



**Maynooth
University**

National University
of Ireland Maynooth

**Investigating the potential use of sparse-input reanalyses to homogenise
long-term land surface air temperature records**

Ian Gillespie

Thesis Submitted for the Degree of Doctor of Philosophy

Irish Climate Analysis and Research Units (ICARUS), Department of Geography,
Maynooth University, National University of Ireland

February 2021

Head of Department:

Professor Gerry Kearns

Research Supervisors:

Professor Peter Thorne

Professor Leopold Haimberger



Irish Climate Analysis and Research Units

Table of Contents

ABSTRACT.....	v
Acknowledgements.....	vii
List of Figures	ix
List of Tables	xix
Chapter 1 Introduction	1
1.1 Background and rationale.....	1
1.2 The International Surface Temperature Initiative databank	2
1.3 Reanalyses.....	3
1.4 Research aims and thesis structure	3
Chapter 2 Literature Review	5
2.1 A brief history of observations of land surface air temperature	5
2.1.1 Early meteorological measurements – the birth of modern meteorology.....	5
2.1.2 The instigation and proliferation of National Weather Services	9
2.1.3 Standardisation of methods of observation through the early 20 th Century	10
2.1.4 Automation, the growth of third party data providers, and increased heterogeneity in recent decades	15
2.2 Database management efforts and the International Surface Temperature Initiative databank	18
2.3 Land surface air temperature homogenisation approaches	21
2.3.1 Breakpoint Detection.....	21
2.3.2 Adjustment of series at identified breakpoints	26
2.4 Assessing the strengths and limitations of modern homogenisation approaches.....	27
2.5 Brief Summary of state-of-the-art land surface air temperature datasets	29
2.6 Uncertainty characterisation	34
2.6.1 Theory	34
2.6.2 Derivation of uncertainty estimates	36
2.7 Reanalysis products.....	41
2.7.1 Full-input reanalysis products.....	42
2.7.2 Sparse-input reanalysis products.....	44
2.8 Radiosonde adjustments using full-input reanalyses	48
2.9 Summary	50
Chapter 3 Preparation of the ISTI databank holdings.....	52
3.1 Introduction	52

3.2 Summary of ISTI version 1.0.0 merging process and updates in version 1.1.0	54
3.2.1 Updates to v1.1.0	58
3.3 Removal of short period records from V1.1.0	61
3.4 Selection of case study stations and their neighbours for initial assessment	63
3.5 Investigation of duplication of station data throughout the ISTI databank	70
3.5.1 Simple paired duplicates	73
3.5.2 Complex Cases	76
3.5.3 Resolution of cases of stations with identical coordinates	80
3.6 Discussion	83
3.7 Conclusion	86
Chapter 4 Assessment of the utility of 20 th Century Products as a reference series for homogenisation of land surface temperatures	87
4.1 Introduction	87
4.2. Possible approaches to constructing comparator series	90
4.3 Interpolation of sparse-input reanalysis gridded series to station locations	95
4.4 Analysis of relative performance of sparse-input reanalysis ensemble averages and ensemble members	99
4.5 Analysis of suitability for undertaking homogenisation	104
4.5.1 Case study stations analysis	105
4.5.2 Regionally aggregated analyses	110
4.5.3 Comparison between densely and sparsely sampled regions	116
4.6 Discussion	120
4.7 Conclusion	125
Chapter 5. Application of homogeneity assessments using sparse-input reanalyses fields and comparison to existing approaches	127
5.1 Introduction	127
5.2 Quality control and breakpoint detection	128
5.2.1 Removal of gross outliers	128
5.2.2 Breakpoint detection	130
5.2.3 Break Assignment	135
5.3 Application of adjustments	139
5.3.1 20CRv3 _{long}	140
5.3.2 20CRv3 _{short}	140
5.3.3 Neighbour _{segments}	141
5.3.4 Neighbour _{double-diffs}	142

5.3.5 Adjusting Climatology to 1961-1990.....	144
5.4 Assessing the efficacy of the approach.....	145
5.4.1 Assessment of adjustments	145
5.4.2 Evaluation of impacts on individual station series.....	148
5.4.3 Spatial anomalies	159
5.4.4 Spatial trends	163
5.4.5 Summary of assessment of the efficacy of approaches.....	169
5.5 Regional, hemispheric and global analysis	169
5.5.1 Regional analysis	170
5.6 Intercomparison to other products	183
5.7 Discussion.....	187
5.8 Conclusion.....	195
Chapter 6 Summary and discussion	196
6.1 Context.....	196
6.2 Key Findings	196
6.2.1 Assessment of the ISTI Databank.....	196
6.2.2 Suitability of sparse-input reanalyses for reference series construction	197
6.2.3 Homogenisation using sparse-input reanalysis products	198
6.3 Implications.....	199
6.4 Limitations and possible future work	200
REFERENCES.....	203

ABSTRACT

The correction of meteorological observational records (homogenisation) for non climate artefacts is an important task. Very few, long-term meteorological station series are entirely free of non-climatic influences. Climate data homogenization aims to identify and remove these non climate factors. Numerous methods of homogenisation have been developed over the decades. Current state of the art approaches generally proceed using pairwise difference series between observations from a network of reference stations and the station under assessment. Such methods work well in well sampled regions such as Europe and North America, but are less successful in poorly sampled regions and epochs. Reanalyses are produced by assimilating available observations into a forecast model, producing complete fields that are consistent with: the input data, the model physics, and any external boundary conditions prescribed. Full-input reanalyses which assimilate data from all available sources have previously been used to homogenise radiosonde data records. This work sets out to investigate if sparse-input reanalysis products that only assimilate surface pressure and use prescribed sea-ice, sea surface temperatures and changes in atmospheric composition, can act as a suitable reference series for the homogenisation of land surface air temperatures and to compare the results to established methods. It is found that sparse-input reanalysis products have successively improved in their quality with each new generation. The most recent product from NOAA-CIRES – 20CRv3 – has comparable overall statistical properties when interpolated to station locations and differenced to pairwise differences. In well sample regions neighbour-based comparisons remain favourable, but in sparser regions and epochs –20CRv3 may be preferable. The 20CRv3 product is therefore used to identify breakpoints and then 4 distinct approaches are used to adjust the series. Two of these directly use the 20CRv3 fields to estimate adjustments, while the remaining pair use apparently homogeneous neighbour station series wherever possible. The resulting set of estimates show reasonable overall behaviour looking at station series behaviour, spatial anomalies and spatial trends. The thesis highlights the potential for sparse-input reanalysis to provide a substantial methodological degree of freedom in the homogenisation of

global land surface air temperature estimates, but further work is required in developing an operational version.

Acknowledgements

In the first instance, I wish to thank Prof. Peter Thorne, my supervisor, who created the opportunity to undertake this research project and whom as Director of the Irish Climate Analysis and Research Unit (*ICARUS*) of the Department of Geography at Maynooth University created and encouraged a friendly, informal and encouraging ethos in the unit. I also wish to acknowledge and express my sincere thanks to Prof. Thorne for the endless support, expert advice, endless patience and encouragement that was freely given at all times during this incredible, challenging but enjoyable journey.

I also wish to thank my second supervisor Prof. Leopold Haimberger of the University of Vienna for the advised, input and encouragement. I especially wish to thank Prof. Haimberger who hosted my visit to Vienna and for his welcome and making my visit so enjoyable.

This research was funded by a scholarship from the President of Maynooth University

I wish to thank all my colleagues at ICARUS, both PhD research candidates and professional staff who always offered friendship, encouragement and assistance during my research.

I would like to mention Miss Corinne Voces, Manager/Administrative Officer, who was always at hand to help in any way she could, for her constant good humour and for her efforts to make life easier for all PhD candidates.

I wish to thank Dr Gil Compo at NOAA for his advice and for making 20CRv3 available to me before the official release.

I wish to offer a special thanks to my wife Rebecca who made room for me by taking on all domestic duties that I should have completed so that I was free to work on this thesis. To my daughter Jordan who always encouraged me who made the space available to me so that I could undertake this work and who supported me through difficult times. My family made it possible for me to complete a life of study for endless years and well before the undertaking of this research and allowed me to fulfil a life-long ambition.

Finally, I dedicate this work to my late brother Paul and his late wife Ruth for what they did for me that made this effort possible.

List of Figures	Page	
Figure 2.1	Top left Aristotelis's Meteorologica. Top right is a drawing of the thermoscope believed to have been invented around 1603. The bottom panel is a picture of the remaining Little Florentine Thermometers - the measurement quality of which is remarkable for the period.	6
Figure 2.2	Evangelista Torricelli, a student of Galileo, best known for the invention of the barometer.	7
Figure 2.3	William Derham who first started outdoor temperature measurements and whose records contributed significantly to Manley's Central England Temperature record reconstruction.	8
Figure 2.4	Extract from Robert Fitzroy's "The Weather Book" published in 1863. Fitzroy established barometers around the coast, convinced that falling air pressure was indicative of upcoming storms.	10
Figure 2.5	Graphic comparison of the three prime temperature scales developed between 1700 and 1750. Lord Kelvin proposed the Kelvin scale a century later. In the early days of temperature measurement, a wide variety of regional and local temperature scales were used. It was in the mid-1700s that the wish for standards started to coalesce around the three most famous scales. The Reaumur scale fell out of favour around the mid 1800s.	11
Figure 2.6	George Symons conducted trials on 9 different types of thermometer stands (screens) between 1868 and 1870 and found the Stevenson Type screen to be superior.	12
Figure 2.7	The different screens employed in the late nineteenth and early twenty century. An observer reads thermometers in a Glaisher stand, next to a thermometer house (middle). The Stevenson screen (far left) has become an almost universal standard.	13
Figure 2.8	American Cotton Region Shelter, variant of the Stevenson screen (left) and Stevenson screen (right). Both apparatus are similar but becoming less common as automated instrumentation using dome multiplate enclosures replace them.	14
Figure 2.9	A typical amateur weather station cheaply available that is used to upload to various weather sites such as weather underground.	16
Figure 2.10	De-Bilt experiment (a) 1989-1991, (b) 1992-1995, in which ten screens were compared. The screens operated in parallel with a reference screen for two years. Both images show traditional Stevenson type screens and new multiplate screens under extensive testing.	17
Figure 2.11	Different screens tested at De Bilt: (a) KNMI multiple, (b) Vaisla multiplate, (c) Young Gil Multiplate, (d) Young aspirate (type 1 and 2), (e) Socrima multiplate, (f) Stevenson Screen P.V.C version, (g) Standard wooden Stevenson screen (Van der Meulen and Brandsma, 2008).	18
Figure 2.12	Guy Callendar published the first estimation of global temperature changes in 1938 (Callendar, 1938).	19

