooled mean of valid points before and after the break. The test is typically applied iteratively to consecutive segments of the series of equal intervals. In the present analysis, the segment length either side of the tested point is set to 5 years (60 monthly values) such that the statistic is returned for points only within the segment bounded by the first and last 5 years of series availability. Following Haimberger *et al.* (2012) a missing mask matching is applied so that each tested segment contains the same number of points and seasonality of sampling is identical. That is, if a specific month in segment one is missing the corresponding month in segment two is set to missing and vice versa. This is critical if data biases have substantive seasonal signatures. The test is applied only if more than 20 points still remain in both segments.

When applied to series exhibiting autocorrelation, the SNHT test has no universally recognized critical threshold. In the present analysis, the SNHT test was run for critical values ranging from 6 to 20 on all stations at intervals of 2 in an attempt to determine the most suitable critical value. The distribution of implied segment adjustments and cumulative adjustments using the average of the station minus 20CRv3 difference series for each segment were examined (not shown for brevity, see Gillespie, 2021). As expected, the number of breakpoints detected decreases with increasing critical value. Critical thresholds of 12–16 are equivalent to a break every 15–20 years and would be consistent with the typical frequency of breakpoints reported in GHCNMv4 (Menne *et al.*, 2018). However, benchmarking exercises (Venema *et al.*, 2012; Williams *et al.*, 2012) highlight that all current algorithms tend to underestimate the number of breakpoints in synthetic test series. It is reasonable to assume that this also extends to the real-world. In the absence of other criteria, a conservative critical value of 16 was selected for the present analysis while recognizing that this will miss some real-world breakpoints.

Finally, long gaps in the digital record, unless resulting solely from poor records management, will typically be associated with a change of one or more of: station observer, instrumentation, or station location. However, the SNHT, at least as applied, could not detect breaks of this nature. To account for this, the test statistic has been manually set to a very large value (99.9) at each resumption following a break of >36 contiguous months, thus forcing a breakpoint to be assigned.

2.3 | Break assignment

Breaks are only assigned if three or more consecutive values exceed the SNHT critical value threshold to minimize false-positives. Breakpoints are associated with the timing of the maximum test statistic value attained within such a contiguous string. To further minimize the effects of time series noise, if two such breaks are assigned within a single 12-month interval only the largest of the pair is retained.

Figures 2–4 illustrate examples of applying the algorithm to three selected stations from data sparse regions where the algorithm may have the most value over traditional pairwise homogenisation methods (Gillespie *et al.*, 2020). For a broader view, we also include De Bilt in the Netherlands, as an example from a highly sampled region (Figure 5).

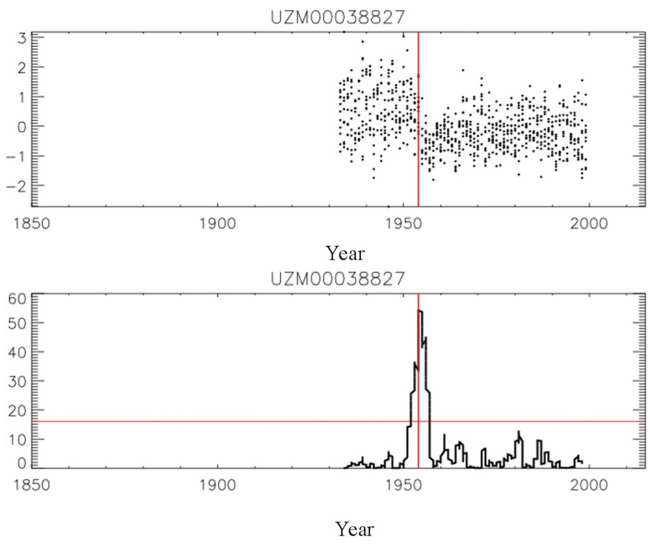

FIGURE 2 Station at Baisun from Uzbekistan (38.2°N, 67.2°E, 1,241 m.a.s.l) with 803 observations over December 1932 until October 1999 with a single break assigned in November 1954. The top panel shows the station minus 20CRv3 difference series where each monthly value is plotted as a dot. The lower panel shows the SNHT scores trace with the threshold denoted by the horizontal red line and the break location returned, denoted by a vertical red line [Colour figure can be viewed at wileyonlinelibrary.com]

In Figure 2, a single detected break exists in the series which is visually obvious and associated with a large exceedance of the critical threshold for the SNHT test. The assigned breakpoint is in good accordance with the apparent break location from visual inspection. Figure 3 shows a much more complicated example of a long-running station series for which multiple breaks have been assigned. In this case, the visual basis for each break is less clear, but some of the breaks are apparent and, for these, the assigned locations seem reasonable. Figure 4 shows a case with a break in the series availability, over the second world war, and highlights how the forced insertion of a breakpoint upon resumption ensures an adjustment will be estimated. Visually there is a potential minor discontinuity between the series before/after the world war. Figure 5 has two detected breaks over 1895–1905. The break around 1900 appears to be associated with a change in not just the mean but also the variance of the series (and is implicitly associated with the splicing of a nearby series into the record according to the metadata, see Gillespie *et al.*, 2020), although the SNHT test is only able to detect the mean shift aspect.

Visual inspection of several hundred additional series confirms the results shown in the above examples. At an SNHT score threshold of 16, the SNHT algorithm appears to detect obvious breaks in the series and does not obviously overestimate break occurrence.

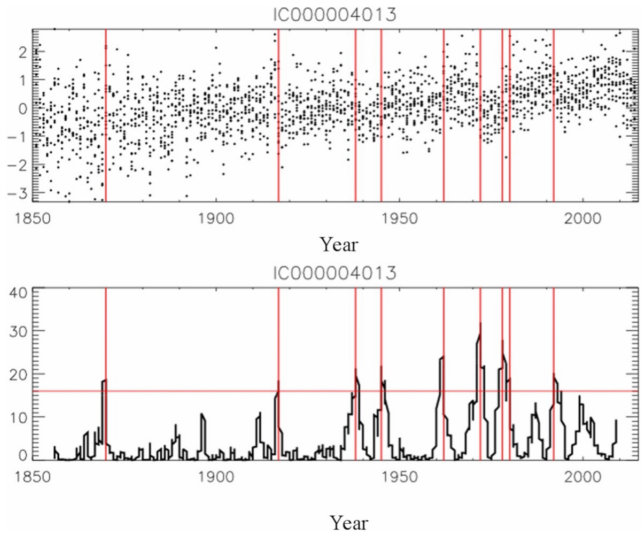

FIGURE 3 As figure but for Stykkisholmur, western Iceland (65.073°N, 22.725°W, 15 m.a.s.l). The site has 2051 observations, commencing before January 1851. The site has minor data gaps from August to December 1921 and between December 1940 and April 1941 that are not clearly visible. Breakpoints were detected in December 1869, January 1917, September 1938, September 1945, February 1962, February 1972, July 1978, February 1980 and November 1992 [Colour figure can be viewed at wileyonlinelibrary.com]

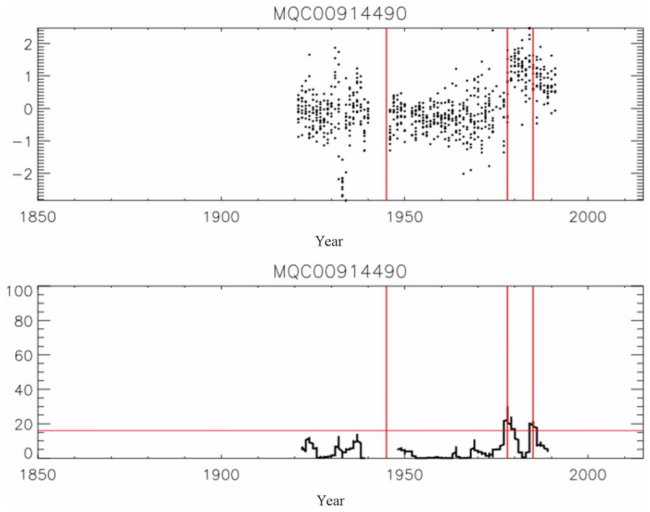

FIGURE 4 As Figure 2 but for a site on Midway Sand Island at Midway in the Pacific Ocean (28.217°N, 177.35°E, 3 m.a.s.l.). The site has 701 monthly observations commencing in December 1920 until August 1991. Note the gap in the observations from December 1940 to December 1945 (month 1,080–1,140) over WW2. Breakpoints were detected in April 1978 and at May 1985 in addition to the assignment of a breakpoint upon time series resumption after WW2 [Colour figure can be viewed at wileyonlinelibrary.com]

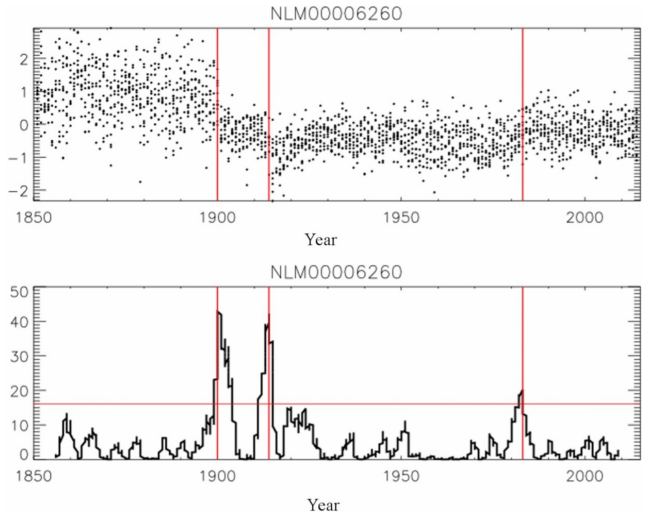

FIGURE 5 As Figure 2 but for De Bilt (52.1014°N, 5.1867°E 2 m.a.s.l). As per Figures 4–6 the series from January 1851 to December 2014 is shown only. However, the De Bilt site time series extends further back in time than January 1851, the period under examination for this work [Colour figure can be viewed at wileyonlinelibrary.com]

3 | APPLICATION OF ADJUSTMENTS

Adjustments are applied to all points preceding each identified breakpoint in a series. They are applied progressively backwards such that the resulting series mean state, if the adjustments are adequate, should become homogeneous relative to the final identified homogeneous segment. Adjustments, irrespective of the approach, are applied as seasonally invariant mean shifts. No attempt is made to adjust for any variance effects or seasonality effects, although such artefacts undoubtedly remain in some series (e.g., Figure 5). The adjustment methods are based upon the RAOBCORE and RICH approaches described in Haimberger *et al*. (2012) and references therein. RAOBCORE uses the station minus reanalysis background forecast series directly, whereas RICH uses statistical characteristics of apparently homogeneous neighbour segments. We modify those procedures to create four distinct adjustment estimates: two based upon RAOBCORE and two based upon RICH.

3.1 | 20CRv3$_{long}$

20CRv3$_{long}$ is based upon RAOBCORE but using a reanalysis estimate directly rather than a background forecast. In RAOBCORE, the background forecast must be used because the reanalysed field assimilates the target data and its use would introduce an overt circularity. In 20CRv3 surface temperature observations are not assimilated, and thus the reanalysis estimate is independent. Arguably because it never assimilates the temperature observations from any station it is more independent than the RAOBCORE approach where the background contains some residual information from prior observations at the site being homogenized. The same difference series as was used to determine the break locations is used to estimate segment mean adjustments. The full segment series irrespective of length between the test break and its predecessor break (or the start of the series) and the test break and its successor break (or the end of the series) is used to estimate the means prior to and after the break to infer the required adjustment. The difference in segment means (after minus prior) is sequentially added backwards to all valid data points prior to the current breakpoint in the series.

3.2 | 20CRv3$_{short}$

20CRv3$_{short}$ is identical to 20CRv3$_{long}$ except that segments are cut if longer than 5 years. If there is long-term drift in the 20CRv3 reanalysis product at the station location then this minimizes the impact that drift can have on the adjustment estimates and the resulting homogenized series. However, use of short segments may introduce noise in the resulting series because, all else being

equal, segment means will be more uncertain owing to the smaller sample sizes available to estimate the true mean value of each segment.

3.3 | Neighbour$_{segments}$

The neighbour$_{segments}$ adjustment procedure is broadly based upon the RICH$_{obs}$ method of Haimberger *et al.* (2012). A search is made at each breakpoint identified in the candidate series through each of the 250 nearest neighbours, irrespective of direction or correlation, defined by a simple great circle distance search. Each neighbour that contains sufficient data within ±5 years of the breakpoint and itself does not have identified breakpoints within 20 months (chosen to ensure an adequate sample from which to derive an estimate of the true mean) either side of the breakpoint is used to create an adjustment estimate (with the segment if necessary cut at the break in the neighbour). The difference series between the target station anomalies and the neighbour series anomalies (both calculated relative to their own station series availability to maximize station retention) is used as the basis for this estimate (again using the difference in means after minus prior). There must be at least 20 valid data points prior to and after the break to estimate an adjustment. Differences in station data availability lead to slight station-to-station climatology differences which are assumed sufficiently small as to be unimportant and when averaged over a sufficient sample to become pseudo-random.

Assuming one or more neighbour-based estimates are returned, the median of the individual estimates is applied as the adjustment. This minimizes the impacts of any individual outlier estimates when sample sizes are sufficiently large. If no estimates are available then the 20CRv3$_{short}$ estimate is used in its place under the assumption that this is preferable to no adjustment being applied. This occurs for 5,154 breakpoints across a total of 4,986 stations. The overall proportion of instances falling back on 20CRv3$_{short}$ is just under 10% of all identified breaks. Some of these may arise because there are fewer than 20 months between consecutive breaks in the target series, but in such cases 20CRv3 is assumed to be a better basis for adjustment. Figure 6 illustrates that the propensity of deferral to 20CRv3 generally increases back in time as the availability of neighbours decreases.

Figure 7 shows maps of stations that deferred at least once to 20CRv3 to homogenize a break. These maps show some surprises, particularly in North America post 1950 (lower right map). There are a substantial number of stations deferred to 20CRv3 at least once. This appears to relate to the quasi-contemporaneous change to MMTS sensors across the COOP network (Quayle *et al.*, 1991; Hausfather *et al.*, 2016) whereby no suitable neighbours that were themselves homogeneous presumably remained.