List of Figures	Page	
Figure 2.13	An elementary example of homogenisation by pairwise reference. The chart tells that a break around 1940 can be seen in the B-A series and in the C-A series but not in the C-B series. The break probably occurred in A (Sourced from http://variable-variability.blogspot.com/2012/08.html).	24
Figure 2.14	The maximum (t_{\max}), minimum(t_{\min}) and mean(t_{mean}) of the USCRN, USHCN raw, and USHCN adjusted data left column. Right column USHCN raw series minus USCRN (blue) and USHCN adjusted minus USCRN in red (Hausfather et al., 2016).	29
Figure 2.15	Figure 2.15. Comparison of several LSAT datasets normalised to 1981-2010 (https://www.metoffice.gov.uk/hadobs/monitoring/temperature.html)	30
Figure 2.16	Flow of the ensemble generation for CRUTEM dataset ensemble taken from Morice et al (2012).	38
Figure 2.17	Time series of the number of pressure observations per year in version 2 of the International Surface Pressure Databank (ISPDv2) from 1870 to 2010. Note the logarithmic scale along the y-axis. Inset panel: time series during the same period showing the number of observations in the Northern Hemisphere (blue curve) and Southern Hemisphere (red curve), Caption & image taken from Cram et al. (2015)	47
Figure 3.1	The proposed ISTI dataset construction process and stages of development taken from Rennie et al. (2014).	53
Figure 3.2	Timeseries of percentage of 5° by 5° grid boxes that contain land which has at least one station present plotted against year. The black curve represents coverage in the precursor GHCNv3 product. The red curve is ISTI v1.0.0 and the blue curve is ISTI v1.1.0. Taken from Rennie et al. (2015).	60
Figure 3.3	Global distribution and period of record (colours) of stage 3 monthly stations. Note the concentration of long record stations (reds and blacks) in N.America and Western Europe. Long records overplot shorter period records. Taken from Rennie et al. (2015).	60
Figure 3.4	Map of all removed stations (red dots) from V1.1.0 based upon removing all records with fewer than 120 monthly average obs. Most removed stations are from North America.	61
Figure 3.5	The locations 495 stations with fewer than 120 months observations in the ISTI, release v1.1.0 that commenced observations between 1880 and 1900 are shown in red. The 25 stations that commenced observations before 1880 are shown in the blue squares.	62
Figure 3.6	The 29 case study stations(marked with red crosses) and their 25 neighbours, (blue asterisks), (top panel) and (lower panel) with a stipulation of 50% data record overlap being required. Spread in neighbour location with a 50% overlap requirement is particularly marked for case study stations outside Europe and North America. Some case stations share neighbours	66

List of Figures	Page	
Figure 3.7	The differences in anomalies between the candidate station and each of its nearest 25 neighbours. Each difference is offset from by 1°C vertical intervals for clarity. If the candidate and neighbour station series do not intersect there is no candidate– neighbour pairs difference shown.	67
Figure 3.8	Plot of Helsinki Kumpula difference series to its 25 nearest neighbours (coloured and each successively offset vertically by 1 degree for clarity). Note the strong variations in behaviour particularly marked prior to 1970.	69
Figure 3.9	Station at Tromdolsnges Norway which shows a single duplication event with annually repeating offsets indicative of a homogenised series.	69
Figure 3.10	An example of a series that was assessed as not having sufficient similarity to warrant further action. The candidate minus neighbour series is the black trace in the center. The lower traces show in colour-code which segments of each station arise from each source and the identifier used in that source (if given and distinct from the station identifier itself).	74
Figure 3.11	Example from Argentina. As the station at Villa Gesell is made up from a single source and the station at Balcarce airport is a combination of three sources, with source 59013210 overlapping with the string of zeros, source 59013210 was removed in resolving the issue.	75
Figure 3.12	An example of exact data match that must result from a poor decision in the ISTI databank merge algorithm as these arise from the non-GHCND source deck.	76
Figure 3.13	Example where a single station’s observations are duplicated with many neighbouring stations in a complex arrangement. In this case, Station FIE00142226, Helsinki Kumpula, shares data with twelve other stations. Only 4 cases are shown here.	77
Figure 3.14	This figure illustrates the interconnection and duplication between Helsinki Kumpula, Finland, station FIE0042226 and twelve of its neighbours, several of which also share commonalities with additional stations shown on the right. Helsinki is by far the most complicated case with multiple stations being ghosted into one another in full or in part and several stations containing segments arising from multiple other stations in ways that would require expert local knowledge to satisfactorily resolve.	79
Figure 3.15	The data sources that made up Powell River duplication. Each trace shows data availability. In this case station CA001046392 was removed in resolving.	81
Figure 3.16	As Figure 3.15 but for Metoryuk/Nunivak example where there is no overlap between the two series and a merge was performed.	82
Figure 3.17	Summary of the locations of removed files due to: the NCEI blacklisting (black squares); accounting for station series ghosting (red diamonds) and exact location matches (blue triangles). This map does not include files deleted because of fewer than 120 months of observations. For that see figure 3.4.	84

List of Figures

Page

- Figure 4.1 Summary of neighbour station data availability for De Bilt since 1850 (the series extends to the 1700s but for the present study the interest is in the period since the 1850s driven by the availability of sparse-input reanalysis products and globally representative observations). This series is a centennial station series with almost continuous availability (bottom black) since the 1850s although prior to 1897 data arises from Utrecht and then several additional sources:
http://projects.knmi.nl/klimatologie/daggegevens/antieke_wrn/index.html. Within the ISTI databank data, 1901 to date arises from the KNMI hosted E-OBS. Data prior to 1901 arises from GHCNMv2 collection which appears to arise directly from KNMI. The 25 nearest neighbours (other colours) are shorter with no suitable neighbour amongst them to use for homogenisation in the 1850 to 1900 period. There is one potential neighbour for the period of 1900 to 1945, after which there are several possible neighbours for pairwise homogenisation. Effectively pairwise homogenisation techniques are not possible for the period of 1850 to 1945 without expanding the neighbour search radius due to a lack of suitable neighbours. 92
- Figure 4.2 As Figure 4.1 for Vartan ,Sweden since 1850 (the series again extends back before 1850). A limited set of pairwise comparisons would be possible throughout the series but with a marked step-change in capability around 2/3 of the way through the series when a substantial number of neighbour series becomes available. 93
- Figure 4.3 As Figure 4.1 but for Albany, New York (the series again extends back earlier than the 25 nearest neighbours (other colours) which are much less complete and frequently drop in and drop out with no suitable neighbours amongst them to use for homogenisation for the entire period. 94
- Figure 4.4 As Figure 4.4 Maps showing ISPD V3.7 & V4.7 for 1930 and 1950 (Source <https://psl.noaa.gov/data/ISPD/>) 97
- Figure 4.5 An example analysis from Bombay, India (18.9°N, 72.8°E) of correlation (r) and standard deviation (sigma) of the different series of the 56 ensemble members of 20CRv2c to determine if the ensemble mean or individual ensemble members are most suitable for further comparison to pairwise homogenisation. This analysis for this station is over the full period of January 1851 to December 2014. There are a total of 1628 observations out of a possible total of 1968 observations. The correlation v standard deviation are plotted for the 25 nearest neighbours, the three reanalysis products and the 56 20CRv2c ensemble members. Values closer to [1,0] would constitute increasingly valuable comparators which are highly correlated with low variability of the different series 102
- Figure 4.6 As Figure 4.5 but for Perth, Australia (32° S, 115.9°E) This analysis for this station is over the full period of January 1917 to September 2013. There are a total of 1028 observation out of a possible total of 1161 observations 103

List of Figures	Page
Figure 4.7	As Figure 4.5 but for Santiago, Chile (33.5°S, 70.7°W). 103
Figure 4.8	Top panel: Anomaly difference series between the long-running De Bilt series in the Netherlands (although note caveats around splicing stations identified in Figure 4.1) for the sub-period of record since 1850 and the sparse-input reanalysis-based estimates. Middle panel: anomaly difference series using De Bilt's 25 nearest neighbours. Bottom panel: anomaly difference series using De Bilt's 25 nearest neighbours with a minimum 50% data overlap. Comparisons are now available for the entire post-1850 portion of the De-Bilt data record, but at a cost to correlation and the standard deviation of the difference series (Table 4.3) In the two lower panels. Each neighbour difference series is a different colour for illustrative purposes in the middle and lower panels. 109
Figure 4.9	Map of regions used for the analysis as defined by Giorgi and Francisco (2000),((Tian et al 2018). 110
Figure 4.10	The 27,639 long-term stations in the ISTI dataset, following removal of questionable series as detailed in chapter 3, split out into Giorgi region groupings. Note the extra grouping of 'Not in Giorgi' which captures the Antarctic, remote islands, and some Arctic sites not included in the original 21 Giorgi regions. 111
Figure 4.11	The median value (50th percentile) of the regionally-aggregated differences series between 20CRv2c ensemble mean and the station anomalies at each timestep aggregated over the Giorgi regions. Each series is vertically offset for clarity. There is a marked degradation in apparent performance over many regions in the mid 20 th Century. For region definitions see main text. 112
Figure 4.12	As figure 4.11 but for 20CRv3. The 20CRv3 product shows better performance than either ERA-20C or 20CRv2c across all regions with stable behaviour back to at least 1900 across all regions. The mid 20 th Century is much more stable than either of the other sparse-input reanalysis products. 113
Figure 4.13	As Figure 4.11 but for ERA-20C which starts in only 1900. This is a clear limitation on the use of ERA-20C compared to the two NOAA sparse-input reanalysis products. Although ERA-20C contains apparent decadal variations in the mid 20 th century, the degradation in this case, is much less marked than for 20CRv2c in most regions (c.f. Figure 4.10). 114
Figure 4.14	For each of the 22 Giorgi regions, the bars summarize the standard deviation of timeseries shown in Figures 4.10 through 4.12 for the three reanalysis products from 1851(Top Panel.) The 20CRv3 product exhibits the lowest standard deviation for almost all regions . The bottom panel is the same analysis for a reduced period from 1900 to concur with the start of ERA-20C 115

List of Figures

Page

- Figure 4.15 Top Panel is a pooled comparison of the correlations (r) (Blue x's) between each station and its 25 nearest neighbours across for both the 100 densely-sampled and sparsely-sampled stations (200 times 25 independent values). Overplotted are correlations between the 20CRv3 product and the candidate series (Red Diamonds). These are each displaced in the x-axis by the distance to the median neighbour such that for stations in densely sampled regions the reanalysis is closer to $X=0$ and for progressively sparser station locations the reanalysis estimate is further displaced from $X=0$. The bottom panel is the same comparison as in the top panel, but for the standard deviation of the difference series. Neighbour-based pairwise comparisons are likely better when the distance from a candidate station to its neighbours is less than 350km and, conversely, 20CRv3 reanalysis performs better when the distances are c. 700km or greater. 118
- Figure 4.16 Correlations between the candidate station anomalies and the median neighbour compared to the correlation between the candidate station anomalies and that of 20CRv3 in well sampled (top panel) and sparsely sampled (bottom panel) regions. The median neighbour value is denoted as a black star and the 20CRv3 value as a red box. Higher values denote better agreement and thus greater suitability as a reference series to remove common climatic events to perform relative homogenisation. 119
- Figure 4.17 As Figure 4.16 but now considering the standard deviation of the difference series. Lower values would, all else being equal, lead to smaller breaks being able to be detected and reliably adjusted in the candidate series. 120
- Figure 4.18 Histogram of the occurrence of the frequency of overlap between each station and its 25 nearest neighbours aggregated over the poorly sampled candidate station-neighbour pairs (top panel) and well sampled regions (bottom panel) from 1850 to 2012. The most frequent occurrence is for no overlap for both regions and the median value is 6 out of 25 comparisons being possible at any given timestep in well sampled regions, falling to 3 in poorly sampled regions. 123
- Figure 4.19 Stations with 25 or more stations within 350 km radius (yellow) for which pairwise approaches may be preferable. Stations with the 25 nearest neighbours within 700km (blue) in which pairwise and 20CRv3 based approaches may be of comparable power according to the present analysis. Stations in more data sparse regions (red) which likely will be more amenable to homogenisation using 20CRv3. As successive sparse-input reanalysis products improve over time progressively more points may become blue or red in similar future maps. 125
- Figure 5.1 Map detailing locations of stations Quality Control using the same intervals as Table 5.1. Stations with more data removed overplot those with fewer data removed. 129

List of Figures	Page	
Figure 5.2	Examination of the SNHT scores for critical values from 6 to 100 producing a highly skewed smooth distribution that provides no clear rationale for the selection of a critical SNHT score value. The spikes at almost 100 relate to breakpoints assigned to account for timeseries cessation and resumption over a period of >36 months duration which are given a value of 99.9. to force a breakpoint to be assigned.	132
Figure 5.3	Histograms of breakpoint sizes inferred from the station minus 20CRv3 difference series at each breakpoint identified for different SNHT critical values (panels). Within each panel, the mean calculated adjustment and the standard deviation of the distribution are shown.	133
Figure 5.4	Analysis of the cumulative segment adjustments for each critical value from 6 to 20 in intervals of 2. Shown in-line within each panel is the total number of breaks identified, the mean cumulative adjustment and the standard deviation.	134
Figure 5.5	The analysis of the application of the SNHT test to the 100,000 simulations of homogenous time series with similar statistical properties to the difference series for the case study stations used in Chapter 4. The maximum SNHT score has been retained from each series and is plotted against the autocorrelation of the synthetic series (top panel) and the sigma of the synthetic series (lower panel).	135
Figure 5.6	Station at Baisun from Uzbekistan (38.2° N, 67.2° E, 1241 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.	136
Figure 5.7	As Figure 5.6 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.	137
Figure 5.8	As Figure 5.6 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 1080 to 1140) over the second world war. Breakpoints were detected in April 1978 and at May 1985 in addition to the assignment of a breakpoint upon time series resumption after WW2	138
Figure 5.9	As Figure 5.6 but for De Bilt (52.1014°N, 5.1867°E 2 m.a.s.l). The De Bilt site is continuous before and after 1851 to 2014, the period under examination for this work.	139

List of Figures	Page	
Figure 5.10	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.	142
Figure 5.11	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.	143
Figure 5.12	As Figure 5.6 but for GM000001474 Bremen, Germany (53.0464°N, 8.7992°E 4 m.a.s.l) with the highest number of returned breaks at 19.	146
Figure 5.13	Distribution comparison of adjustments for the four adjustment approaches employed in this analysis using an SNHT critical value of 16.	148
Figure 5.14	Figure 5.14 The resulting set of 20CRv3 minus station difference series for station Baisun Uzbekistan, UZM00038827 (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 5.5). The single breakpoint location identified in the present analysis is denoted by the solid red vertical line. Individual monthly values are shown.	149
Figure 5.15	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.	150
Figure 5.16	As of Figure 5.15, but only showing one adjustment (20CRv3 _{long} , the ISTI raw series and 20CRv3 interpolated anomalies	151
Figure 5.17	As Figure 5.15 but for Stykkiosholmur western Iceland. Differences in the final homogeneous segment relate to QC differences for raw and GHCNMv4 which alter some annual values.	152
Figure 5.18	As Figure 5.15 but for Midway Island.	153
Figure 5.19	As Figure 5.14 but for station NLM00006260 De Bilt, Netherlands,	154
Figure 5.20	As figure 5.15 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.	154
Figure 5.21	As Figure 5.15 but for Kisumu, Nyanza, Kenya (0.1 N°, 34.75 E° 1146 m.a.s.l).	151
Figure 5.22	As Figure 5.15 but for LG000026422 at Riga ,Latvia at (56.9625°N and 24.04°E, 17 m.a.s.l). Note the temperature drop circa 1940. This is not assessed to constitute a break.	156

List of Figures	Page	
Figure 5.23	As Figure 5.15 but for ITE00001729 at Parma, Italy at (44.8°N, 10.54°E, 54 m.a.s.l).	157
Figure 5.24	As Figure 5.15 but for USW00014837 at Madison Dane County Airport, Wisconsin, USA 43.(14°N 85.32°W at 264 m.a.s.l).	158
Figure 5.25	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.	160
Figure 5.26	Maps of June 1900 gridbox anomalies from a 1961-1990 climatology for (from top left to bottom right): double differencing; neighbour segment; 20CRv3 long; 20CRv3 short; NOAA NCEI's GHCNMv4 product and the original raw ISTI databank holdings. Plots produced using Panoply version 4.10.12 for windows.	161
Figure 5.27	As Figure 5.26 but for June 1970.	162
Figure 5.28	As Figure 5.26 but for June 2000.	163
Figure 5.29	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). Maps from top left to bottom right are for: Neighbour _{doublediff} , Neighbour _{segment} , 20CRv3 _{long} , 20CRv3 _{short} , GHCNMv4 and the raw ISTI databank holdings.	165
Figure 5.30	As Figure 5.29 but for the period 1900 to 2014.	166
Figure 5.31	As Figure 5.29 but for the period 1951 to 2014.	167
Figure 5.32	Intercomparison of trend analysis between the different methods of homogenisation and GHCNMv4 on a global basis for the period 1980 to 2014.	168
Figure 5.33	Top Panel Annual anomalies relative to 1961-1990 for the European domain defined as 35° N to 70° N and 10° W to 70° 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. Bottom panel is difference between GHCNMv4 and other data series	172
Figure 5.34	As Figure 5.33 but for the North American region defined as 25°N to 60°N and 45°W to 135°W.	175
Figure 5.35	As Figure 5.33 but for the Australian region defined as 10°S to 45°S and 110°E to 155°E. Note that there are very few observing sites early in this series and great caution should be exercised in interpretation.	177
Figure 5.36	Northern Hemisphere Annualised time series shown for all 4 homogenised adjustments, 20CRv3 sparse input reanalysis, GHCNMv4 and the raw unadjusted time series.	179
Figure 5.37	As Figure 5.36 but for the Southern Hemisphere.	180
Figure 5.38	Annualised global analysis of all homogenised and raw time series.	183

List of Figures

Page

- Figure 5.39 Top panel in a comparison of the established datasets of Berkeley, CRUTEMv5, GHCNMv4, GISTEMP, C-LSAT, with the 4 variants constructed herein: Neighbour_{double-diff}, Neighbour_{segment}, 20CRv3_{long}, and 20CRv3_{short}. All series have been normalised to 1901-2000 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. Bottom is the same as the panel but normalised to 1961-1990
- Figure 5.40 Trend analysis comparison for the period 1900 to 2014 using OLS trend estimation with AR(1) correction following Santer et al. (2008) for 20CRv3_{long} for the default version (top left), the same using only stations with 1961-90 stations (top right) and with an SNHT crit value of 12 (bottom left).

185

194

List of Tables

Table 2.1	A summary of the key characteristics of the five modern global gridded datasets highlighting similarities and distinctions between them.	31
Table 3.1	Changes in thresholds between ISTI release v1.0.0 and release v1.1.0 (Rennie, 2015).	59
Table 3.2	Summary of the stations with fewer than 120 monthly averages that were removed from further analysis by period, the total removed monthly averages, the minimum station length removed and the mean station length removed (the maximum in all cases is the 119 limit applied). The ISTI databank does not contain any stations with fewer than 24 monthly values	62
Table 3.3	A summary of the initial subset of 29 case study stations used herein including their identifier, name, country, geolocation, regional characteristics and local environment. For this analysis observations between January 1851 and December 2014 are used. Station records may start and end outside the selected dates	63
Table 3.4	Comparison of standard deviation and correlation (based upon the selection of neighbours based on two criteria (1) by at least 50% time series overlap and then (2) by 25 nearest by distance. The former will expand the search region and include, on average, stations further away. Differences in sigma, correlation and distance between the two selection methods are shown in the final three columns.	65
Table 3.5	Summary of initial case study series found to suffer from exact series replication between the case study station and one or more of the 25 nearest ISTI databank neighbours.	68
Table 3.6	Summary of duplication detection and type of duplication of data detected in the ISTI dataset version 1.1.0 along with a summary of the resolution.	72
Table 3.7	Summary of those files removed in full or in part to resolve Helsinki Kumpula identified overlaps.	78
Table 3.8	The Powell River example of duplication in Canada.	80
Table 3.9	The Metoryuk/Nunivak example, deemed to be the same site with two different names.	82
Table 3.10	Station location exact match resolution summary.	82
Table 3.11	A summary of the processing steps undertaken in the pre-processing of the ISTI databank performed herein.	84
Table 4.1	Comparison of interpolated reanalysis minus station difference series using inverse linear distance and inverse linear squared distance for correlation and Sigma using the 20CRv2c ensemble mean product for interpolation to selected stations (Chapter 3). The difference between the methods on both an individual basis and an aggregate basis for both sigma and correlation are small with a slight overall improvement when using the inverse squared distance approach. Given that this product is the coarsest resolution reanalysis, differences are smaller for other reanalysis products considered (not shown).	

Table 4.2	Summary of comparison of correlations and standard deviation of ensemble members to the ensemble mean and a summary of the comparison of the standard deviation of ensemble member's differences to the ensemble mean differences. The summary shows the maximum and minimum values obtained for the correlation across the entire ensemble versus the station anomalies in comparison to the correlation value of the ensemble mean to the station anomalies and the same for the standard deviation of the difference series of station anomalies minus 20CRv2c ensemble-mean interpolated anomalies. St = Station values r = Correlation coefficient Sigma – Standard Deviation	101
Table 4.3	A summary of the correlations (r) and the standard deviations (sigma, °C) of the anomaly difference series between the station anomalies and the reference which is either a reanalysis data set or the median of the 25 nearest neighbours for <i>high-density stations (italic)</i> : Intermediate stations (regular font); and stations located in sparsely sample areas (bold) .	108
Table 5.1	Summary of the frequency of different percentage intervals of observations removed by Quality Control from individual stations.	129
Table 5.2	Summary of the preponderance of breakpoint detection at an SNHT critical value of 16 across the raw ISTI databank stations retained following the analysis undertaken in chapter 3. This count includes cases where a breakpoint has been forced to account for a gap of 36 months or longer duration.	146
Table 5.3	Table 5.3 trend analysis for four time periods for the European region defined as 35°N to 70°N and 10°W to 70°E. 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 1850-1900 and 2005-2014 (final column).	173
Table 5.4	As Table 5.3 but for the North American region defined as 25°N to 60°N and 45°W to 135°W.	176
Table 5.5	As Table 5.3 but for the Australian region defined as 10°S to 45°S and 110°E to 155°E.	178
Table 5.6	As Table 5.3 but for the Northern Hemisphere.	181
Table 5.7	As Table 5.3 but for the Southern Hemisphere.	182
Table 5.8	Global trend analysis comparison with other Land Surface Air Temperature datasets.	186
Table 5.9	Possible sources of uncertainty which should be considered in the construction of a parametric uncertainty ensemble and an expert based estimation of their possible impact.	190
Table 5.10	Global trend analysis sensitivity assessment using three versions. Top set is those used in Section 5.5. The middle set is using the same settings but restricted solely to the subset of stations for which a 1961-90 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.	194

Chapter 1 Introduction

1.1 Background and rationale

Climate change is a topic of huge scientific and societal importance as evidenced by the Paris Agreement and national and European level policy interventions.