3.4 | Neighbour$_{double\text{-}diffs}$

Neighbour$_{double\text{-}diffs}$ is broadly based upon the RICH$_{tau}$ method of Haimberger *et al.* (2012). It differs from

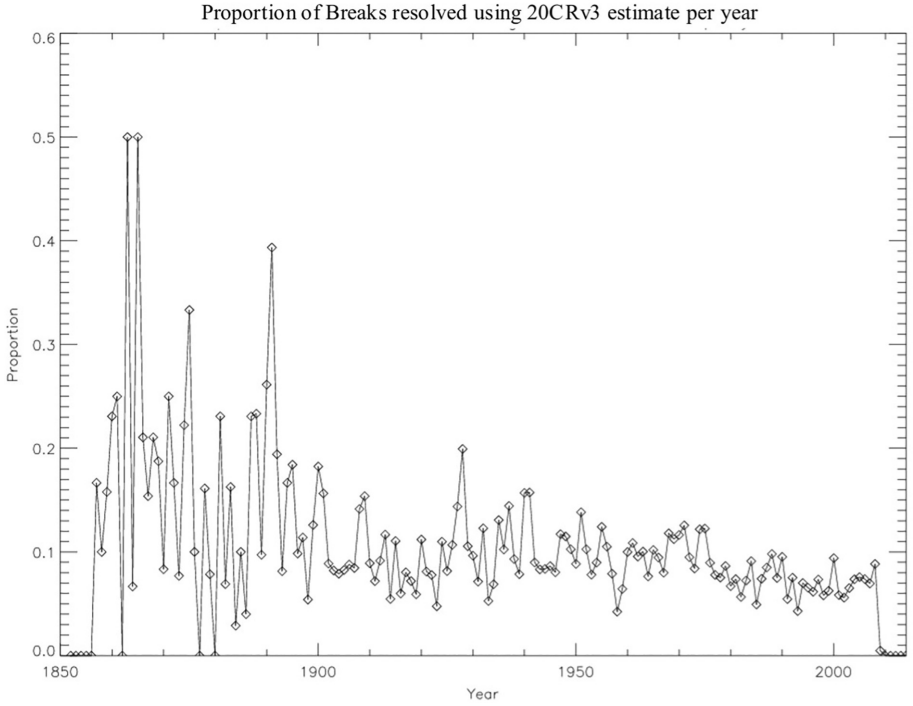

FIGURE 6 Proportion of deferral to 20CRv3 for homogenisation with time as a proportion of total break counts in each given year (which increases substantially as the station density increases after 1950). Note that no breakpoints are detected and hence no adjustments applied in the first and last 5 years of the series

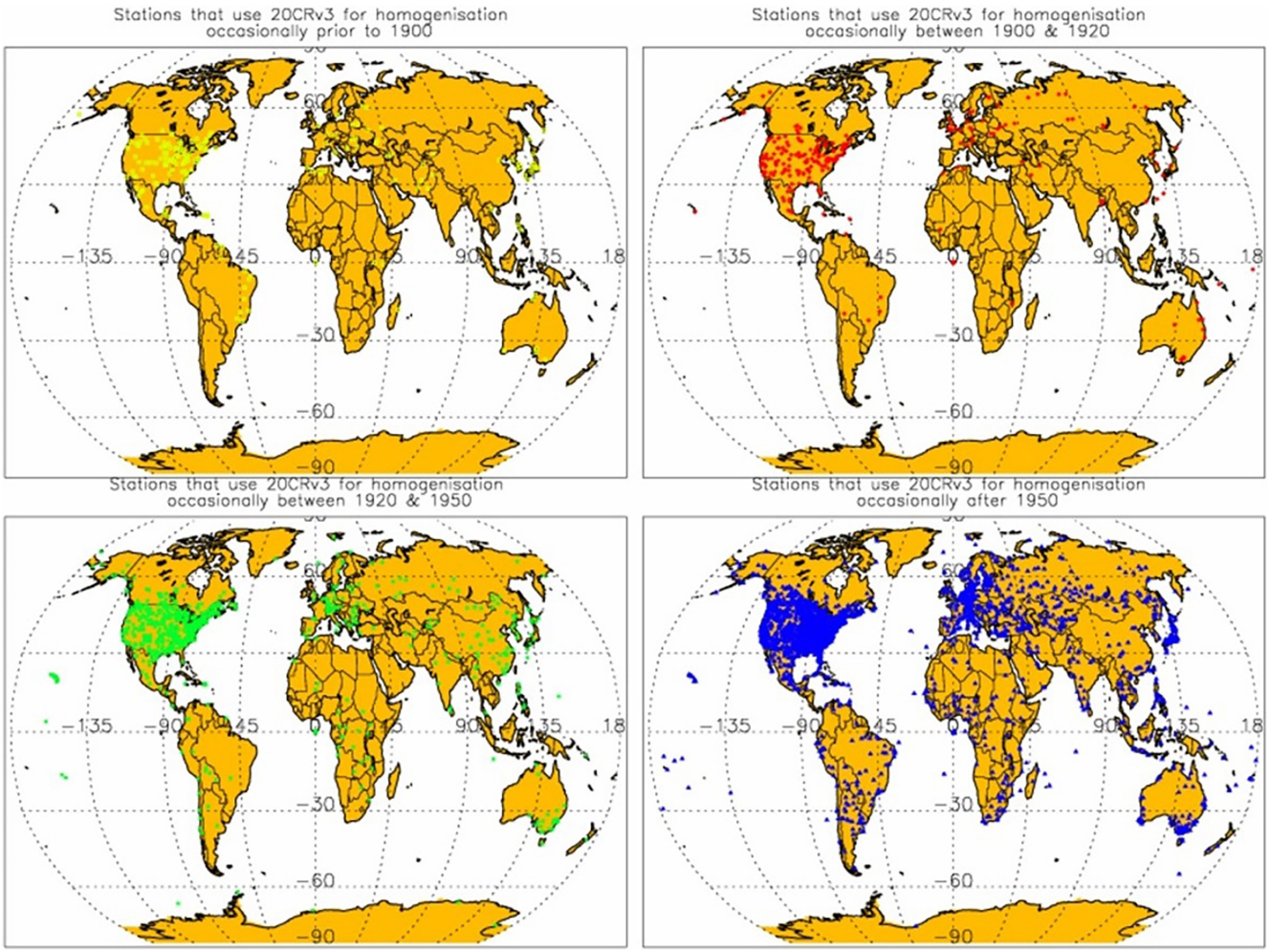

FIGURE 7 Map showing stations the deferred to $20CRv3_{short}$ for homogenisation using the two neighbour-based approaches before 1900 (top left), between 1900 and 1920 (top right), between 1920 and 1950 (bottom left) and after 1950 (bottom right) for the homogenisation of at least one identified breakpoint [Colour figure can be viewed at wileyonlinelibrary.com]

$neighbour_{segments}$ in that it uses the differences between the station-20CRv3 series for the target and neighbour stations. This is methodologically broadly similar to double differencing techniques used widely in Numerical Weather Prediction (Kanamitsu *et al.*, 1991; Saha *et al.*, 2010). The assumption is that although the source being used (in this case 20CRv3 reanalysis fields) may be biased, this bias varies smoothly. Taking the difference of the differences between reasonably proximal locations removes the common bias component in the reference source leaving a "purer" estimate of the true instrumental differences assuming that the source reasonably approximates true geophysical gradients and variations.

Other than the series being used, all other details of $neighbour_{double\text{-}diffs}$ are identical methodologically to those used in $neighbour_{segments}$. This includes the defaulting to the use of $20CRv3_{short}$ as the basis for adjustments when a neighbour-based estimate is unavailable.

3.5 | Adjusting climatology to 1961–1990

Station time series up to and including adjustment have been used either as actuals or as anomalies relative to their own data availability. This serves to maximize the station count that can be used in these steps. However, to further analyse the long-term series and perform aggregations to regional and global series requires the application of a consistent climatology. The method employed herein is to calculate a climatology based on 1961–1990. However, only approximately 11,000 of the just over 27,000 adjusted stations have sufficient data available during 1961–1990 to calculate a climatology for that period directly where criteria for inclusion are at least 20 years of data in each calendar month over the period. To incorporate remaining stations, temporally complete station-equivalent timeseries estimates were calculated from 20CRv3 for the full period of January 1851 to

December 2014 for each such station location. The 1961–1990 climatology and the climatology matched to the station availability over 1850–2014 were calculated for 20CRv3. The difference between these two estimates was then subtracted from the station anomalies to normalize the homogenized data series to 1961–1990. To check that this step did not unduly bias the analysis, several subsequent analysis steps were undertaken using both all station series and the subset from which a climatology could be directly calculated and the results compared (section 7).

4 | ASSESSING THE EFFICACY OF THE APPROACH

4.1 | Characterization of adjustments

Applying the SNHT with a critical value 16 leads to 58,325 breakpoints being identified. Of the 27,639 series analysed, 8,398 had no breaks and 1,909 had breaks only associated with missing data segments, with most remaining stations having 1–3 breaks (Table 1). There is a relatively long-tailed distribution with a total of 350 stations containing 10 or more identified breakpoints and a preponderance for these to be centennial-scale station series. The maximum number of breakpoints found in a single station is 19 at station GM000001474 from Bremen, Germany.

TABLE 1 Summary of the preponderance of breakpoint detection at an SNHT critical value of 16 across the raw ISTI databank stations retained following the analysis described in Gillespie *et al.* (2020)

Number of breaks identified	Number of stations
0	8,398
Data gaps breaks of >36 continuous months only	1,909
1	5,915
2	4,292
3	2,944
4	1,464
5	919
6	598
7	416
8	274
9	160
10+	350
Total	27,639

Note: This count includes cases where a breakpoint has been forced to account for a gap of 36 months or longer duration.

The resulting set of adjustment estimates applied at each breakpoint are summarized in Figure 8 for the four adjustment approaches. All four histograms contain an identical total number of adjustment estimates by construction. The four distributions are distinct from one another, the neighbour$_{segment}$ based approach has a feature not seen in the other three in that it contains no dip centred close to zero adjustment size—a feature present in all remaining distributions and also the PHA technique as applied to GHCNMv4 (Menne *et al.*, 2018). This so-called "missing middle" dip is largest in 20CRv3$_{short}$ adjustments. This effect arises because the "missing middle" is inherently a methodological result of being unable to detect the small breaks owing to signal to noise limitations. It follows that the further methodologically the detection and adjustment steps are from one another based upon differing data and/or time-windows the more this "missing middle" feature would a *priori* be infilled.

The neighbour$_{segment}$ approach also has the largest standard deviation and more large adjustments (fatter tails) than the remaining three techniques. The neighbour$_{double\text{-}diffs}$ technique shows the next highest standard deviation. The lowest standard deviation is from the 20CRv3$_{long}$ adjustment technique. All four techniques show a very slightly negative mean adjustment of between −0.011 and −0.022°C, which are comparable to the mean adjustment of −0.023°C reported for GHCNMv4 (Menne *et al.*, 2018).

Overall, despite some differences in the adjustment characteristics between the different techniques, none appear to be obviously unreasonable approaches based upon the distribution of applied adjustments. However, the distinct behaviour of the neighbour$_{segment}$ technique, with many adjustments close to zero, means some caution may be warranted in its application.

4.2 | Evaluation of impacts on individual station series

Having ascertained that in aggregate the four adjustment approaches appear reasonable, next the impacts on individual station series must be ascertained. Recourse is made also to the GHCNMv4 neighbour based adjusted series returned by NCEI's PHA algorithm (Menne and Williams, 2009). For illustrative purposes the adjusted series are further adjusted to be equal over the final homogeneous segment and then compared. This better highlights the time-varying nature of adjustments than comparing series normalized to 1961–1990, as well as illustrating the effectiveness of the intent of homogenisation to make all segments comparable to the most recent (and ongoing in operational stations) segment. The purpose of

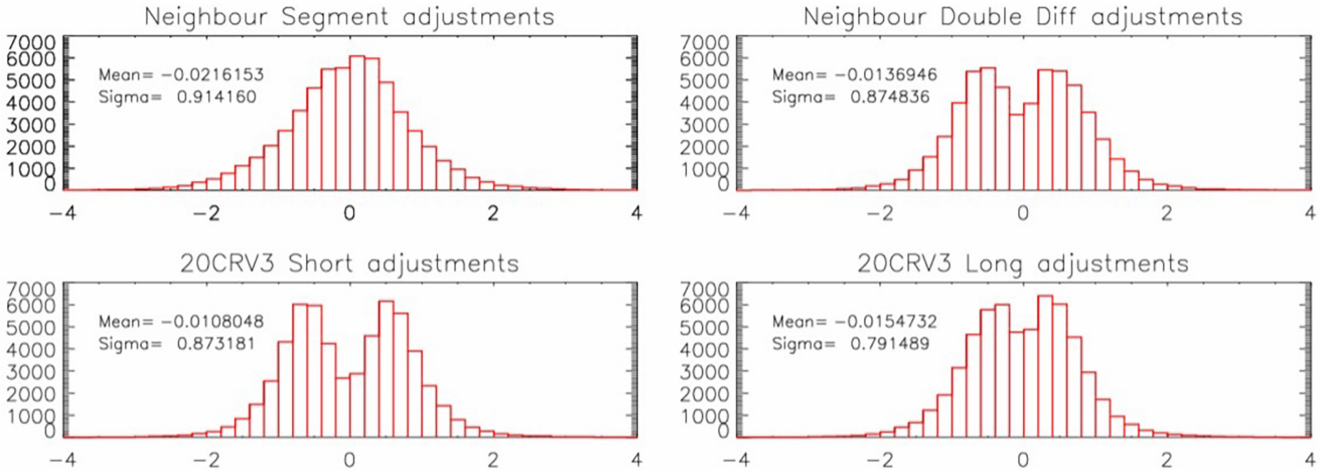

FIGURE 8 Distribution comparison of adjustments for the four adjustment approaches using an SNHT critical value of 16 [Colour figure can be viewed at wileyonlinelibrary.com]

this analysis is simply to determine overall reasonableness of the adjustment approaches via comparison. New benchmarking studies (e.g., Venema *et al*., 2012; Williams *et al*., 2012) may permit a more absolute characterization were they to include 20CRv3 as their background field, but are beyond the scope of the present analysis.