Underpinning many aspects of climate science is the observational evidence basis. Of all this evidence basis foremost in the public mind are records of global surface temperature change. Yet, when you dig down into the details there are surprisingly few available estimates and many of these estimates are not entirely independent from one another. There is significant scope to consider in novel ways this totemic record. Such investigation may yield new insights, and should ultimately improve scientific confidence in our estimates of changes to date. To address this, the focus of this thesis is to assess the utility of modern-day sparse input reanalysis products as a suitable reference series for the adjustment of long-term land surface air temperature data sets.

When an observation is taken at a weather station, that observation is of that moment. It can never be repeated. While the observed value may indeed remain constant for a period, it is never the same observation. Repeatability in science is desirable and even paramount. But unlike laboratory samples under controlled conditions that can be retested, climate observations cannot. Changes in aspects such as observational practices, siting, equipment upgrades etc have been ubiquitous across the global network. Such changes in very many cases introduce substantial spurious systematic and random effects into the observational series which must be identified and adjusted for.

Imperfect as they are, these measurements constitute the only available means to gauge the extent of temperature changes over the global land surface since the 19th Century. Comparing each station to the records of several local stations can highlight non-climatic artefacts. Each individual station making up a comparator group may contain non-climatic artefacts arising at different points in time. Thus a sufficient sample of pairwise comparisons should provide a basis to uniquely identify breaks in a given network of sites. This is the basic principle of state-of-the-art approaches to

the homogenisation of land surface air temperature records. However, long-term station series are only available for limited locations where meteorological measurements were performed since the 19th century (Bronnimann et al., 2013), with modern coverage only since the mid-to-latter part of the 20th Century. It is, therefore, a challenge to demonstrate the extent of climate change that has occurred since the industrial revolution by direct inference from long-term records (Hawkins et al., 2017).

The Intergovernmental Panel on Climate Change (IPCC) Working group 1 in their 5th Assessment Report state that “*It is certain that global mean surface temperature has increased since the late 19th century. Each of the past three decades has been successively warmer at the earth’s surface than any previous decades in the instrumental record*” (Hartmann et al., 2013). The datasets which lay behind this assessment all used some combination of station-wise and pairwise homogenisation techniques. This relative paucity of diversity in approaches, coupled to overlap in stations used, reduces independence in the available estimates. It is of huge potential value to increase the diversity of approaches. The current thesis aims to do so taking advantage of two advances since the IPCC fifth assessment report: i) improved holdings of fundamental data; and ii) advances in sparse-input reanalyses products.

The fundamental premise is that sparse-input reanalysis products can be used instead of neighbour based approaches to identify and potentially to adjust for breaks in the data holdings. This builds upon pioneering work using full-input reanalysis products to homogenise radiosonde data (Haimberger et al., 2012) which has gained broad-scale traction in a range of applications.

1.2 The International Surface Temperature Initiative databank

In 2010 scientists recognised the need for a comprehensive collection of ‘raw’ land surface air temperature data not unlike the database that exists for surface ocean measurements – the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) (Woodruff et al., 2011). There are millions of land surface observations arising from tens or even hundreds of thousands of stations both past and present sitting in repositories around the world. Many of these series extend back for decades

and even centuries. The International Surface Temperature Initiative (ISTI) databank results from significant efforts to collate and reconcile these sources and contains more than 35,000 stations from around the world (Rennie et al., 2014). However, while this extended global data collection is invaluable for research, the impacts of non-climatic artefacts still need to be removed prior to climate applications. To date, only one method – the pairwise approach from NOAA NCEI (Menne and Williams, 2009)– has been applied to attempt to homogenise these holdings through the creation of GHCNMv4.

1.3 Reanalyses

Reanalyses combine Numerical Weather Prediction models (NWP) with sophisticated data assimilation methods and selected available observations to reconstruct the state of the atmosphere upon a common grid (Kalnay et al., 1995, Bosilovich et al., 2012, Compo et al., 2011). Full input reanalyses use a combination of surface, upper-air and satellite measurements. They generally extend back only as far as at best the mid-twentieth Century. More recently, sparse-input reanalysis products have emerged that use only surface observations and extend back to at least 1900, and for some products the early 19th Century (Poli et al., 2016, Slivinski et al., 2019, Laloyaux et al., 2018, Compo et al., 2011). In this thesis, several state-of-the-art sparse-input reanalysis products (20CRv2C, 20CRv3, ERA-20C and CERA-20C) are considered as candidate series to be used to perform homogenisation on centennial-scale land surface air temperature records made available through the ISTI databank

1.4 Research aims and thesis structure

The principal aim of this thesis is to investigate if modern sparse input reanalysis (20th-century) products can act as a suitable reference series for the homogenisation of land surface air temperatures.

The remainder of the thesis is structured as follows:

- Chapter 2 provides a literature review that details the broader context of the present thesis highlighting the scientific context within which the specific work carried out rests.
- Chapter 3 introduces the ISTI databank, outlines its construction and undertakes an initial analysis of the data. A number of issues pertaining to the databank are highlighted leading to the removal of many short-term stations and about 5% of multi-decadal station records either in full or in part. Reasons for removal are justified.
- Chapter 4 assesses the suitability of sparse input reanalysis products (20th-century) as a reference series for assessment of long-term surface temperature homogeneity. It addresses the question of which current sparse input reanalysis products, if any, are most appropriate to act as a reference series and performs a comparative assessment of their performance compared to the pairwise comparison method.
- Chapter 5 goes on to perform a homogenisation of the surface temperature records using the 20CRv3 sparse-input reanalysis product. The methods applied borrow heavily from Haimberger et al. (2012) and estimates using both the reanalysis and apparently homogeneous neighbour segments for adjustments are considered. The resulting estimates are compared to the range of existing available products.
- Chapter 6 closes with a discussion reflecting critically upon the lessons learnt and a consideration of possible next steps.

Chapter 2 Literature Review

2.1 A brief history of observations of land surface air temperature

2.1.1 Early meteorological measurements – the birth of modern meteorology

Humans have long been interested in weather and climate. Throughout history, significant events, including early migration, the colonisation of new lands, the rise of agriculture, and the rise and fall of many civilisations were influenced by climate (Nicholson and Flohn, 1980). Aristotle drawing on the work of several Babylonian, Egyptian and other Greek scholars wrote *Meteorologica* around 340 BC, the first publication dedicated to meteorology (Figure 2.1, top left). In the first three books of the series of publications, Aristotle postulated as to the origin of the wind, the formation of rain, storms and other weather events (Modise and Mphale, 2018, Zen-de-Figueiredo-Neves et al., 2017). The Greeks were the first to publish meteorological records in the form of almanacs prominently centred around wind, reflecting the importance of navigation at sea (Bowker, 2011).

Interest in the weather was not confined to Europe. Other civilizations, remote from Europe, such as China, India and the Middle East engaged in speculation as to the origin of climate and how to measure it (Lee-Di, 1978). Archimedes and Hero of Alexandria both quoted the work of Philo of Byzantium, describing a primitive apparatus which many scholars now identify as a very basic thermoscope (Hellmann, 1908).

But records of direct measurement of meteorological phenomena with instrumentation only began in the past few hundred years. The earliest known pioneers of meteorological observations were William Merle of England and Marcin Biem of the Krakow Academy who recorded observations meticulously in the thirteenth and fourteenth centuries (Zen-de-Figueiredo-Neves et al., 2017).

Despite restrictions put in place on science by the catholic church during the dark ages, interest in measurements including those of temperature stretches far back. The search for a device of some form to make measurements include the early thermoscope (Figure 2.1 top right panel), the invention of which has been attributed

to Galileo, Santorio and others around 1603. Sometime during this early period, a scale was added to the thermoscope effectively inventing the thermometer, and although measurements were made, they were seldom recorded.

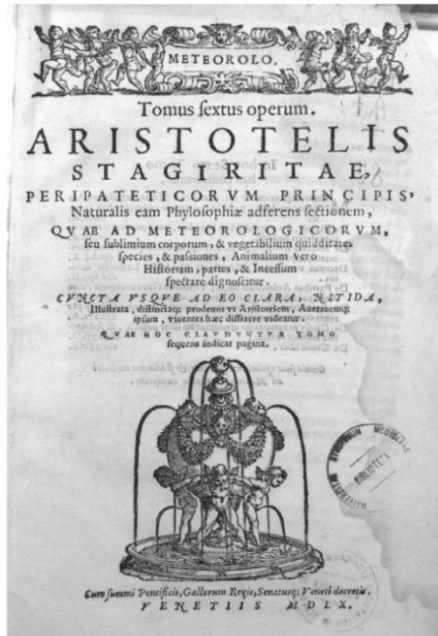


Figure 2.1 Top left Aristotelis's Meteorologica. Top right is a drawing of the thermoscope believed to have been invented around 1603. The bottom panel is a picture of the remaining Little Florentine Thermometers - the measurement quality of which is remarkable for the period.

The first known recorded long-term time series of daily temperature is the Medici Network which was run by the Grand Duke of Florence and his brother Prince

Leopold who set up a network consisting of eleven stations in 1654 including Florence, Milan, Pisa, Paris and Warsaw, all employing a strict protocol and identical “The Little Florentine Thermometer” instruments (LFT, Figure 2.1 bottom panel) (Camuffo and Bertolin, 2011). Measurements were taken under a strict protocol specifying the time and method of measurement and thus effectively introducing the first standardisation and enabling comparability of measurements across the network (Vittori and Mestitz, 1981). Indeed, these measurements were of sufficient quality and described with metadata so well that we can calculate the equivalent Celsius values from the units of the day. Tests made by Vittori and Mestiz in 1981 on the 15 surviving LFT implied a standard deviation among the LFT readings of just 0.5°C (Camuffo and Bertolin, 2011).

It was the invention of the barometer by Evangelista Torricelli during the 1640s (Figure 2.2) that kick-started the practice of regular atmospheric observations among gentlemen citizen scientists. Early adopters included Newton, Locke, Boyle, and Derham among others, who also came together to form the Royal Society in 1660. (Hoppen, 1976).



Figure 2.2 Evangelista Torricelli, a student of Galileo, best known for the invention of the barometer
In January 1699 William Derham (Figure 2.3) first started making outdoor observations. Earlier observations by Locke and others were taken in unheated

indoor rooms (Parker et al., 1992). The Derham observations series extended from 1699 through to 1706 with some missing or lost data for the year 1707. It is believed that Derham continued his observations through to 1730, although these records are yet to be rescued (Cornes et al., 2012). Derham took care to position his thermometer out of the direct sun on a shaded north wall and he meticulously recorded metadata which included his times of observations. However, questions arise as to the quality of Derham's observations due to his crude means of measurement and, despite his efforts, the possible impact of solar radiation on his thermometer. Derham's work was rapidly followed by other observers. From the efforts of these pioneers, Manley was later able to produce the 1659 to 1973 Central England Temperature monthly mean time series (Manley, 1974), which constitutes the longest available instrumental temperature record in the world (Parker et al., 1992).



Figure 2.3 William Derham who first started outdoor temperature measurements and whose records contributed significantly to Manley's Central England Temperature record reconstruction.

By the mid 1700s, due in no small part to the ability of artisan instrument manufacturers to produce instruments of near-identical properties on a large scale, interest in recording meteorological observations began to flourish (Zen-de-Figueiredo-Neves et al., 2017). By the late eighteenth century, France had begun to establish a large network of Meteorological stations across Europe and even in North America and Greenland in the belief that health and air quality were interlinked (Demaree et al., 2002). The first recorded observation in Belgium commenced in 1763 by Abbot Jean Bapiste Chevalier (Demaree et al., 2002). Spanish records

commenced in 1776, while Russian records date back to 1743 (Camuffo and Jones, 2002). These early networks can help us place 21st Century data in a historical context in terms of pre-industrial climate and its variability (Hawkins et al., 2017). Brönnimann et al. (2019) have recently brought a renewed interest in these early records via the collation of an inventory of all known pre-1850 instrumental records. Many of these records remain in hardcopy or image form requiring data rescue (Brunet and Jones, 2011, Allan et al., 2011).

2.1.2 The instigation and proliferation of National Weather Services

By the mid-1800s Robert Fitzroy, once captain of the HMS Beagle, developed the fundamental techniques of weather forecasting, with the main focus being to save lives at sea. Fitzroy became the first Director General of the UK Meteorological Office. The Meteorological Office obtained data from the growing network of weather stations set up in the UK, Europe and the US for analysis from which Fitzroy issued forecasts (Modise and Mphale, 2018, Murphy, 1998) and published his observations in “The weather book: a manual of practical meteorology” in 1863 (Figure 2.4). But the criticism that Fitzroy suffered from both the public and the gentlemen of the Royal Society when some of his forecasts proved inaccurate ultimately contributed to the taking of his own life. Fitzroy devised a code of meteorological telegraphy in cypher to transmit observations from various sites to a central station (Murphy, 1998). His work encouraged others to start the mapping of surface pressure and other meteorological observations that was the genesis of weather maps (Modise and Mphale, 2018).

By 1850 national weather services had been set up in many countries across Europe including Prussia, Austria, France, and the UK (Brönnimann et al., 2019). In 1853 Matthew Fontaine Maury, a lieutenant in the US navy convened the Brussels Conference on Meteorological data collection leading to the birth of the “public weather services” and the international sharing of data (Zillman, 2005). Such was the concern and need for shipping as the industrial revolution accelerated that the Brussels conference focused on maritime weather and most attendees were navy personnel. Maury’s first proposal to the conference was that observations from ships should be made available to all shipping. As a direct result of the Brussels initiative, and despite several false starts, the First International Meteorological Congress met

in September 1873 in Vienna. The 1873 Congress is considered a milestone in international cooperation in meteorology (WMO, 1973).

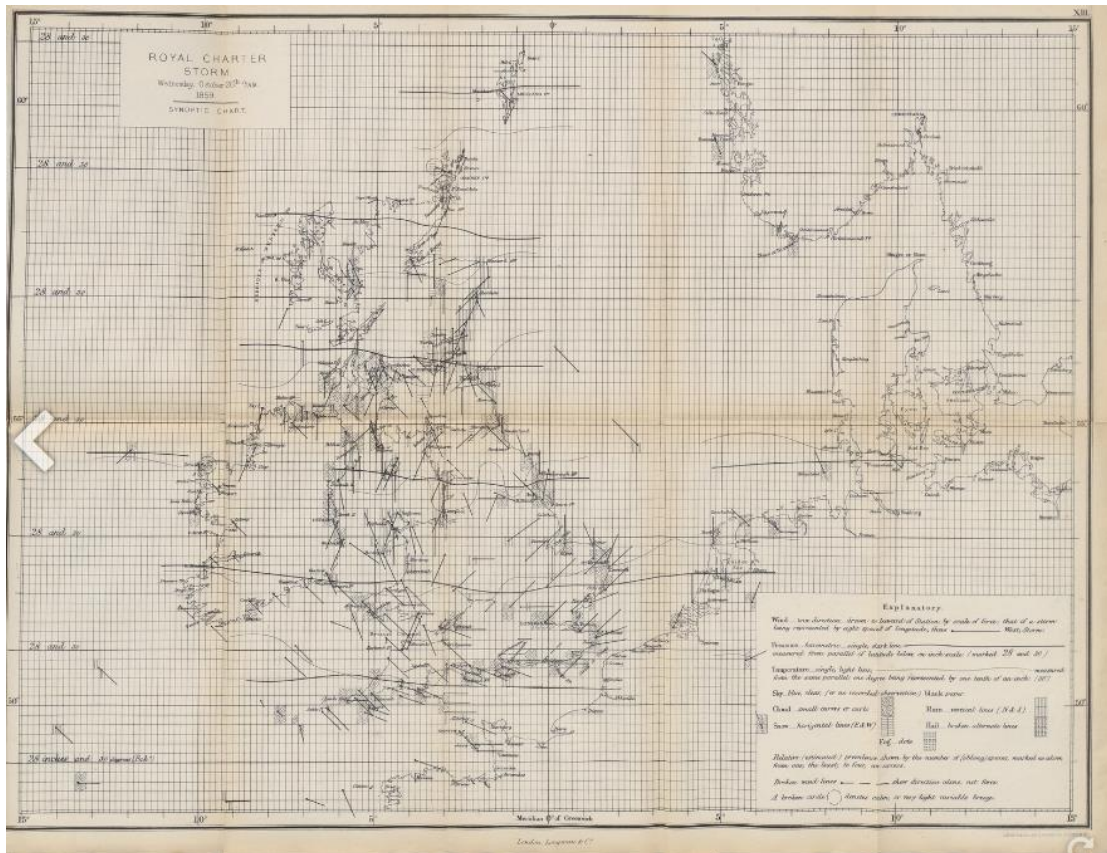


Figure 2.4 Extract from Robert Fitzroy’s “The Weather Book” published in 1863. Fitzroy established barometers around the coast, convinced that falling air pressure was indicative of upcoming storms.

2.1.3 Standardisation of methods of observation through the early 20th Century

Early observers of meteorology often had to devise their own protocols and measurement scales in the absence of recognised standards. By the mid 1700s, there were numerous different scales in use. This served to greatly inhibit the comparability of measurements. In 1774 Louis Cotte published a table that compared 15 of the most popular temperatures scales in use at the time (Camuffo and Jones, 2002). As the use of thermometers to measure air temperature became more widespread and common, attention turned to the need for standardised scales. The Reaumur scale was devised in 1713, followed by Fahrenheit in 1724 and then the Celsius scale which emerged in 1742 (Figure 2.5). All of these temperature scales were based on interpretations of the properties of water. Before that, a broad range of

alternative scales was in use, many of which were deployed nationally and even regionally. It is thanks to the metadata left by the curators of the individual time series, including the scale employed, that we can attempt homogenisation of these old time series (Eden, 2009).

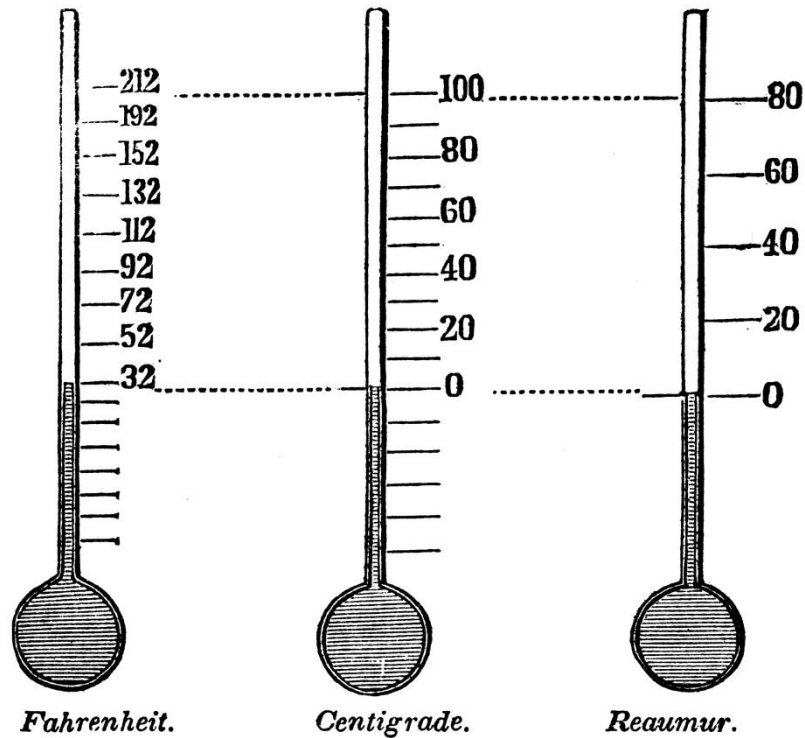


Figure 2.5 Graphic comparison of the three prime temperature scales developed between 1700 and 1750. Lord Kelvin proposed the Kelvin scale a century later. In the early days of temperature measurement, a wide variety of regional and local temperature scales were used. It was in the mid-1700s that the wish for standards started to coalesce around the three most famous scales. The Reaumur scale fell out of favour around the mid 1800s.

The variety of temperature scales used, however, was not the only issue that inhibited comparability of the earliest available temperature measurements. Other issues included the siting of the thermometer, times of observations, and the protection from solar radiation. Some observers positioned the thermometer on a north wall (or in the Southern Hemisphere a south wall) away from direct sunlight for at least the majority of the year. Others chose to place the thermometer in unheated rooms open to the elements (Parker et al., 1992). Often the height of the thermometer could vary from one meter to several meters above the ground, resulting in differences between individual observer reports (Manley, 1974, Parker, 1994). Not only would different observers choose different times of day to make

observations, but daily and monthly averages were also calculated using a broad range of methods (Manley, 1974).