For the station with a single-break (see Figure 2), all four homogenisation methods produce a series that reduces the mean shift, as does GHCNMv4 which identified the same breakpoint (Figure 9). None of the techniques (including GHCNMv4), by construction, is able to deal with the obvious associated shift in variance. Figure 10 shows that all techniques adjust the early period of record to be cooler than the original record. This adjustment is visually obvious, and the various adjustment estimates are in broad agreement, appearing to improve station series homogeneity.

Moving on in complexity to Stykkiosholmur in western Iceland which had multiple breaks identified (see Figure 3), the resulting adjustments show considerable spread (Figure 11) that increases back in time. The difference in adjustments is acute prior to a breakpoint identified in the early 1960s and for a breakpoint identified in the late 1930s. Multidecadal variability, presumably driven by the Atlantic multidecadal oscillation (Knight *et al*., 2006; Allison *et al*., 2014; Yang *et al*., 2020) remains in all the resulting series. Whether, and if so how much, this location has warmed since the mid-19th century is highly sensitive to the choice of adjustment approach. GHCNMv4 tends to stay closest to the original series. Both neighbour-based adjustment approaches show greater spread in adjustments for certain periods than remaining estimates. Given the relative remoteness of this Icelandic site, this behaviour is perhaps unsurprising and highlights the potential value of using sparse-input reanalyses not just to detect but also to adjust for breakpoints in such regions. Conversely, how well SST and sea-ice variations in the early record are recreated may impact the fidelity of 20CRv3 in this region (Przybylak *et al*., 2016).

For the Midway Island site with a long duration cessation of operation over WW2 (see Figure 4), GHCNMv4 has removed the entire pre-WW2 segment (Figure 12) whereas all four of the adjustment approaches herein retained this segment. Pre-WW2 all the solutions are similar and systematically warmer than the raw data. The degree of adjustment of the overall series varies widely across solutions with GHCNMv4 clearly identifying at least one apparent breakpoint not detected by the present approach (a systematic shift in mid-1950s). GHCNMv4 also either (a) has a much larger adjustment estimated for the brief homogeneous segment between the late 1970s and mid-1980s than any of the four solutions developed herein; or (b) did not detect and adjust for a break here.

Turning attention to De Bilt (see Figure 5) which, unlike the other three stations considered thus far, is in a well sampled region of the globe with several centennial neighbouring sites available for its full period of record, even if they are at some distance, all methods are in broad agreement. Again, by construction, none of the methods accounts for variance effects although GHCNMv4 appears to do so to some extent (Figure 13). The individual time series plots are all similar to one another (Figure 14) with differences being considerably smaller than the inter-annual variability. It appears that GHCNMv4 captured broadly the same breakpoints as the present method with the exception of a period around WW2 when two additional breakpoints may have been assigned by GHCNMv4. Pre 1900 all solutions continue

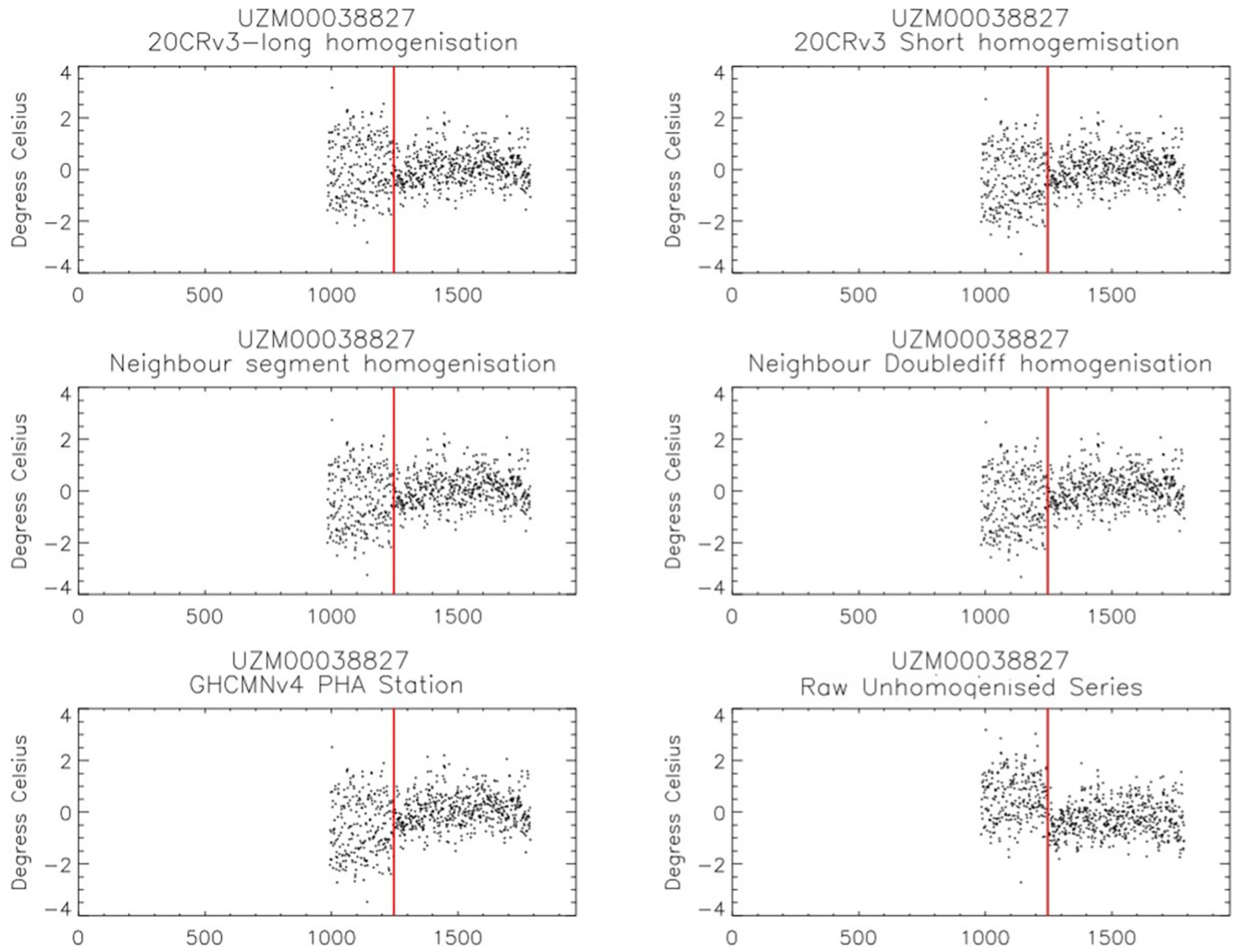

FIGURE 9 The resulting set of 20CRv3-station difference series for station UZM000038827 (top left panel 20CRv3 long; top right panel 20CRv3 short; middle left panel neighbour segment; middle right panel doublediff; bottom left panel GHCNMv4 adjusted; bottom right panel raw unadjusted series is reproduced from Figure 5). The single breakpoint location identified in the present analysis is denoted by the solid red vertical line. Individual monthly values are shown with months indexed from January 1851 [Colour figure can be viewed at wileyonlinelibrary.com]

to agree closely but adjust to be substantively cooler than the raw data, by of the order 1°C.

The above examples are broadly indicative of a much larger, but still far from complete, sample of stations considered manually. These include intermittent stations, stations with large numbers of breaks and from a range of regions (Gillespie, 2021). Overall, the approaches appear reasonable. Where there are potential issues, they tended to occur more prominently in the two neighbour-based adjustment approaches, but not always so. Overall, it would be hard to consistently question the value of any of the approaches in a manner which may lead to their rejection as a scientifically reasonable approach. Furthermore, qualitatively the resulting adjusted series in most cases (including many shown in preceding figures) correspond better with the neighbour-based GHCNMv4 analysis than with the original raw data, particularly so in well-sampled regions where PHA is expected to perform best, and in those cases where breaks in the raw data are visually obvious. Given the rich heritage of the PHA technique, the similarity of station series adjustments builds some confidence in the verity of the present method. The four adjustment methods do, however, show considerable spread in some cases (e.g., Figures 11 and 12) justifying analysis as distinct possible approaches to series adjustment.

4.3 | Spatial anomalies

Having ascertained that all techniques appear to provide reasonable adjustments at the station level, the

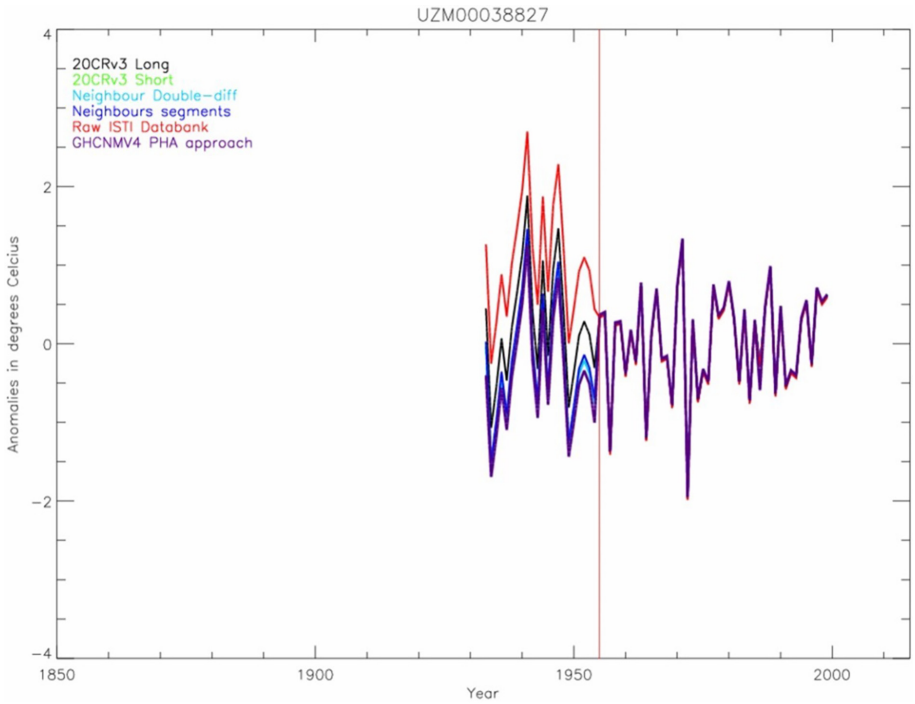

FIGURE 10 Annual time series of anomalies following application of adjustments (except for the raw series) and renormalisation to a 1961–1990 climatology followed by matching all series to be identical for the final homogeneous portion for illustrative purposes. Locations where breakpoints have been assigned and thus adjustments applied are denoted by solid red vertical lines [Colour figure can be viewed at wileyonlinelibrary.com]

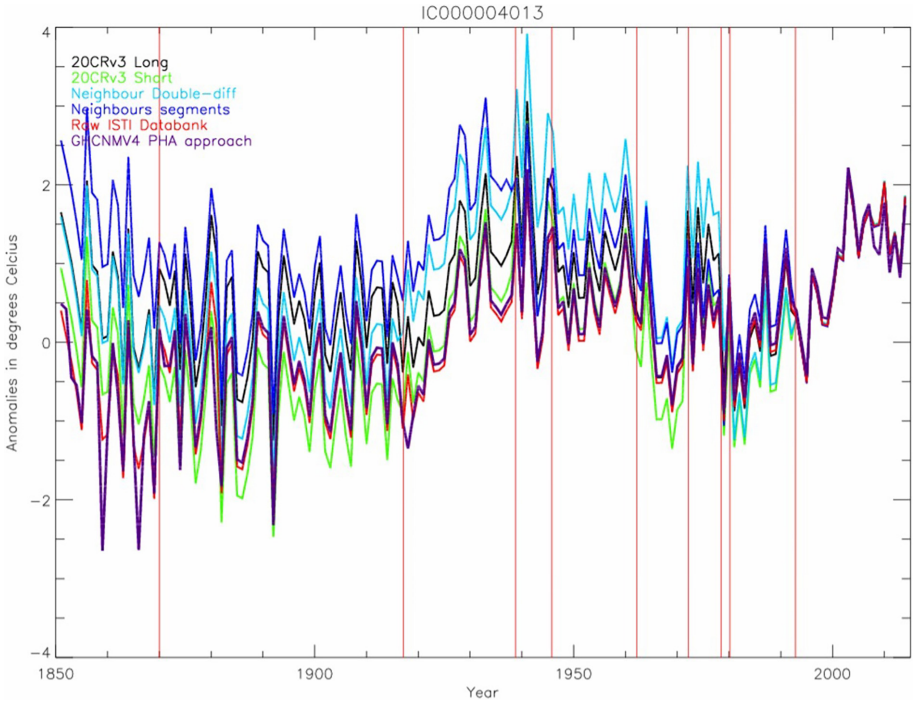

FIGURE 11 As Figure 10 but for Stykkiosholmur western Iceland. Differences in the final homogeneous segment relate to QC differences for raw and GHCNMv4 which alter some annual values [Colour figure can be viewed at wileyonlinelibrary.com]

resulting series were gridded for the "raw" data, GHCNMv4 and all four adjustment techniques. This gridding used simple grid box binning of 1961–90 normalized series (section 3) to a 5° by 5° resolution with no attempt made at interpolation. There is a degree of mismatch between the GHCNMv4 station availability and that in the four adjusted versions (Figure 15). Stations present in GHCNMv4 but absent in the new estimates arise from the analysis described in Gillespie *et al.* (2020). Stations present in the current analysis that are not included in the GHCNMv4 dataset arise due to not meeting GHCNM inclusion criteria. This analysis only includes those stations present in the current analysis and so excludes those stations with an estimate available solely in GHCNMv4. But the inverse matching has not been applied (as we did not wish to

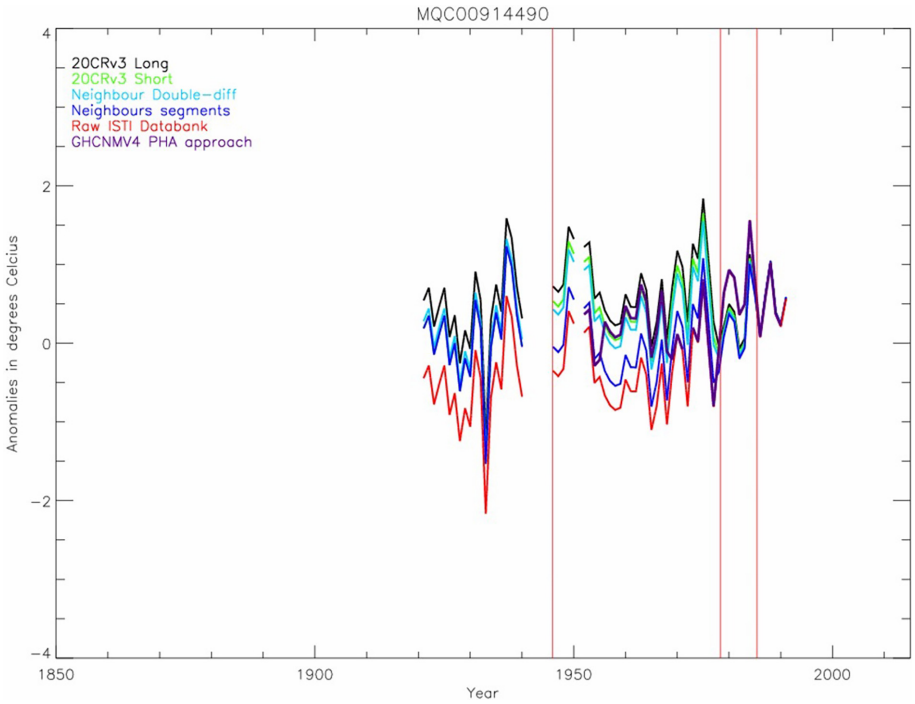

FIGURE 12 As Figure 10 but for Midway Island [Colour figure can be viewed at wileyonlinelibrary.com]

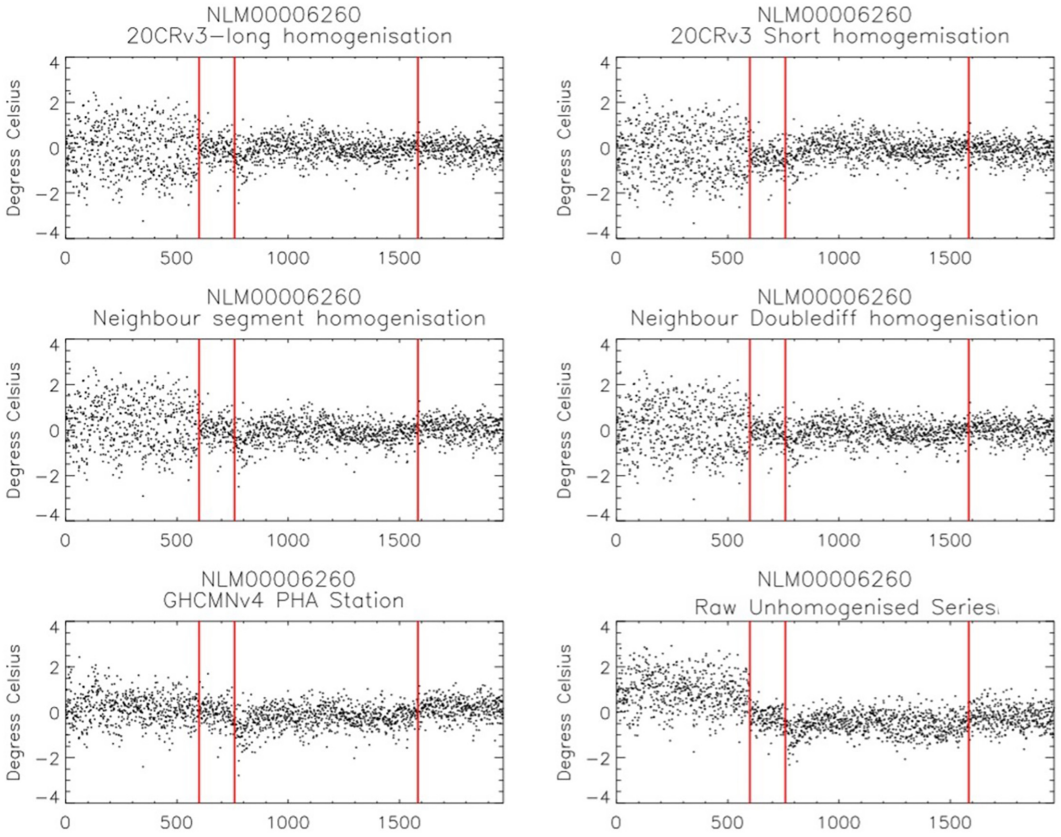

FIGURE 13 As Figure 9 but for station NLM00006260 De Bilt, Netherlands [Colour figure can be viewed at wileyonlinelibrary.com]

degrade the analysis of these new estimates), explaining some missing grid boxes in the GHCNMv4 maps that follow.

Anomaly fields were visually compared for all four homogenized series produced herein, GHCNMv4 and the raw data across a broad sample of months (Figure 16

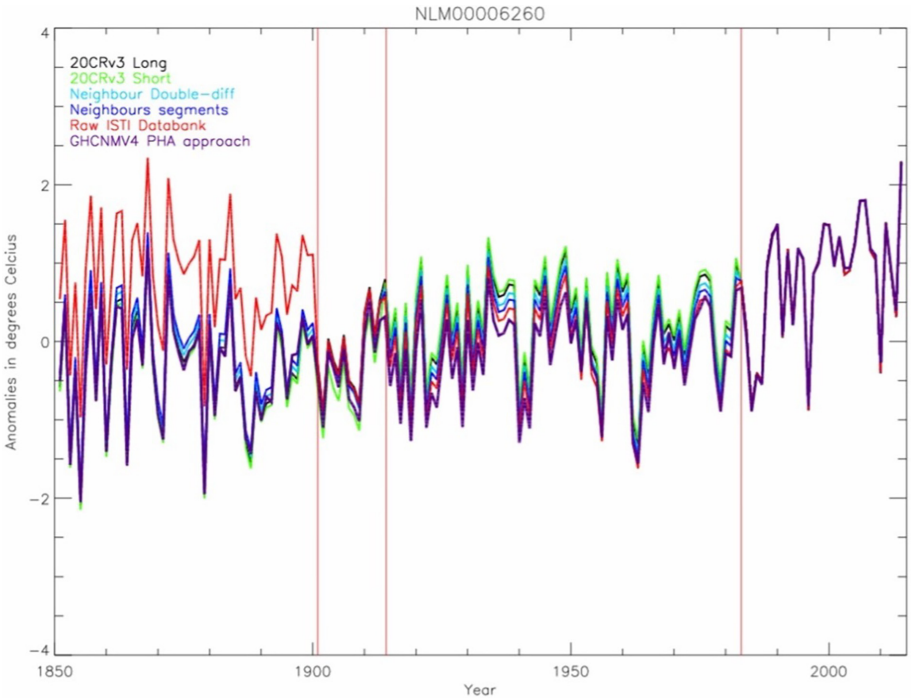

FIGURE 14 As Figure 10 but for De Bilt, Netherlands. Note that this series extends back further than 1850 but the assessment of homogeneity herein has been truncated to 1850 so the series is accordingly truncated here [Colour figure can be viewed at wileyonlinelibrary.com]

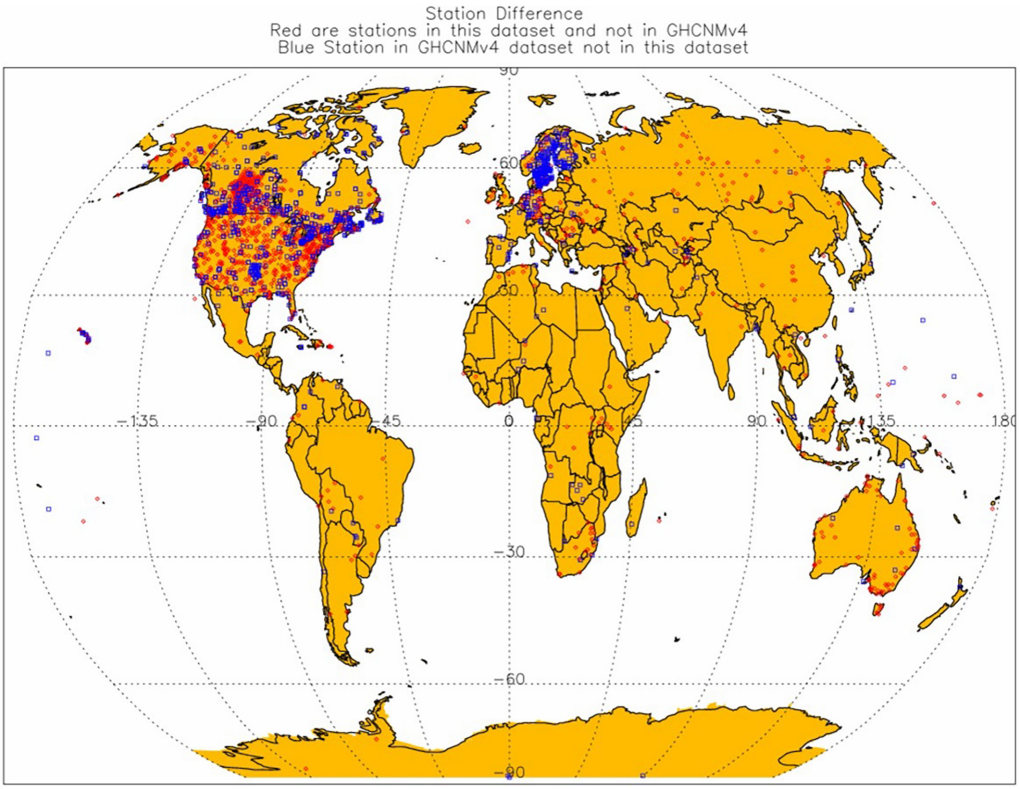

FIGURE 15 Summary of differences in station inclusion between GHCNMv4 and the present analysis. Red stations are present only in the current analysis. Blue stations are present only in GHCNMv4 [Colour figure can be viewed at wileyonlinelibrary.com]

shows June 1970). Individual monthly anomalies are large relative to the long-term trends and thus any apparent impact on monthly gridbox anomalies would be a cause to seriously question the efficacy of one or more of the adjustment approaches. In general, distinctions between the products are smaller than the colour-scale

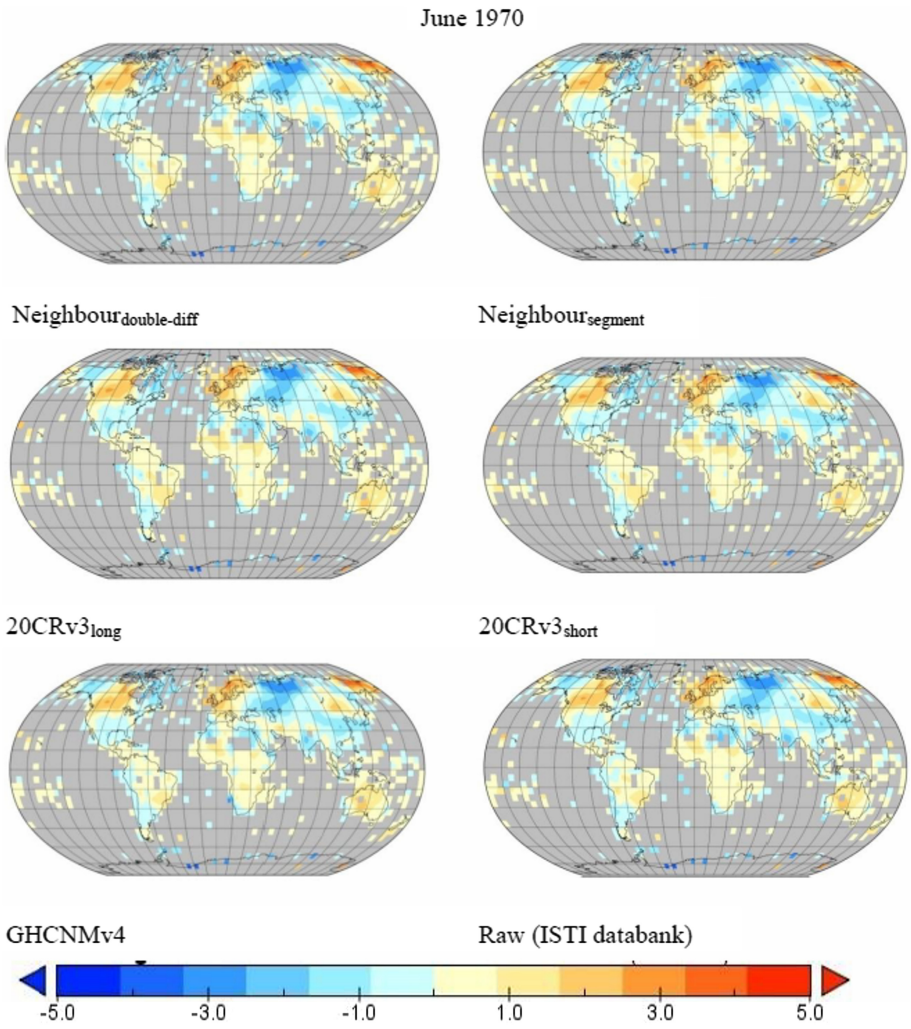

FIGURE 16 Maps of June 1970 gridbox anomalies from a 1961–1990 climatology for (from top left to bottom right) double differencing, neighbour segment, 20 CRv3 long, 20CRv3 short, NOAA NCEI's GHCMNv4 product and the original raw ISTI data bank holdings. Plot produced using Panoply version 4.10.12 for windows [Colour figure can be viewed at wileyonlinelibrary.com]

resolution necessary to span the full range of anomalies. The impact of a particular data product upon large-scale monthly mean anomalies is thus small relative to monthly variance. The monthly anomaly field behaviour does not give rise to any concerns about the methods.

4.4 | Spatial trends

Having ascertained that there are minimal differences in individual monthly anomaly fields, next spatial trends are considered. Differences arising from homogenisation choices, which act as red noise, should project much more strongly onto spatial trends than onto individual monthly anomaly fields.

Starting with the very longest intervals from 1851 to 2014, for which relatively few grid boxes contain sufficiently complete records (Figure 17), over Europe and Russia, the trend estimates are all broadly similar both in terms of the magnitude and the significance of the inferred trends. Most long-term stations exist in this region and so many grid boxes are constrained by multiple station estimates. Over North America, there are somewhat larger differences between the four estimates produced herein, and also with GHCNMv4. There are fewer stations available to inform grid box estimates in general in this region and so the impacts of the different homogenisation algorithms would be expected to be larger. To varying extents, neighbour$_{double-diff}$, neighbour$_{segment}$ and 20CRv3$_{short}$ lead to trend estimates with considerably less warming or even long-term cooling in the continental interior. They also find far fewer areas exhibiting significant trends than does GHCNMv4. Outside these two regions, the differences in the handful of individual gridboxes, which generally will consist of single station series in the early portions of their record, show few systematic differences.