Early attempts at standardisation of methods of observation include the work of the Societas Meteorologica Palatina in 1870 who dispensed calibrated equipment which included thermometers and barometers to various institutions with instructions that observations should be taken at 07, 14 and 21 hours. They also provided logbooks free of charge in return for copies of the observations (Zen-de-Figueiredo-Neves et al., 2017). In 1860 George Symons (Figure 2.6) by then an employee of the Royal Meteorological Society under the supervision of Robert Fitzroy initiated a trial of various meteorological instruments with the aid of volunteers. Symons was particularly interested in rain gauges. But he was also unhappy with the variety of thermometer stands and screens in use (Naylor, 2018). He enlisted the aid of Rev. Griffith between November 1868 and April 1870 and a trial was carried out with nine different types of thermometer screens which included an early precursor to modern-day Stevenson screens (Naylor, 2018). A report was eventually published in the *Quarterly Weather Report of the Meteorological Office* favouring the Stevenson and the Kew screen techniques and resulted in the Stevenson screen becoming widely adopted in the British and Irish Isles by the mid-1870s. By 1882 the Royal Meteorological Society required the Stevenson screen to be used in all their stations. In 1883 Mawley enlarged the Stevenson screen to provide better air circulation (Naylor, 2018), and his fundamental design is still widely used today.



Figure 2.6 George Symons conducted trials on 9 different types of thermometers stands (screens) between 1868 and 1870 and found the Stevenson Type screen to be superior.

At a meeting of the Royal Society in November 1873, Symons, Griffith and Stow presented a twelve point criteria that should be observed when siting a thermometer in a screen in a quest for “Uniformity”. These twelve points were accepted albeit with some amendments and formed the basis of standardisation to be adhered to throughout the British Empire (Naylor, 2018).

Standardisation in meteorology globally was first brought to completion at the second meeting of the Permanent Committee of the International Meteorology Congress held in Utrecht in September 1884. At this meeting, there were thirteen items on the agenda, five of which dealt with standardisation of instruments, observations and meteorological units (WMO, 1973). Despite the early recognition of the need for standardisation, different types of screens were still in use well into the twentieth century (Figure 2.7). It took until 1954 for the publication of the first version of “Guide to meteorological instruments and observing practice”. The World Meteorological Organisation, the successor to the IMO, continues this work today, issuing updates to technical guidance regularly. However, despite these efforts, climate records still require homogenisation to remove artefacts of a non-climatic nature. This in part is because, despite international guidelines being set out to obtain uniformity in measurements, the guidelines do not extend to specifying, for example, standard thermometers and screens (Brandsma and van der Meulen, 2008), the logic being that such specification may inhibit advances in measurement technology and lead to vendor lock-in (WMO, 1973).



Figure 2.7 The different screens employed in the late nineteenth and early twenty century. An observer reads thermometers in a Glaisher stand, next to a thermometer house (middle). The Stevenson screen (far left) has become an almost universal standard.

Even though the WMO did not specify the Stevenson screen as a standard, the latter part of the 19th century and early part of the 20th century saw a broad adoption of screened outdoor measurements using some form of Stevenson screen. The exact design varied from place to place. In the United States of America (Figure 2.8), a cotton region shelter became the standard, while in some British tropical colonies, a large thatched enclosure was common (Trewin, 2010) in the belief that the Stevenson screen was unsuitable for the tropics and that different shelters were more suitable for different latitudes (Parker, 1994).



Figure 2.8 American Cotton Region Shelter, variant of the Stevenson screen (left) and Stevenson screen (right). Both apparatus are similar but becoming less common as automated instrumentation using dome multiplate enclosures replace them.

Stevenson screen type measurements of various designs had almost become universal by 1972 (Sparks, 1972, Warne, 1999). The instruments were generally housed at 1.5 to 2m above the local ground surface, although with climatological snow depth in North America and parts of Europe being a determining factor (Parker, 1994), and readings were taken from maximum and minimum temperature thermometers in either Fahrenheit or Celsius. Contemporary to this, other similar standardisation for other meteorological parameters was introduced. The degree of adoption of these new techniques varied regionally, nationally, and even locally.

There was also variation in on site practice and maintenance which can give rise to temperature drift because of contamination, screen discolouration, or fabric degradation among other issues (Van der Meulen and Brandsma, 2008).

2.1.4 Automation, the growth of third party data providers, and increased heterogeneity in recent decades

Since the latter half of the 20th Century instrument manufacturers have increasingly developed semi-automated or fully automated sensors, deployable at low-cost, often with autonomous power solutions and telemetry (Eden, 2009). A spin off of this development was a proliferation of increasingly sophisticated personal weather stations, resulting in meteorological measurements being made by an increasing number of public bodies, private enterprises and citizens. There thus exists a greatly increased range of available observations, taken under an increasingly diverse range of auspices, siting, and methods of observation (Meier et al., 2017).

These observations are often uploaded to web sites such as Weather Underground, the UK Met Office Web site (weather WOW) or the Citizen Weather Observer Program (CWOP) to name the three most popular. Although quality checks are generally performed, the details of these checks are not always made public (Butler, 2018). In 2012 there were over 400 amateurs in the UK and Ireland uploading data to the UK Met Office web site (www.metoffice.gov.uk) and over 1350 contributing to the Weather Underground site (www.wunderground.com) (Bell et al., 2015).

Weather Underground now has more than a quarter of a million amateur subscribers worldwide providing weather data from privately owned automatic weather stations (AWS). Eden (2009) emphasises that the adjective “amateur” in this context refers not to inexperienced practitioners, but rather those that are not employed as professional observers, and notes that all our long term records were compiled by “amateurs” up to the early twentieth century. These records are no less valuable as a result (Eden, 2009). The introduction of these automated measurements into observation networks enhances long term observations insofar as that continuous and uninterrupted monitoring can be achieved and observations can be made in remote, uninhabited and inhospitable locales and extend the area of coverage considerably, providing a much more comprehensive, often higher quality (Hunziker et al., 2017)

and extended global network of observations going forward (Milewska and Hogg, 2010).

Investigations of the performance of off the shelf AWS (Figure 2.9) equipment that comes with reported performance variations of less than 0.3 °C found that when these units are operated under test conditions to manufacturers specifications their performance is much improved on that of many old max/min thermometers in shelters (Meier et al., 2017, Fenner et al., 2017, Lagouvardos et al., 2017).



Figure 2.9 A typical amateur weather station cheaply available that is used to upload to various weather sites such as weather underground.

But very often poor maintenance, siting and failure to operate as per manufacturer's instructions in the field have led to a high degree of uncertainty (Meier et al., 2017, Fenner et al., 2017, Bell et al., 2013). Furthermore, where automatic systems with improved performance have been introduced into official networks their introduction has often resulted in the introduction of inhomogeneities into the data record (Brandsma and van der Meulen, 2008). This is because the new instruments have different systematic and random uncertainties than those they replace. For example, when changes were made in the USA from liquid in glass to electronic resistance thermometers this gave a spurious cooling of the maximum temperatures and warming of the minimum temperatures (Quayle et al., 1991, Wendland and Armstrong, 1993). However, as the transition was also normally associated with a

site relocation due to the need for access to a power supply the situation is more complicated than just equipment change over, and each site had other additional issues (Williams et al., 2012).

For these reasons, the Commission for Instruments and Methods of Observation (CIMO) have recommended periodic international comparisons of commonly-employed temperature sensor & screen combinations. One such experiment was conducted at De-Bilt in the Netherlands between 1989 and 1995 (Figures 2.10 and 2.11) (Brandsma and van der Meulen, 2008). These controlled studies of the differences between the liquid in glass thermometers and electronic sensors found differences of up to approximately 0.2°C , with a standard deviation of 0.6°C . The response time, accuracy, the siting of the equipment, sensitivity of the equipment, and the calibration practices all contribute to these differences (Milewska and Hogg, 2010, Warne, 1999, Bell et al., 2013).

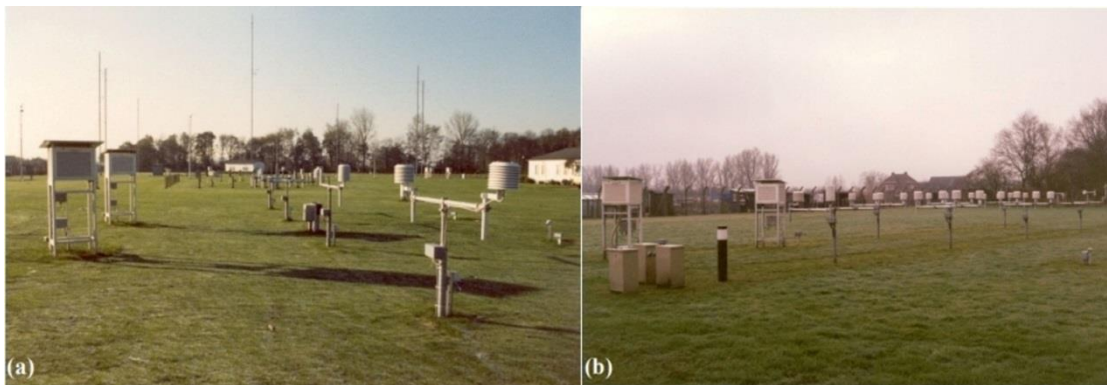


Figure 2.10 De-Bilt experiment (a) 1989-1991, (b) 1992-1995, in which ten screens were compared. The screens operated in parallel with a reference screen for two years. Both images show traditional Stevenson type screens and new multiplate screens under extensive testing.

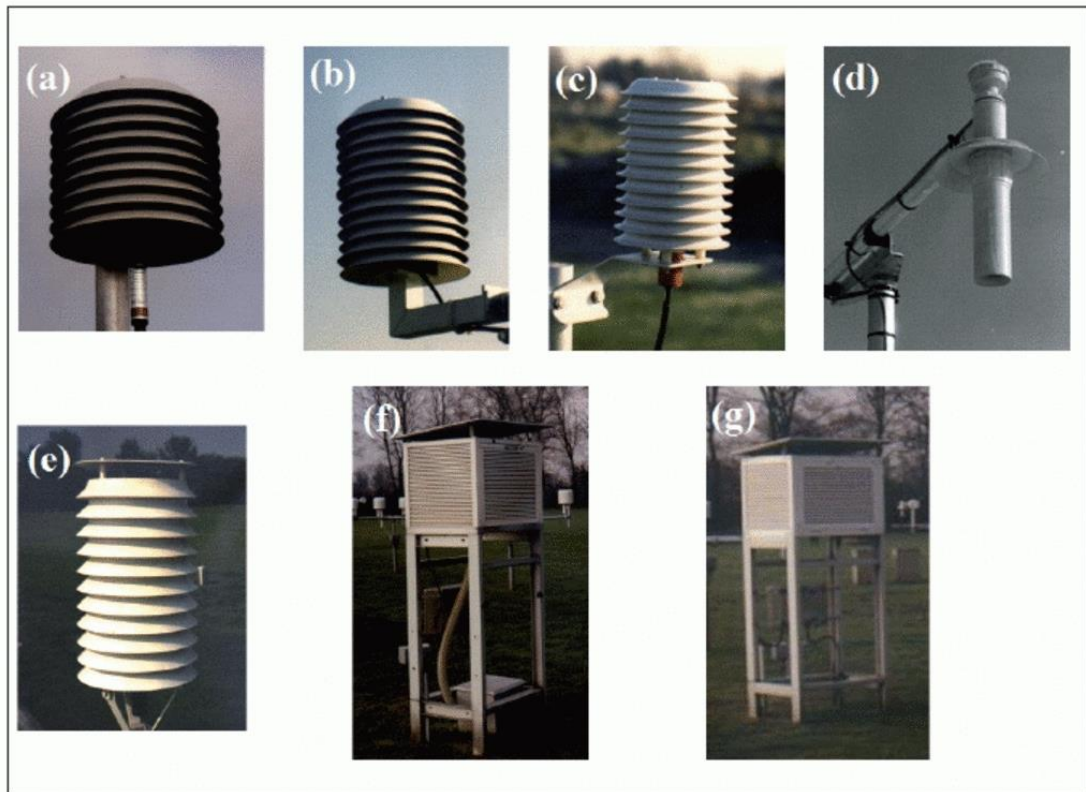


Figure 2.11 Different screens tested at De Bilt: (a) KNMI multiple, (b) Vaisla multiplate, (c) Young Gil Multiplate, (d) Young aspirate (type 1 and 2), (e) Socrima multiplate, (f) Stevenson Screen P.V.C version, (g) Standard wooden Stevenson screen (Van der Meulen and Brandsma, 2008)