There are far more grid boxes for which trends can be inferred for the 1900 to 2014 period allowing a much more exhaustive consideration of sensitivity to homogenisation choices (Figure 18). Trends over Europe are, again, largely consistent across all five estimates both in terms of trend magnitude and trend significance. However, 20CRv3$_{long}$ suggests less warming than the other

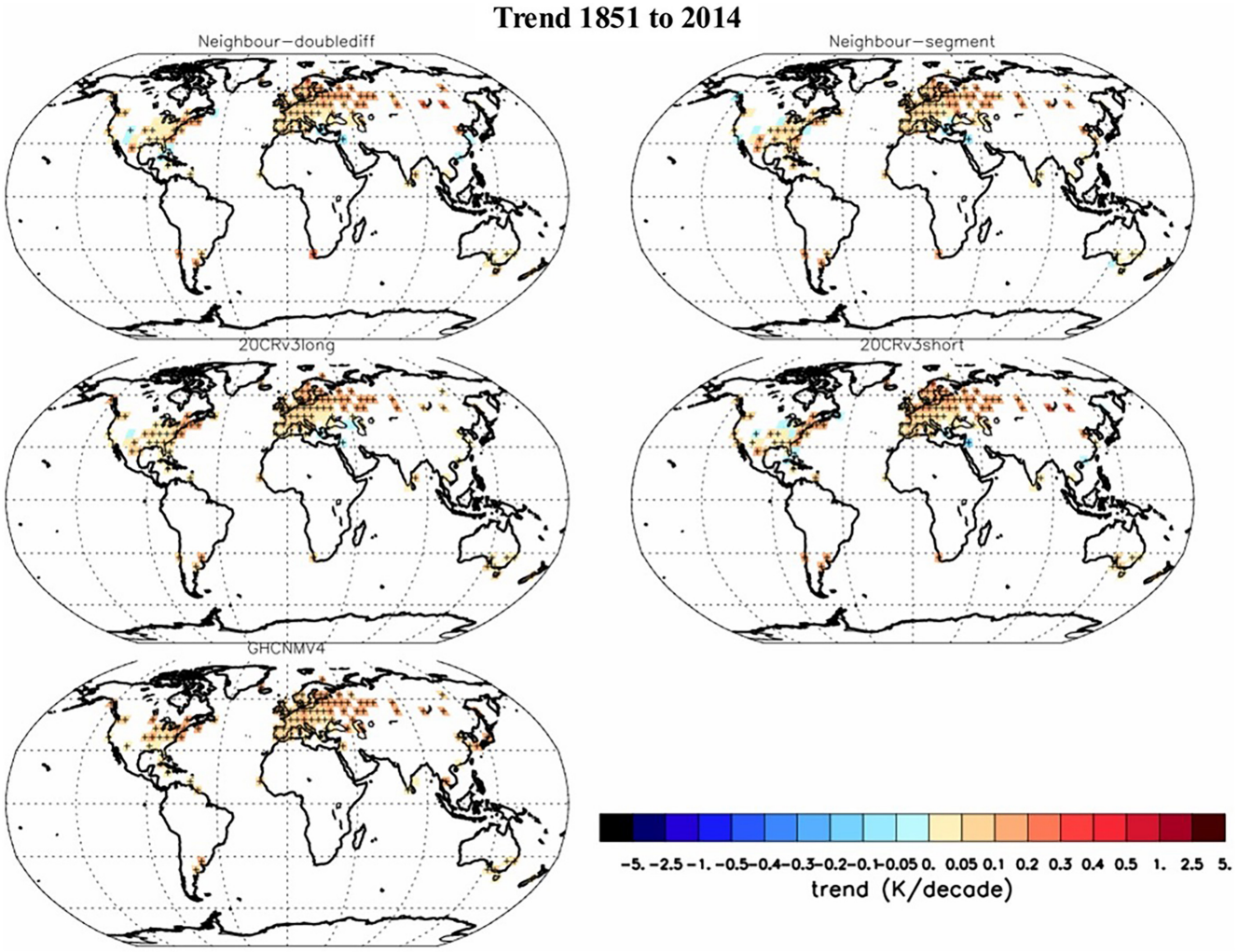

FIGURE 17 Gridbox trend analysis from 1851 to 2014. Trends have been calculated using OLS regression and based upon a requirement for 70% reporting with some reports in the first and final decile. Trend significance is denoted by + signs and ascertained from AR(1) corrected uncertainty estimation following Santer *et al.* (2008) [Colour figure can be viewed at wileyonlinelibrary.com]

estimates in the Balkans. The same also holds true for Australia and the region of Japan/Korea/Eastern China as well as southern South America. To the extent spatial coverage permits, this also broadly holds true over Africa.

The most obvious differences arise over and around the Indian subcontinent and in North America. Over the Indian subcontinent, a consistent feature across $20CRv3_{short}$ and the two neighbour-based approaches is a significant local cooling. The cooling is also present, but to a much lesser degree, in $20CRv3_{long}$. Conversely, GHCNMv4 estimates a robust warming in this region. Over North America, the patterns, significance and even sign of the trend differ between the five estimates in a zone from the southeast of the United States through the central to upper plains. GHCNMv4 warms everywhere, but that warming is not significant across the southeast—the well documented warming hole (Pan *et al.*, 2004; Kunkel *et al.*, 2005; Mascioli *et al.*, 2017). $20CRv3_{short}$ and $neighbour_{double\text{-}diff}$ agree with GHCNMv4 over the lack of significance of this regional warming, whereas $20CRv3_{long}$ and $neighbour_{segment}$ approaches show significant warming. In the central to upper plains both $20CRv3_{short}$ and $neighbour_{double\text{-}diff}$ show a slight cooling, in contrast to all remaining estimates.

Over the more recent periods of 1951–2014 and 1980–2014 (not shown for brevity; see Gillespie, 2021), when almost complete global land domain coverage is achieved, at least over the inhabited continents, trends are much more consistent between the four estimates produced herein and with GHCNMv4 than for the earlier periods.

In summary, globally, all four adjustment techniques show reasonable trend agreement between each other and with GHCNMv4 across a range of timescales. Where substantial differences arise, these are most pronounced in less well sampled regions and epochs. Differences also grow back in time as would be expected given the dataset construction techniques which progressively accumulate adjustments, and thus homogenisation uncertainty.

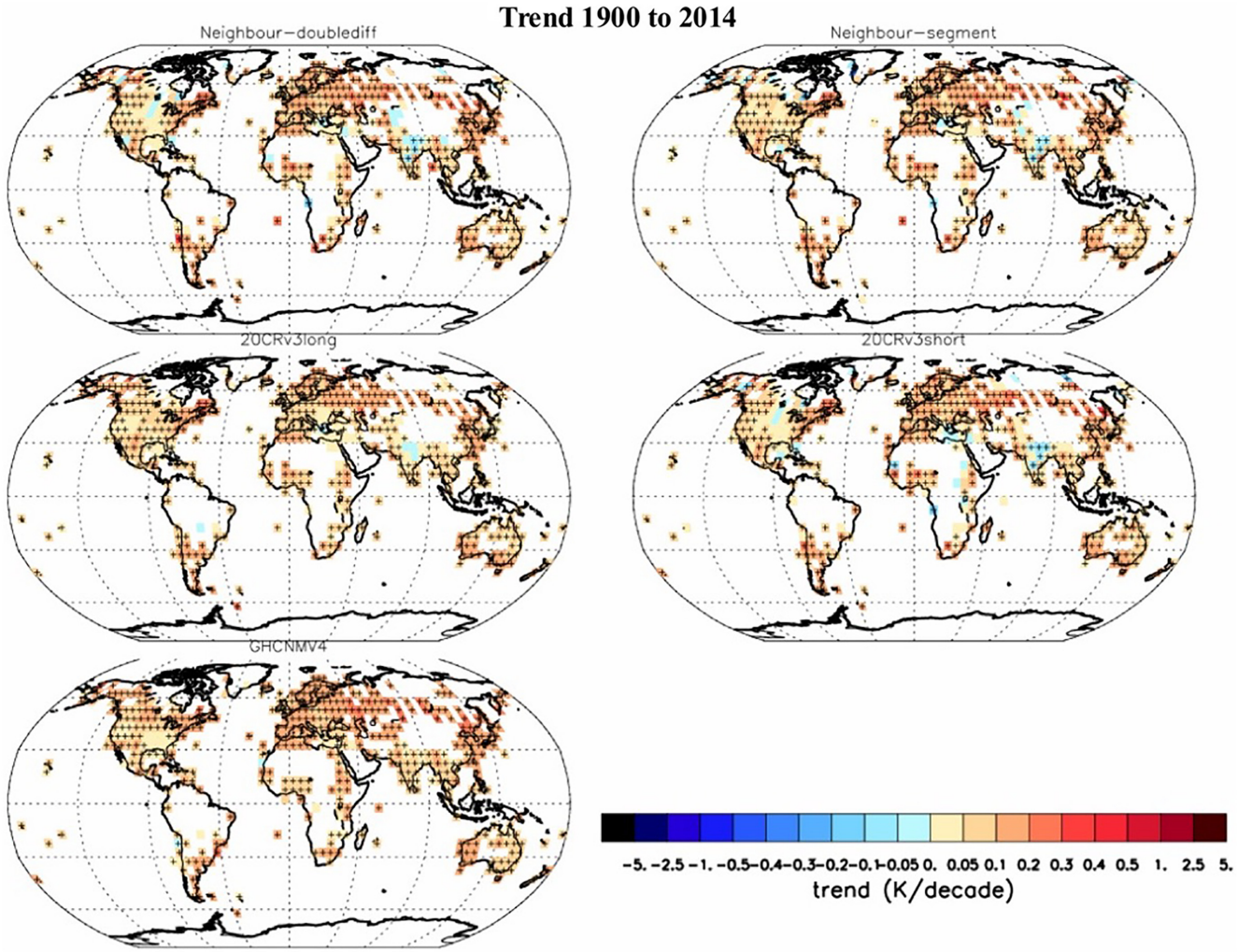

FIGURE 18 As Figure 17 but for the period 1900 to 2014 [Colour figure can be viewed at wileyonlinelibrary.com]

While there exist interesting differences, which are most apparent in trends over 1900–2014, and predominantly arise over the Indian subcontinent and North America, these differences are probably not sufficient to disqualify any of the approaches as constituting a reasonable homogenisation technique. If instead, one or more of the approaches had been systematically distinct from all other estimates either in broad-scale trend patterns or introducing spurious spatiotemporal structure ("spottiness"), then it would have provided grounds for rejection of that approach.

5 | REGIONAL AND HEMISPHERIC ANALYSIS

The regionally aggregated analyses consider the same series as section 4 but, in addition, includes 20CRv3 sampled to the station locations and data availability. This series may help in understanding any differences that arise between GHCNMv4 and the new set of estimates. Further plots and analyses, including difference series, are available in Gillespie (2021) Regional analyses are restricted to Europe, North America and Australia, where in all cases limited (in the case of Australia, extremely limited) data extend back to 1851. These regions are defined by simple bounding boxes and for each month the available grid boxes have been averaged using cosine of latitude weighting. Following these regional analyses, a hemispherically aggregated analysis is performed which can bring in additional information from more data sparse regions. In these comparisons, GHCNMv4 has been aggregated from the station series to ensure that any implied differences arise from a combination of any station selection mismatches (section 4.3) or the impacts of differences in station series adjustments, and not from additional postprocessing choices to create hemispheric and global averages undertaken by NOAA NCEI.

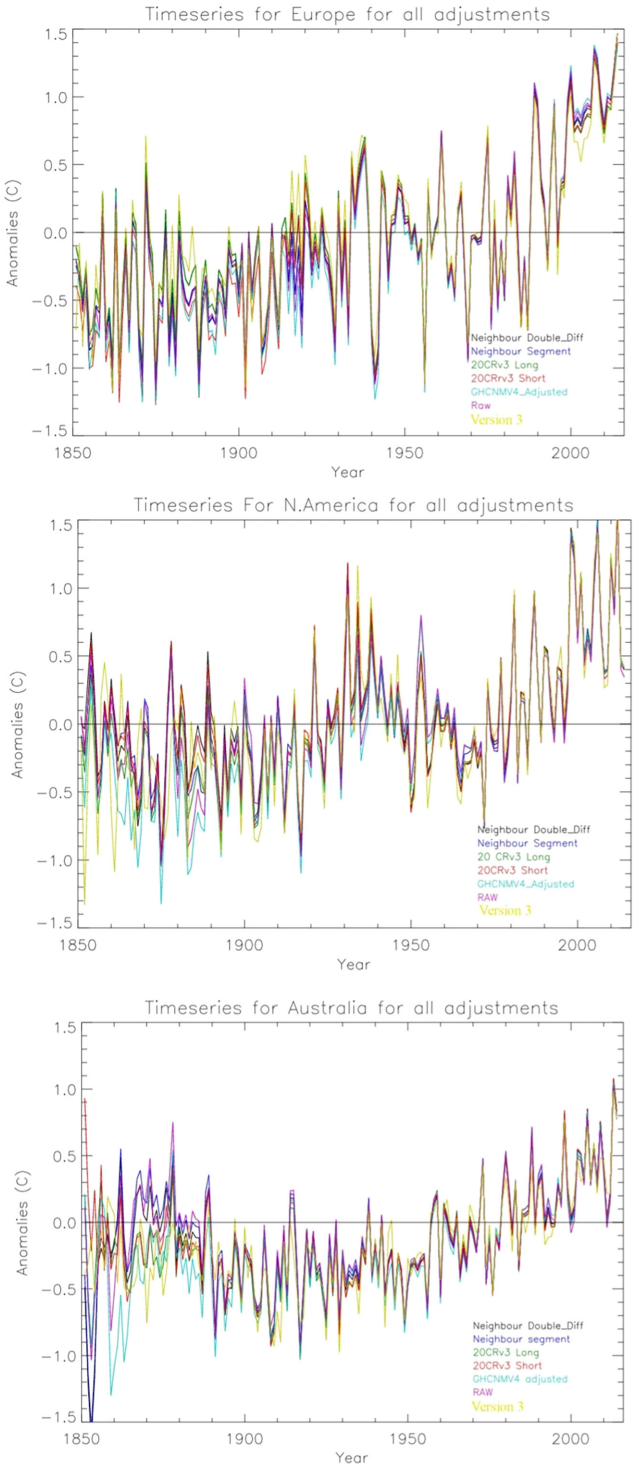

FIGURE 19 Annual anomalies relative to 1961–1990 for the European domain defined as 35°–70°N and 10°W–70°E (top panel), North America (middle panel) defined as 25°–60°N and 25°–60°W and Australia (lower panel) defined as 10°–45°S and 110°–155°E relative to a 1961–1990 climatology for the four products developed herein, GHCNMv4, the raw ISTI databank and 20CRv3 interpolated to station locations and spatially matched to observational availability [Colour figure can be viewed at wileyonlinelibrary.com]

5.1 | European domain

Over Europe, all series are overall in good agreement on interannual to decadal timescales. Occasionally the 20CRv3 reanalysis is a slight outlier (e.g., early 2000s, 1920s) and all series diverge to some extent prior to 1900 (Figure 19, top panel). The 20CRv3 reanalysis is systematically a little warmer in this period, although it continues to be strongly correlated. A concern that could therefore apply to $20CRv3_{long}$ and $20CRv3_{short}$ adjustment approaches used herein is that any such systematic 20CRv3 offsets might be incorporated via adjustment. Such concern over RAOBCORE led Haimberger *et al.* (2012) to produce the RICH approaches which assure a degree of independence in adjustment. In RAOBCORE/RICH there are visually obvious differences between the products following spatial aggregation (Haimberger *et al.*, 2012 and updates) which is not obvious here between the four adjusted products. Further, it is not obvious that the adjustment techniques that directly use 20CRv3 are being pulled unduly towards 20CRv3 behaviour when it diverges, particularly so for $20CRv3_{short}$. However, Europe is a region of plentiful surface pressure observations where 20CRv3 will a *priori* be well constrained throughout the record.

Considering long-term trends, all four solutions broadly agree with GHCNMv4 (Table 2). Over 1851 to 2014, $20CRv3_{short}$ is very similar to GHCNMv4, but the other three solutions indicate slightly less warming. This persists if, instead of an OLS trend, the change between 1850–1900 and 2005–2014 is considered. Over 1900 to 2014 $20CRv3_{long}$ suggests substantially less warming than the remaining three solutions and GHCNMv4. For both the 1950 to 2014 and the 1980 to 2014 periods all solutions including GHCNMv4 substantially agree. For most periods considered, the four new solutions collectively bracket the raw data trends, and to a lesser extent, GHCNMv4.

5.2 | North American domain

Over North America post-1970 there is strong agreement (partially forced by the choice of 1961–1990 climatologies) between all timeseries, including GHCNMv4 (Figure 19, middle panel). Between 1940 and 1970 all four approaches to homogenisation herein closely agree, but GHCNMv4 is slightly warmer around the 1950s. Over the 1930s and 1940s (the period of the dust bowl), GHCNMv4 is somewhat cooler than the remaining estimates. This study's four distinct approaches start to show sufficient systematic differences prior to ca. 1920 to be

TABLE 2 Trend analysis for four time periods for the European region defined as 35°–70°N and 10°–70°E

Dataset	OLS trends in °C per decade				Change 1851–1900 to 2005–2014
	1851–2014	1900–2014	1950–2014	1980–2014	
$Neighbour_{double\text{-}diff}$	0.077 ± 0.016	0.106 ± 0.031	0.213 ± 0.068	0.418 ± 0.153	1.55
$Neighbour_{segment}$	0.076 ± 0.016	0.106 ± 0.031	0.215 ± 0.068	0.418 ± 0.154	1.54
$20CRv3_{long}$	0.067 ± 0.016	0.089 ± 0.031	0.194 ± 0.067	0.400 ± 0.153	1.39
$20Crv3_{short}$	0.088 ± 0.016	0.106 ± 0.031	0.201 ± 0.067	0.413 ± 0.153	1.65
GHCNMv4	0.087 ± 0.017	0.123 ± 0.031	0.230 ± 0.067	0.413 ± .156	1.67
Raw data	0.081 ± 0.016	0.109 ± 0.032	0.217 ± 0.067	0.406 ± 0.156	1.57
20CRv3	0.064 ± 0.015	0.080 ± 0.031	0.180 ± 0.063	0.402 ± 0.143	1.34

Note: Linear trend estimates are calculated using ordinary least squares regression (OLS) following Santer *et al.* (2008) technique accounting for AR(1) effects on the d.o.f. Also shown is the simple change in means between 1850 and 1900 and 2005 and 2014 (final column).

TABLE 3 As Table 2 but for the North American region defined as 25°–60°N and 45°–135°W

Dataset	OLS trends in °C per decade				Change 1851–1900 to 2005–2014
	1851–2014	1900–2014	1950–2014	1980–2014	
$Neighbour_{Doublediff}$	0.046 ± 0.018	0.076 ± 0.030	0.194 ± 0.055	0.225 ± 0.134	1.00
$Neighbour_{Segment}$	0.050 ± 0.016	0.077 ± 0.028	0.187 ± 0.054	0.250 ± 0.131	1.06
$20CRv3_{long}$	0.060 ± 0.016	0.084 ± 0.028	0.195 ± 0.054	0.230 ± 0.134	1.16
$20CRv3_{short}$	0.048 ± 0.018	0.077 ± 0.030	0.196 ± 0.052	0.227 ± 0.132	1.02
GHCNMv4	0.073 ± 0.016	0.089 ± 0.025	0.168 ± 0.061	0.247 ± 0.136	1.33
Raw data	0.054 ± 0.016	0.072 ± 0.025	0.149 ± 0.061	0.227 ± 0.133	1.08
20CRv3	0.064 ± 0.017	0.087 ± 0.032	0.205 ± 0.053	0.226 ± 0.137	1.19

TABLE 4 As Table 2 but for the Australian region defined as 10°–45°S and 110°–155°E

Dataset	OLS trends in °C per decade				Change 1851–1900 to 2005–2014
	1851–2014	1900–2014	1950–2014	1980–2014	
$Neighbour_{doublediff}$	0.038 ± 0.017	0.085 ± 0.015	0.139 ± 0.030	0.153 ± 0.080	0.75
$Neighbour_{segment}$	0.032 ± 0.018	0.082 ± 0.015	0.142 ± 0.030	0.155 ± 0.080	0.68
$20CRv3_{long}$	0.038 ± 0.013	0.088 ± 0.014	0.126 ± 0.030	0.151 ± 0.080	0.75
$20CRv3_{short}$	0.033 ± 0.015	0.089 ± 0.015	0.139 ± 0.030	0.165 ± 0.080	0.72
GHCNMv4	0.051 ± 0.014	0.090 ± 0.015	0.145 ± 0.030	0.126 ± 0.079	0.91
Raw data	0.028 ± 0.016	0.076 ± 0.014	0.128 ± 0.030	0.113 ± 0.080	0.61
20CRv3	0.039 ± 0.012	0.089 ± 0.013	0.116 ± 0.026	0.176 ± 0.068	0.75

able to clearly distinguish between them. These differences become much more marked in the 19th century. At times in the early record, the 20CRv3 reanalysis shows marked inter-annual distinctions from all remaining observationally-based estimates. Again, there is no obvious visual evidence that this leads to any biases in the two adjustment methods that rely upon the 20CRv3 differences directly. GHCNMv4 is systematically cooler over North America than all other estimates prior to 1900 by several tenths of a degree C. The early period divergence leads to GHCNMv4 reporting greater warming between 1850–1900 and 2005–2014 by between 0.17 and 0.3°C than the four new adjustment techniques (Table 3, final column). Trends from 1900 onwards agree reasonably.

5.3 | Australian domain

Over Australia, all temperature series show good correspondence since the start of the 20th century, with offsets between series apparent prior to this (Figure 19, lower panel). The very earliest period data relies upon a single

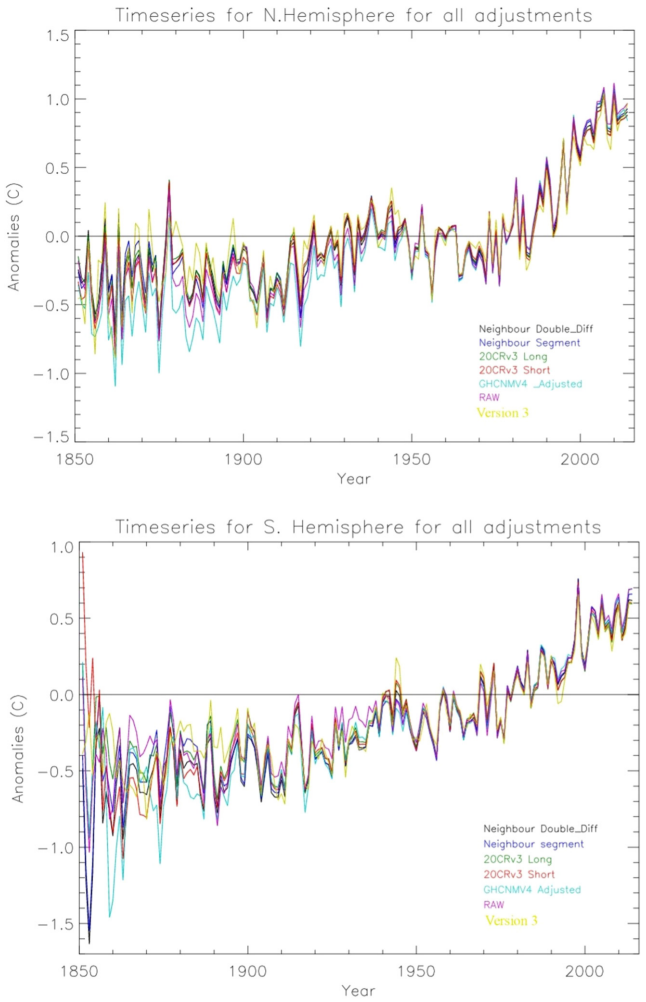

FIGURE 20 Northern Hemisphere annualized (top panel) and Southern Hemisphere (lower panel) time series shown for all four homogenized adjustments, 20CRv3 sparse input reanalysis, GHCMNv4 and the raw unadjusted time series [Colour figure can be viewed at wileyonlinelibrary.com]

record from Tasmania and thus should be treated with extreme caution as it is not truly representative of the broader Australian domain. All estimates show greater long-term warming than the raw series (Table 4). GHCNMv4 indicates somewhat greater warming between 1851–1900 and 2005–2014 by 0.16–0.23°C than the four new techniques. From 1900 onwards all solutions align reasonably closely in their trend estimates. However, GHCNMv4 is again an outlier in the 1980 to 2014 period suggesting slightly less warming per decade than the remaining solutions and being much closer to the raw data.

5.4 | Hemispheric analyses

In the well-sampled Northern Hemisphere (Figure 20, top panel), all series are indistinguishable from one another after the mid-20th century. Prior to 1950, GHCNMv4 becomes systematically cooler and the effect grows back into the late 19th century when it becomes of the order 0.2°C cooler. All other series are barely distinguishable from one another all the way back to 1850. Even in 1850, there exist numerous stations in the Northern Hemisphere. The same cannot be said for the Southern Hemisphere (Figure 20, bottom panel) where the earliest records in the ISTI databank currently arise from the single station in Tasmania (Gergis *et al.*, 2021), although data rescue efforts can, hopefully, improve this situation in the future (Brönnimann *et al.*, 2019).

In the Southern Hemisphere (Figure 20, bottom panel), different products can periodically be distinguished from one another throughout the series. Differences become particularly marked prior to the 20th century. The 20CRv3 reanalysis estimates interpolated to the locations and times of station observations is distinguishable from remaining products prior to 1950. Again, there is no obvious indication that this biases those estimates that directly or indirectly rely upon it for adjustments compared to those which do not. Differences between GHCNMv4 and all remaining estimates are considerably smaller and less systematic in the Southern Hemisphere than is the case in the Northern Hemisphere. Nevertheless, the tendency is, again, for GHCNMv4 to be slightly cooler prior to 1950 than the new estimates.

Considering trends, GHCNMv4 estimates more long-term warming than the other four adjustment techniques in the Northern Hemisphere (Table 5). The disparity is larger for periods prior to 1950 but persists to a degree even over 1950–2014. Then, in the most recent period, GHCNMv4 shows less warming than the four new estimates. In all periods, the four new estimates are more similar to each other than they are to GHCNMv4. For the longest periods, they warm less than the raw ISTI databank holdings whereas GHCNMv4 warms more. This may, in part, relate to the station sampling being distinct for GHCNMv4 (section 4.3). In the Southern Hemisphere (Table 5), there is more spread in estimates of the 1851–1900 to 2005–2014 changes across the four new estimates than is the case for the NH (compare the final columns in the two subtables). This is also reflected in the OLS-regression based trend estimates for the same period.