2.2 Database management efforts and the International Surface Temperature Initiative databank

Over time, there have been numerous efforts to gather and curate observational records, including the World Weather Records (WWR) initiative that started in 1923 (Peterson and Vose, 1997). But broad-scale international cooperation and the sharing of data only really comprehensively evolved in the latter part of the 20th Century with increases in connectivity and computing capability that enabled the storage and sharing of data. Rights holders concerns over data sharing, combined with short-term funding often concerned with single variables, regions and timescales, have led to a fragmented approach to the sharing of records over time. Individual station's observations, if retained at all, were often stored locally or at best in regional or national archives and as a result, historically many meteorological records have been scattered, neglected and even lost. What data are available have been curated in

multiple national, regional and global repositories and in a broad variety of different formats (Smith et al., 2011). The lack of coordinated curation has led to much sharing and duplication that means the same data may be present in multiple archives, often processed distinctly and with differences in the associated metadata such as station names and identifiers as well as coordinates (Rennie et al., 2014).

Early studies of climate change include Kincer (1933) who concluded that the North American climate has warmed over the previous decades. But the very earliest compilation of a globally representative set of measurements was undertaken by Guy Callendar (Figure 2.12) who in 1938 made use of the Smithsonian World Weather Records (Callendar, 1938) and relied upon manual transcription of data (Hawkins and Jones, 2013). There followed several further efforts, including an updated analysis by Callendar himself (Callendar, 1961), but these generally lacked truly global coverage. The first truly global database efforts with global and regional data coverage close to what we have today were undertaken by the UK Climatic Research Unit (under contract to the US Department of Energy) (Jones et al., 1985b) and with the 1992 establishment by the US National Oceanic and Atmospheric Administration of the Global Historical Climatology Network (GHCN) (Lawrimore et al., 2011).

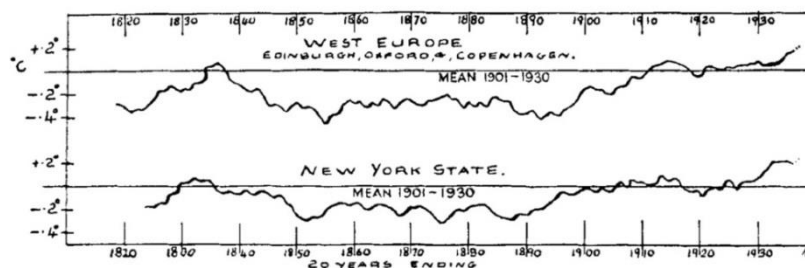


FIG. 3.—The most reliable long period temperature records. Twenty-year moving departures from the mean, 1901-1930.

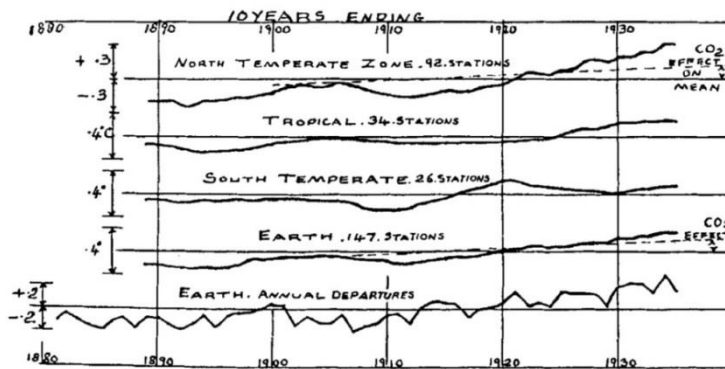


FIG. 4.—Temperature variations of the zones and of the earth. Ten-year moving departures from the mean, 1901-1930, °C.

Figure 2.12 Guy Callendar published the first estimation of global temperature changes in 1938 (Callendar, 1938)

However, it has long been known that many series exist that were not available in these early compilations. For example, in 1997 it was estimated that globally over 100,000 locations were collecting or have collected monthly mean surface temperature observations at some point in time, and often for extended periods (Peterson and Vose, 1997) and that much of this data remained to be rescued. Much of the early meteorological data remains in paper or image form only (Allan et al., 2011). This data must be digitised, which is a slow and painstaking process (Brunet and Jones, 2011). Historically, data rescue has been carried out in a broad variety of manners, primarily through National Meteorological Services, interested individuals, and by funded projects such as the EMULATE daily temperature and precipitation project (Brunet and Jones, 2011). More recently, there has been a proliferation of approaches. Dedicated citizen scientists can now carry out much of this work via projects such as weatherrescue.org (Hawkins et al., 2019). Furthermore, efforts to integrate data rescue into the classroom have been pioneered by Maynooth University (Ryan et al., 2018). The Atmospheric Circulation Reconstructions over the Earth (ACRE) Initiative are involved in coordinating many data rescue projects (Allan et al., 2011) and have been responsible for a wealth of new data provision over the past decade or so. Nevertheless, much data remains to be rescued.

A proposal to the World Meteorological Organisation's Commission for Climatology in 2010 for a comprehensive data bank of land surface holdings leading to new and improved estimates of changes in surface temperatures received full support and resulted in the establishment of the International Surface Temperature Initiative (ISTI) (Thorne et al., 2011, Lawrimore et al., 2015). The ISTI databank is an international effort to gather, merge and collate data from all identified sources to produce a set of historical data holdings (Rennie et al., 2014), thus providing an invaluable resource to interpret how the global climate has changed. The databank brings together in excess of 70 underlying data sources, many of which themselves consist of compilations of underlying sources (Thorne et al., 2011, Rennie et al., 2014, Lawrimore et al., 2015). It is the largest set of data holdings of monthly resolution land surface temperatures available to date containing over 35,000 individual station records that are as close as possible to the originally recorded values without homogenisation. Many records, however, are of short duration. This databank and its compilation is described further in Chapter 3.

2.3 Land surface air temperature homogenisation approaches

Most of the observations used in the study of climate were originally taken for non-climatic purposes and they include non-climatic influences of both random and systematic nature that cannot be unambiguously quantified in most cases (Dee et al., 2011a). Typically measurements were made to meet the needs of weather forecasting, agriculture, hydrology or other specific stakeholders, where systematic uncertainties and shifts may not have been of great importance for the original purpose of the observation (Williams et al., 2012, Vose et al., 2012). But for climate monitoring and reconstruction, it is necessary to distinguish the shifts caused by climatic factors, such as the eruption of a volcano, changes in the North Atlantic Oscillation and an El-Nino, etc, from the shifts caused by a change in the observational practices, siting or instrumentation over time. The term homogenisation is taken from Greek, meaning, to make everything similar. In climate science, it refers to the process of the application of adjustments to account for non-climatic factors that otherwise bias and obfuscate the record (Stepanek et al., 2013).

2.3.1 Breakpoint Detection

It has long been recognised that a time series is homogenised (or considered homogeneous) when the climatological variability is caused by variation of the weather and climate alone (Conrad, 1946). It would be rare for any long term meteorological station time series not to have experienced some form of change, such as a move, a replacement of instruments or another type of change that may introduce some form of bias into the record (Menne and Williams, 2009, Guttman, 1998, Peterson et al., 1998, Aguilar et al., 2003). These biases are often referred to as breaks in the time series. Perhaps the most discussed sudden breaks are associated with station moves (Jones and Briffa, 1992, Trewin, 2010). While the most common trend-type effect is associated with the incremental growth of towns and cities. This growth gives effect to gradual warming, via the urban heat island effect (Rohde et al., 2013a, Van der Meulen and Brandsma, 2008). Night-time temperatures tend to be particularly biased as urban fabric releases stored solar energy at night (Chun and Guhathakurta, 2016, Sahin and Cigizoglu, 2010, Brunet et al., 2006). The

acceleration in aviation over the 20th Century often led to the relocation of stations from expanding urban centres to out of town airports that manifest itself as an upward trend followed by a sudden break downwards when the station is relocated to the airport (Trewin, 2010). In the real world, breaks are believed to occur on average every fifteen to twenty years (Freitas et al., 2013, Venema et al., 2012) across the global network taken as a whole, although at individual locations this varies widely, and this estimate is tacitly recognised to be poorly constrained.

Homogenisation methods can be very broadly divided into direct and indirect methods (Ribeiro et al., 2016, Aguilar et al., 2003). Direct methods would be the ideal world scenario, whereby all consequential changes to an observation system would be diligently recorded and there would be a period of parallel measurements that would directly quantify any introduced biases (Peterson et al., 1998, Aguilar et al., 2003). But sadly in the real world, this is very rarely if ever the case, and to compound matters yet further metadata is rarely complete and detailed. This impedes the identification and removal of non-climatic artefacts in the record and the norm is to have to resort to indirect methods (Thorne et al., 2005b).

Indirect homogenisation begins with breakpoint detection which involves searching for statistical evidence of changes in the station mean, variance, or both (Venema et al., 2018). Absolute homogenisation techniques look at station series in isolation whereas relative homogenisation techniques consider differences to a reference with shared variability - typically surrounding locations series. The earliest homogenisation methods generally considered absolute homogeneity and relied on tests to check the non-stationarity of a single climatological series assuming that the climate is stable (Mamara et al., 2012). Such methods should be avoided since this assumption is unrealistic (Guijarro, 2014, Mamara et al., 2012). These methods are seldom used today, having been replaced with relative homogenisation methods (Peterson et al., 1998, Peterson and Easterling, 1994, Karl and Williams, 1987). These methods are predicated on the assumption that, for example, surrounding stations experience broadly the same climate signal as the candidate station and that any deviations between pairs that are not explained by a constant climatological offset constitute a breakpoint (Costa and Soares, 2008) in either the candidate or the reference. The time series is considered homogeneous if the mean of the difference series does not vary significantly through time (Steffensen et al., 1993,

Alexandersson and Moberg, 1996). The larger the variance in the difference series the more difficult it is to detect smaller breaks irrespective of the choice of breakpoint detection test to be employed (Venema et al., 2018, Venema et al., 2012). If multiple references are used via a series of pairwise comparisons, then a subsequent process of logical elimination can place the breaks in the correct stations (Domonkos and Coll, 2017, Ribeiro et al., 2016).