6 | INTERCOMPARISON OF GLOBAL SERIES TO OTHER PRODUCTS

In this section, we compare five peer-reviewed LSAT trend estimates, all of which were assessed in the recent

TABLE 5 As Table 2 but for the Northern and Southern Hemispheres

Dataset	OLS trends in °C per decade				Change 1851–1900 to 2005–2014
	1851–2014	1900–2014	1950–2014	1980–2014	
Northern Hemisphere					
$Neighbour_{doublediff}$	0.057 ± 0.014	0.090 ± 0.026	0.190 ± 0.042	0.310 ± 0.060	1.16
$Neighbour_{segment}$	0.058 ± 0.015	0.089 ± 0.028	0.195 ± 0.042	0.310 ± 0.061	1.19
$20CRv3_{long}$	0.056 ± 0.013	0.084 ± 0.025	0.181 ± 0.041	0.302 ± 0.060	1.13
$20Crv3_{short}$	0.062 ± 0.013	0.089 ± 0.025	0.181 ± 0.041	0.308 ± 0.060	1.20
GHCNMv4	0.080 ± 0.013	0.107 ± 0.024	0.200 ± 0.040	0.300 ± 0.590	1.44
Raw data	0.067 ± 0.015	0.096 ± 0.027	0.200 ± 0.044	0.313 ± 0.058	1.31
20CRv3	0.055 ± 0.0122	0.080 ± 0.0243	0.174 ± 0.037	0.310 ± 0.0568	1.10
Southern Hemisphere					
$Neighbour_{doublediff}$	0.072 ± 0.009	0.088 ± 0.012	0.134 ± 0.019	0.158 ± 0.041	1.10
$Neighbour_{segment}$	0.064 ± 0.010	0.089 ± 0.012	0.141 ± 0.018	0.168 ± 0.039	1.03
$20CRv3_{long}$	0.052 ± 0.010	0.079 ± 0.015	0.124 ± 0.019	0.160 ± 0.039	0.88
$20Crv3_{short}$	0.059 ± 0.013	0.081 ± 0.013	0.129 ± 0.019	0.162 ± 0.042	0.98
GHCNMv4	0.071 ± 0.011	0.088 ± 0.012	0.138 ± 0.019	0.168 ± 0.035	1.15
Raw data	0.056 ± 0.010	0.077 ± 0.014	0.138 ± 0.021	0.181 ± 0.036	0.99
20CRv3	0.051 ± 0.010	0.078 ± 0.013	0.120 ± 0.020	0.158 ± 0.043	0.85

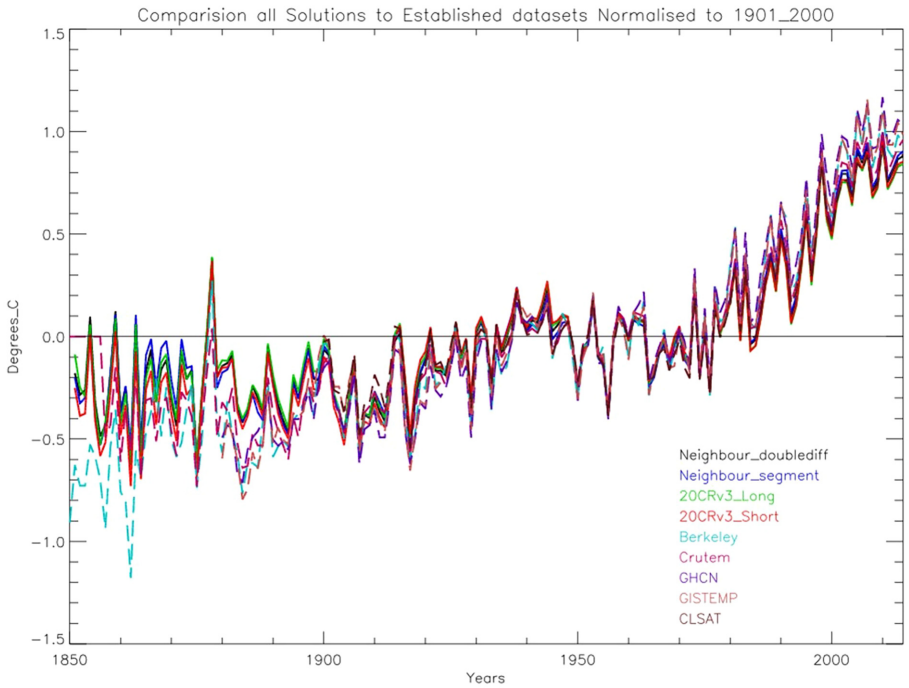

FIGURE 21 Comparison of the established datasets of Berkeley, CRUTEMv5, GHCNMv4, GISSTEMP, C-LSAT, with the four variants constructed herein: $neighbour_{double\text{-}diff}$, $neighbour_{segment}$, $20CRv3_{long}$ and $20CRv3_{short}$. All series have had a 1901–2000 climatology subtracted to try to highlight oftentimes small differences in behaviour. Pre-existing published estimates are given in dashed lines to further accentuate differences between available products and the new estimates constructed herein [Colour figure can be viewed at wileyonlinelibrary.com]

IPCC AR6 report (Gulev *et al.*, 2021). We compare CRUTEMv5 (Osborn *et al.*, 2020), GHCNMv4 (Menne *et al.*, 2018), GISTemp (Lenssen *et al.*, 2019), Berkeley (Rohde *et al.*, 2013) and CLSAT (Xu *et al.*, 2017) to the four new adjusted datasets developed herein (using the averaging as in section 5). For this purpose, global-mean series were sourced from public-facing repositories. Thus differences between the series may arise from some combination of station selections, homogeneity assessments and postprocessing including choices over interpolation and averaging methods applied. For this section, the GHCNMv4 series for consistency is thus swapped out

TABLE 6 Global trend analysis comparison with other land surface air temperature datasets

Dataset	OLS trends in °C per decade				Change 1851–1900 to 2005–2014
	1851–2014	1900–2014	1950–2014	1980–2014	
CRUTEM5	0.074 ± 0.012	0.096 ± 0.021	0.180 ± 0.034	0.259 ± 0.046	1.28
GHCNMv4		0.117 ± 0.020	0.206 ± 0.028	0.273 ± 0.048	
GISStemp		0.104 ± 0.023	0.201 ± 0.030	0.277 ± 0.052	
CLSAT		0.082 ± 0.022	0.170 ± 0.034	0.251 ± 0.042	
Berkeley	0.085 ± 0.012	0.105 ± 0.020	0.196 ± 0.028	0.259 ± 0.052	1.46
$Neighbour_{double\text{-}diff}$	0.056 ± 0.013	0.088 ± 0.021	0.170 ± 0.033	0.260 ± 0.050	1.08
$Neighbour_{segment}$	0.056 ± 0.014	0.088 ± 0.023	0.177 ± 0.032	0.264 ± 0.049	1.09
$20CRv3_{long}$	0.053 ± 0.012	0.081 ± 0.021	0.163 ± 0.032	0.256 ± 0.049	1.03
$20CRv3_{short}$	0.060 ± 0.012	0.085 ± 0.021	0.164 ± 0.031	0.260 ± 0.050	1.11
Raw	0.062 ± 0.013	0.090 ± 0.023	0.179 ± 0.036	0.270 ± 0.047	1.19
20 CRv3	0.052 ± 0.011	0.079 ± 0.020	0.157 ± 0.029	0.260 ± 0.048	0.99

TABLE 7 Global trend analysis sensitivity assessment using three versions

Dataset	OLS trends in °C per decade				Change 1851–1900 to 2005–2014
	1851–2014	1900–2014	1950–2014	1980–2014	
Default settings					
$Neighbour_{double\text{-}diff}$	0.056 ± 0.013	0.088 ± 0.021	0.170 ± 0.033	0.260 ± 0.050	1.07
$Neighbour_{segment}$	0.056 ± 0.014	0.088 ± 0.023	0.177 ± 0.032	0.264 ± 0.049	1.09
$20CRv3_{long}$	0.053 ± 0.012	0.081 ± 0.021	0.163 ± 0.032	0.256 ± 0.049	1.03
$20CRv3_{short}$	0.060 ± 0.012	0.085 ± 0.021	0.164 ± 0.031	0.260 ± 0.050	1.11
11,314 stations with at least 20 years of observations in 1961–1990 period					
$Neighbour_{double\text{-}diff}$	0.055 ± 0.013	0.091 ± 0.021	0.177 ± 0.032	0.263 ± 0.051	1.07
$Neighbour_{segment}$	0.054 ± 0.014	0.088 ± 0.022	0.177 ± 0.032	0.265 ± 0.049	1.07
$20CRv3_{long}$	0.055 ± 0.011	0.082 ± 0.021	0.167 ± 0.031	0.258 ± 0.050	1.06
$20CRv3_{short}$	0.057 ± 0.012	0.084 ± 0.021	0.168 ± 0.031	0.264 + 0.050	1.08
SNHT crit value of 12 not 16					
$Neighbour_{double\text{-}diff}$	0.050 ± 0.013	0.083 ± 0.022	0.167 ± 0.033	0.259 ± 0.049	1.00
$Neighbour_{segment}$	0.049 ± 0.014	0.083 ± 0.023	0.175 ± 0.032	0.261 ± 0.049	1.01
$20CRv3_{long}$	0.051 ± 0.012	0.079 ± 0.021	0.160 ± 0.031	0.255 ± 0.049	1.00
$20CRv3_{short}$	0.055 ± 0.012	0.082 ± 0.210.	0.159 ± 0.031	0.259 ± 0.049	1.05

Note: Top set is those used in section 5. The middle set is using the same settings but restricted solely to the subset of stations for which a 1961–1990 climatology can be directly calculated (about 40% of all stations). The bottom set keeps all settings the same except for using an SNHT critical value of 12 instead of 16.

from that used in sections 4 and 5 to the series made available directly from NOAA NCEI and used in their monitoring activities. Interpolation differences can have substantial impacts (Gulev *et al*., 2021, sect. 2.3) and the new estimates undertake no interpolation which is likely to underestimate long-term warming trends.

There are substantial similarities in behaviour across a range of timescales between the estimates (Figure 21), although note that CLSAT, GHCNMv4 and GISS, at least in the public versions sourced did not extend all the way back to 1851. There are, however, some differences. The four newly produced estimates are systematically lower than the other estimates post the mid-1990s—an affect that may plausibly arise from the lack of interpolation (Morice *et al*., 2021). They are also systematically a little warmer, less than 0.1°C, prior to 1900 and markedly so

over the 1880s and early 1890s by circa 0.2–0.25°C. To some extent, these divergences are forced by rebasing all global series to a 1901–2000 climatology which has been chosen to emphasize any dataset differences to the extent possible and practicable.

For the longest period considered (1851–2014), trends and deltas comparisons can only be made to CRUTEMv5 and Berkeley Earth, both of which suggest considerably more warming than either the ISTI databank raw data or the new 20CRv3-based estimates (Table 6). The effect is a reduction in long-term warming estimates of between 15 and 40% depending upon the choice of comparator product and which of the four adjustment approaches are considered, but this quantification is limited due to the availability of only two published series that extend their public version series back to 1850. With the notable exception of CLSAT all four remaining pre-existing estimates warm more than the raw ISTI databank holdings over 1901–2014 by about 10–25%, whereas the four new estimates all support CLSAT. These two classes represent uninterpolated (new estimates, CLSAT) and interpolated (CRUTEM5, Berkeley) sets of products so the differences may arise from more than just homogeneity adjustments. For trends starting 1950 or later the new set of estimates broadly are comparable to all existing estimates. The increasing divergence between all estimates further back in time results from some combination of the integrative effects of homogenisation uncertainty into the past and additional differences arising from station selection, quality control and postprocessing choices.

7 | DISCUSSION

The present analysis has undertaken a limited exploratory analysis of the application of sparse-input reanalyses to homogenize available LSAT station records. The methods, based upon published approaches to radiosonde homogenisation using full input reanalysis (Haimberger *et al*., 2012), result in four estimates which have the same breakpoint detection step but different adjustment approaches. The estimates pass basic quality checks concerning: the distribution of adjustments, station series inspection, monthly anomaly fields and gridded trend estimates. Comparisons at regional to global aggregations highlight a reduced estimate of long-term warming by 15–40% globally compared to estimates arising from published products depending upon the pair of products and the change metric being considered. This principally relates to estimates of changes prior to the early 20th century when data are sparse and instrumentation and methods of observation were not yet standardized. The new exploratory estimates do not fundamentally alter existing conclusions that the global LSAT has warmed or that this warming has accelerated over recent decades. However, if verified, they may have important implications for how close we now are to Paris Agreement temperature goal thresholds.

While the present analysis has shown the data products produced to be apparently reasonable, this analysis has not been exhaustive. Prior to operationalisation of these products or their use in a policy setting, a range of further analyses would be required with a particular focus upon understanding the divergence with antecedent products that becomes particularly marked prior to the early 20th century as highlighted in section 6. Station series from a greater range of these prior products should be compared to this study's new products. Additional regional analyses should also be undertaken with an emphasis on locations, such as the Indian subcontinent, where the gridded trends appear to diverge from one another and from GHCNMv4. The impact of interpolation choices would also need to be quantified.

Significant efforts would be required to quantify and understand the uncertainty in these new estimates. Parametric uncertainty estimation via the production of an ensemble of plausible solutions would be consistent with state-of-the-art approaches such as GHCNMv4 (Menne *et al*., 2018). Gillespie (2021) includes an initial assessment of what choices have been hardwired into the present system but could be pulled out and allowed to vary in such an ensemble along with an initial judgement as to whether these may prove important. These include uncertainties related to station selection, neighbour selection for the two neighbour-based adjustment approaches, breakpoint detection, adjustment and postprocessing, and serve to highlight a large number of choices which could have an impact upon the resulting series. Of course, until such an ensemble were run these assessments cannot be verified. The availability of one or more alternative state-of-the-art sparse-input reanalyses would also help to understand the impact of uncertainties in the production of sparse-input reanalyses.

In lieu of a full parametric ensemble, sensitivity of results to two plausible sources of high uncertainty have been investigated. The first is if the use of 20CRv3 to adjust the climatology for those stations which cannot derive a 1961–1990 climatology estimate directly. This is assessed by simply gridding solely those stations for which the 1961–1990 climatology can be calculated directly. The second is to use an SNHT critical value of 12 rather than 16 and rerun the entire end-to-end analysis which greatly increases the number of breakpoints returned. The use of 20CRv3 to infill station climatology estimates for stations with insufficient data has only minor impacts upon globally aggregated estimates (Table 7). The use of an SNHT threshold of 12 has a more

considerable impact, would further increase the difference to existing estimates, and suggest slightly less warming. While these two comparisons are very far from an exhaustive assessment of parametric uncertainty they do place a firm lower bound on what this could plausibly be. In particular, the use of the SNHT threshold alone changes many long-term change estimates by 10–15% suggesting that the uncertainty would be at least of this magnitude and possibly considerably larger. Such an uncertainty, assuming it were quasi-symmetrical, could easily reconcile our new long-term warming rate estimates with many of the existing records, especially when accounting for their uncertainty. This highlights the critical importance of quantifying uncertainty in the new products prior to their application in an operational context.

8 | CONCLUSION

Using the 20CRv3 sparse-input reanalysis product, an assessment of the ISTI databank has been undertaken to produce four novel exploratory homogenized estimates building methodologically on similar work by Haimberger *et al*. (2012) for radiosondes. The estimates have been assessed from the individual station series through various aggregations to global and verified on this basis as potentially reasonable estimates. When compared to the full-range of published estimates of LSAT change, the series broadly agree for changes since 1950 but increasingly diverge before then, particularly prior to the early 20th century. Thus while this analysis demonstrates that there is potential for sparse input reanalysis products to contribute to the homogenisation of LSAT series, further work is required to investigate why the divergence occurs between these estimates and other datasets. In addition, further work upon uncertainty quantification and a means to provide regular updates would be required prior to using such estimates in an operational context.

ORCID

Peter W. Thorne https://orcid.org/0000-0003-0485-9798

REFERENCES

Alexandersson, H. and Moberg, A. (1996) Homogenization of Swedish temperature data part 1. Homogeneity test for linear trends. *International Journal of Climatology*, 17, 25–34.

Allison, L., Hawkins, E. and Woollings, T. (2014) An event-based approach to understanding decadal fluctuations in the Atlantic meridional overturning circulation. *Climate Dynamics*, 44, 163–190.

Bronnimann, S. (2015) *Climatic Changes since 1700*. Cham: Springer.

Brönnimann, S., Allan, R., Ashcroft, L., Baer, S., Barriendos, M., Brázdil, R., Brugnara, Y., Brunet, M., Brunetti, M., Chimani, B., Cornes, R., Domínguez-Castro, F., Filipiak, J., Founda, D., Herrera, R., Gergis, J., Grab, S., Hannak, L., Huhtamaa, H., Jacobsen, K., Jones, P., Jourdain, S., Kiss, A., Lin, K.E., Lorrey, A., Lundstad, E., Luterbacher, J., Mauelshagen, F., Maugeri, M., Maughan, N., Moberg, A., Neukom, R., Nicholson, S., Noone, S., Nordli, Ø., Ólafsdóttir, K.B., Pearce, P. R., Pfister, L., Pribyl, K., Przybylak, R., Pudmenzky, C., Rasol, D., Reichenbach, D., Řezníčková, L., Rodrigo, F.S., Rohr, C., Skrynyk, O., Slonosky, V., Thorne, P., Valente, M.A., Vaquero, J.M., Westcottt, N.E., Williamson, F. and Wyszyński, P. (2019) Unlocking pre-1850 instrumental meteorological records: a global inventory. *Bulletin of the American Meteorological Society*, 100, ES389–ES413.

Caussinus, H. and Mestre, O. (2004) Detection and correction of artificial shifts in climate series. *Applies Statistic*, 53, 405–425.

Compo, G., Sardesmukh, P., Whitaker, J., Brohan, P., Jones, J. and Mccoll, C. (2013) Independent confirmation of global land warming without the use of station temperatures. *Geophysical Research Letters*, 40, 3170–3174.

Compo, G., Whitaker, J.S., Sardeshmukh, P.D., Matsui, N., Allan, R.J., Yin, X., Gleason, B.E., Vose, R.S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R.I., Grant, A.N., Groisman, P.Y., Jones, P.D., Kruk, M.C., Kruger, A.C., Marshall, G.J., Maugeri, M., Mok, H.Y., Nordli, Ø., Ross, T.F., Trigo, R.M., Wang, X.L., Woodruff, S.D. and Worley, S.J. (2011) The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, 137, 1–28.

Domonkos, P. and Coll, J. (2017) Homogenisation of temperature and precipitation time series with ACMANT3: method description and efficiency tests. *International Journal of Climatology*, 37, 1910–1921.

Gergis, J., Baillie, Z., Ingallina, S., Ashcroft, L. and Ellwood, T. (2021) A historical climate dataset for southwestern Australia, 1830–1875. *International Journal of Climatology*, 41, 4898–4919.

Gillespie, I. M. (2021) *Investigating the potential use of sparse-input reanalyses to homogenise long-term land surface air temperature records*. National University of Ireland, Maynooth. Available at: https://mural.maynoothuniversity.ie/14901/.

Gillespie, I.M., Haimberger, L., Compo, G. and Thorne, P.W. (2020) Assessing potential of sparse-input reanalyses for centennial-scale land surface air temperature homogenisation. *International Journal of Climatology*, 41, E3000–E3020.

Gulev, S.K., Thorne, P.W., Ahn, J., Dentener, F.J., Gerland, S., Gong, D., Kaufman, D.S., Nnamchi, H.C., Quaas, J., Rivera, J. A., Sathyendranath, S., Smith, S.L., Trewin, B., Von Shuckmann, K. and Vose, R.S. (2021) Changing state of the climate system. In: MassonDelmotte, V.P., Zhai, A., Pirani, S.L., Connors, C., Péan, S., Berger, N. and Caud, Y. (Eds.) *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Geneva: IPCC.

Guttman, N. (1998) *Homogeneity, data adjustments and climatic normals*. Asheville, NC: National Climatic Data Center.

Haimberger, L., Tavolato, C. and Sperka, S. (2012) Homogenization of the global radiosonde temperature dataset through combined

comparison with reanalysis background series and neighboring stations. *Journal of Climate*, 25, 8108–8131.

Hausfather, Z., Cowtan, K., Menne, M. and Williams, C. (2016) Evaluating the impact of U.S. historical climatology network homogenization using the U.S. climate reference network. *Geophysical Research Letters*, 43, 1695–1701.

Hunziker, S., Gubler, S., Calle, J., Moreno, I., Andrade, M., Velarde, F., Ticona, L., Carrasco, G., Castellón, Y., Oria, C., Croci-Maspoli, M., Konzelmann, T., Rohrer, M. and Brönnimann, S. (2017) Identifying, attributing, and overcoming common data quality issues of manned station observations. *International Journal of Climatology*, 37, 4131–4145.

Kanamitsu, M., Alpert, J.C., Campana, K.A., Deaven, D.G., Iredell, M., Katz, B., Pan, H.L., Sela, J. and White, G.H. (1991) Recent changes implemented into the global forecast system at NMC. *Weather and Forecasting*, 6, 425–435.

Karl, T.R. and Williams, C., Jr. (1987) An approach to adjusting climatological time series for discontinuous inhomogeneities. *Journal of Climate and Applied Meteorology*, 26, 1744–1763.

Knight, J.R., Folland, C.K. and Scaife, A.A. (2006) Climate impacts of the Atlantic multidecadal oscillation. *Geophysical Research Letters*, 33, 1–4.

Kunkel, K., Liang, X.-Z., Zhu, J. and Lin, Y. (2005) Can CGCMs simulate the twentieth-century "warming hole" in the central United States. *Journal of Climate*, 19, 4137–4153.

Lawrimore, J., Rennie, J. and Thorne, P.W. (2015) Responding to the need for better global temperature data. *Earth and Space Science News*, 94, 61–62.

Lenssen, N.J.L., Schmidt, G.A., Hansen, J.E., Menne, M., Persion, A., Ruedy, R. and Zyss, D. (2019) Improvements in the GISTEMP uncertainty model. *Journal of Geophysical Research: Atmospheres*, 124, 6307–6326.

Mascioli, N.R., Previdi, M., Fiore, A.M. and Ting, M. (2017) Timing and seasonality of the United States "warming hole". *Environmental Research Letters*, 12, 034008.

Menne, M. and Williams, C. (2009) Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, 22, 1700–1717.

Menne, M., Williams, C.N., Gleason, B.E., Rennie, J.J. and Lawrimore, J.H. (2018) The global historical climatology network monthly temperature dataset, version 4. *Journal of Climate*, 31, 9835–9854.

Morice, C.P., Kennedy, J.J., Rayner, N.A., Winn, J.P., Hogan, E., Killick, R.E., Dunn, R.J.H., Osborn, T.J., Jones, P.D. and Simpson, I.R. (2021) An updated assessment of near-surface temperature change from 1850: the HadCRUT5 data set. *Journal of Geophysical Research: Atmospheres*, 126, e2019JD032361.

Osborn, T.J., Jones, P.D., Lister, D.H., Simpson, I.R. and Harris, I. (2020) Land surface air temperature variations across the globe updated to 2019 the CRUTEM 5 data set. *Journal of Geophysical Research*, 126, 1–22.

Pan, Z., Arritt, R.W., Takle, E.S., Gutowski, W.J., Anderson, C.J. and Segal, M. (2004) Altered hydrologic feedback in a warming climate introduces a "warming hole". *Geophysical Research Letters*, 31, 1–4.

Peterson, T. and Easterling, D.R. (1994) Creation of homogeneous composite climatological reference series. *International Journal of Climatology*, 14, 671–679.

Przybylak, R., Wyszyński, P., Nordli, Ø. and Strzyżewski, T. (2016) Air temperature changes in Svalbard and the surrounding seas from 1865 to 1920. *International Journal of Climatology*, 36, 2899–2916.

Quayle, R., Easterling, D., Karl, T. and Hughes, P. (1991) Effects of recent thermometer changes in the cooperative station network. *Bulletin of the American Meteorological Society*, 72, 1718–1724.

Rennie, J.J., Lawrimore, J.H., Gleason, B.E., Thorne, P.W., Morice, C.P., Meanne, M.J., Williams, C.N., Gambie De Almeida, W., Christy, J.R., Flannery, M., Ishihara, M., Kamiguchi, K., Klein-Tank, A.M.G., Mhanda, A., Lister, D.H., Razuvaev, V., Renom, M., Rustticucci, M., Tandy, J., Worlry, S. J., Venma, V., Angel, W., W., Brunet, M., Dattore, B., Diamond, H., Lazzara, M.A., Blancq, F.L., Luterbacher, J., Machel, H., Revadekar, J., Vose, R.S. and Yine, X. (2014) The international surface temperature initiative global land surface databank: monthly temperature data release description and methods. *Geoscience Data Journal*, 1, 75–102.

Rohde, R., A. Muller, R., Jacobsen, R., Muller, E. and Wickham, C. (2013) A new estimate of the average earth surface land temperature spanning 1753 to 2011. *Geoinformatics & Geostatistics: An Overview*, 1, 1.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y.-T., Chuang, H.-Y., Juang, H.-M.H., Sela, J., Iredell, M., Treadon, R., Kleist, D., Van Delst, P., Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., Van Den Dool, H., Kumar, A., Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Schemm, J.-K., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C.-Z., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R.W., Rutledge, G. and Goldberg, M. (2010) The NCEP climate forecast system reanalysis. *Bulletin of the American Meteorological Society*, 91, 1015–1058.

Santer, B.D., Thorne, P.W., Haimberger, L., Taylor, K.E., Wigley, T. M.L., Lanzante, J.R., Solomon, S., Free, M., Gleckler, P.J., Jones, P.D., Karl, T.R., Klein, S.A., Mears, C., Nychka, D., Schmidt, G.A., Sherwood, S.C. and Wentz, F.J. (2008) Consistency of modelled and observed temperature trends in the tropical troposphere. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 28, 1703-1722. https://doi.org/10.1002/joc.1756

Slivinski, L., Compo, G., Whitaker, J.S., Sardeshmukh, P.D., Giese, B.S., Mccoll, C., Allan, R., Yin, X., Vose, R., Titchner, H., Kennedy, J., Spencer, L.J., Ashcroft, L., Brönnimann, S., Brunet, M., Camuffo, D., Cornes, R., Cram, T.A., Crouthamel, R., Domínguez-Castro, F., Freeman, J.E., Gergis, J., Hawkins, E., Jones, P.D., Jourdain, S., Kaplan, A., Kubota, H., Blancq, F.L., Lee, T.C., Lorrey, A., Luterbacher, J., Maugeri, M., Mock, C.J., Moore, G.W.K., Przybylak, R., Pudmenzky, C., Reason, C., Slonosky, V.C., Smith, C.A., Tinz, B., Trewin, B., Valente, M.A., Wang, X.L., Wilkinson, C., Wood, K. and Wyszyński, P. (2019) Towards a more reliable historical reanalysis: improvements for version 3 of the twentieth century reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, 145, 2876–2908.

Slivinski, L.C., Compo, G.P., Sardeshmukh, P.D., Whitaker, J.S., Mccoll, C., Allan, R.J., Brohan, P., Yin, X., Smith, C.A., Spencer, L.J., Vose, R.S., Rohrer, M., Conroy, R.P., Schuster, D.C., Kennedy, J.J., Ashcroft, L., Brönnimann, S.,

Brunet, M., Camuffo, D., Cornes, R., Cram, T.A., Domínguez-Castro, F., Freeman, J.E., Gergis, J., Hawkins, E., Jones, P.D., Kubota, H., Lee, T.C., Lorrey, A.M., Luterbacher, J., Mock, C.J., Przybylak, R.K., Pudmenzky, C., Slonosky, V.C., Tinz, B., Trewin, B., Wang, X.L., Wilkinson, C., Wood, K. and Wyszyński, P. (2021) An evaluation of the performance of the twentieth century reanalysis version 3. *Journal of Climate*, 34, 1417–1438.

Trewin, B. (2010) Exposure, instrumentation, and observing practice effects on land temperature measurements. *Wiley Interdisciplinary Reviews: Climate Change*, 1, 490–506.

Venema, V., Mestre, O., Aguilar, E., Guilarro, J., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Muller-Westermeier, G., Lakatos, M., Williams, C., Menne, M., Lindau, R., Rasol, D., Rustemier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotte, F., Fratianni, S., Cheval, S., Klancer, M., Brunetti, M., Gruber, C., Prohom-Duran, M., Likso, T., Esteban, P. and Brandsma, T. (2012) Benchmarking homogenization algorithms for monthly data. *Climate of the Past*, 8, 89–115.

Vincent, L.A. (1998) A technique for the identification of inhomogeneities in Canadian temperature series. *Journal of Climate*, 11, 1094–1104.

Williams, C.N., Menne, M. and Thorne, P.W. (2012) Benchmarking the performance of pairwise homogenization of surface temperatures in the United States. *Journal of Geophysical Research: Atmospheres*, 117, D05116.

Xu, W., Li, Q., Jones, P., Wang, X.L., Trewin, B., Yang, S., Zhu, C., Zhai, P., Wang, J., Vincent, L., Dai, A., Gao, Y. and Ding, Y. (2017) A new integrated and homogenized global monthly land surface air temperature dataset for the period since 1900. *Climate Dynamics*, 50, 2513–2536.

Yang, Y.-M., An, S.-I., Wang, B. and Park, J.H. (2020) A global-scale multidecadal variability driven by Atlantic multidecadal oscillation. *National Science Review*, 7, 1190–1197.

How to cite this article: Gillespie, I., Haimberger, L., Compo, G. P., & Thorne, P. W. (2023). Assessing homogeneity of land surface air temperature observations using sparse-input reanalyses. *International Journal of Climatology*, *43*(2), 736–760. https://doi.org/10.1002/joc.7822