For a station(s) to act as a reference for breakpoint detection, it needs to covary with the target series (Menne and Williams, 2009). The greater the distance that two stations are separated by, the more the correlation on average decays (Moberg and Alexandersson, 1996). There is, however, little consensus on what may constitute a reasonable cut-off distance. Recommendations vary widely with examples including 1000 Km (Wang et al., 2018), up to 1200 Km (Hansen et al., 2010) or where the correlation falls below 0.36 (New et al., 1998). But suitability as a neighbour depends not only on distance but on other issues such as elevation, land use, etc. (Willett et al., 2014). There are wide-ranging approaches to the selection of an appropriate number of comparator series, with approaches ranging from a handful of stations to upwards of 40 sites selected from a larger pool based upon an appropriate combination of time series availability and correlation (Peterson and Easterling, 1994, Haimberger et al., 2008, Menne and Williams, 2009). Today with ever-improving computing power, pairwise comparisons using multiple references are the norm in state-of-the-art homogenisation approaches (Venema et al., 2012).

Figure 2.13 is a simplified example of pairwise comparison. In this process, the hypothetical test station A, (top panel) appears to have a break in c.1940 which can be ascertained via three-way comparison. If the anomalies of station A are compared against highly correlated reference stations B and C (middle panel) the break can be seen, whereas it is absent in the B-C pair, placing the likely break in A. In real examples, the break will unlikely be so obvious.

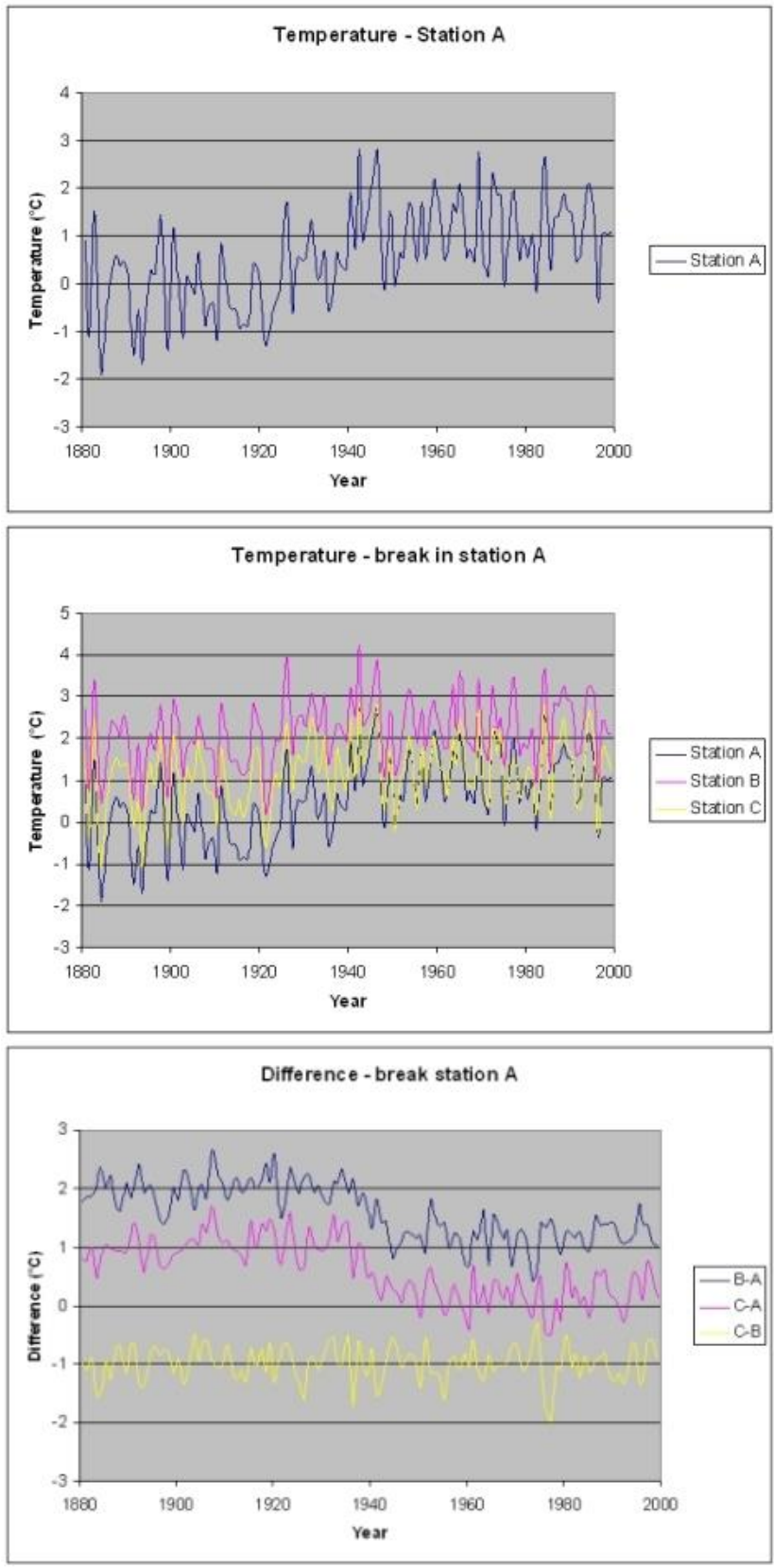


Figure 2.13 An elementary example of homogenisation by pairwise reference. The chart tells that a break around 1940 can be seen in the B-A series and in the C-A series but not in the C-B series. The break probably occurred in A (Sourced from <http://variable-variability.blogspot.com/2012/08.html>)

Most breakpoint detection techniques were developed in regions of high station density such as Europe and the USA (Gubler et al., 2017) where the construction of a reference series of sufficient longevity to detect breaks is less challenging than many remaining areas of the globe. In less well sampled regions where the distance to neighbours may approach or even exceed thousands of Km, the skill of existing methods may be lower (Hunziker et al., 2017). The same challenge is present when attempting to identify breakpoints in time series from the early period of global records when there were far fewer stations available (Aguilar et al., 2003) and even the building of a reference series from segments of data from different stations that are highly correlated may not be possible (Menne et al., 2009, Menne and Williams, 2009). Even when abundant highly correlated sites are present the reliability of the reference series cannot be proven in an absolute sense (Caussinus and Mestre, 2004, Hanssen-Bauer and Forland, 1994, Tuomenvirta, 2002).

Numerous relative homogenisation breakpoint detection tests have been proposed over time, including Potters test and the Standard Normal Homogeneity Test (SNHT) and many of these are summarised in e.g. Peterson et al. (1998). All these tests search for a change in mean and/or variance, and most do so via iteratively applying the test across a fixed-width window of points passing through the series. This means that breakpoints at each end of the series are undetectable. Peterson and Easterling (1995) developed a two phase regression method following Solow (1987). Vincent (1998) employed a technique using a multiple linear regression (MLR) process. It was considered a robust method, scoring marginally higher than the SNHT in a comparison of homogenisation techniques (Ducre-Robitaille et al., 2003). But in efficiency tests its detection skills fell below that of the SNHT (Domonkos, 2011, Domonkos et al., 2012).

The SNHT test (Alexandersson and Moberg, 1996) performs well, particularly in well sampled regions, scored highly on benchmarking tests (Venema et al., 2012) and has become popular (Freitas et al., 2013, Ducre-Robitaille et al., 2003). It has been used in several studies of climate homogeneity (Hanssen-Bauer and Forland, 1994, Slonosky et al., 1999, Tuomenvirta et al., 2000, Klingbjør and Moberg, 2003, Morales et al., 2005). Developed and applied in the first instance to precipitation (Alexandersson, 1986), it makes use of nearby highly correlated neighbouring sites. Because it makes use of multiple sites for the formation of a reference series for

pairwise comparison it can use incomplete neighbouring data series (Alexandersson, 1986). But it can also be carried out using a composite neighbour reference series or another comparator series, even a single reference if there is confidence in the homogeneity of that series (Peterson and Easterling, 1994, Moberg and Alexandersson, 1996). The SNHT can be used efficiently when metadata is unavailable (Ducré-Robitaille et al., 2003) and it is therefore very suitable for automation (Menne and Williams, 2009) and for the homogenisation of large data collections where manual decisions are not feasible (Ribeiro et al., 2016, Aguilar et al., 2003).

Ribeiro et al. (2016) reviewed thirty one homogenisation methods that included the MLR and the SNHT for breakpoint detection and provides a summary of each method's strengths, weaknesses, and in what application(s) they perform well. They find that both MLR and SNHT perform best overall in a range of situations.

2.3.2 Adjustment of series at identified breakpoints

Once breakpoints have been identified, adjustment is carried out to complete the homogenisation procedure. In its simplest form, the mean of the difference series segment before the break is subtracted from the mean of the segment after the break (Alexandersson and Moberg, 1996). Other methods enforce stricter criteria, for example, five continuous years of data without another break to be available each side of a break before an adjustment value can be calculated and applied, or a minimum number of observations to be available within a defined period each side of the break (Trewin, 2018).

The adjustment is generally applied back through the series so that all historical estimates are made equivalent to the most recent homogeneous segment. This makes new measurements directly comparable to older measurements enabling ongoing measurements to be seamlessly compared with those taken in the past. Multiple breaks can be adjusted by consecutive application of the approach multiple times (Tuomenvirta, 2002). Adjustments are generally applied as seasonally invariant deltas to the series. Though this may be problematic under certain circumstances, for example, if there is a high degree of seasonality present in early record biases owing

to solar radiation exposure influences (see earlier sections for discussion and references).

In most cases today pairwise adjustment estimates are estimated using a multitude of reference stations that may or may not be weighted (Menne and Williams, 2009). Multiple potential adjustments are calculated, one for each neighbouring station in use. Menne et al. (2009) suggest that the median estimate from the population of estimates may be more robust than the mean. Others propose using all available adjustments estimates by weighting each based on correlation or perhaps distance (Hanssen-Bauer and Forland, 1994, Steffensen et al., 1993, Hausfather et al., 2016).

2.4 Assessing the strengths and limitations of modern homogenisation approaches

To determine the strengths and weaknesses of a homogenisation method, it must be tested against a benchmark (Willett et al., 2014). A long term series where the true solution is known is required to test homogenisation methods against (Vincent, 1998). This is not naturally available, thus it is necessary to construct a synthetic fully homogeneous time series and introduce into that timeseries inhomogeneities of known size and location that mimic real data issues. Synthetic series can then determine the skill of the methods predicated upon the assumptions underlying the construction of the synthetic series (Willett et al., 2014). Several such tests have been undertaken over the years (Ducré-Robitaille et al., 2003, DeGaetano, 2005, Domonkos and Stepanak, 2009, Williams et al., 2012, Venema et al., 2012).

The European Cooperation in Science and Technology (COST) Action ES0601 undertook the most recent and comprehensive test of several popular homogenisation algorithms (Venema et al., 2012). During this test, an open invitation to climatologists was given to perform a blind homogenisation of synthetic time series that Venema et al. (2012) prepared from true time series taken from Austria, France and Catalonia that were homogenised and detrended. Artificial breaks were then inserted. Venema et al. (2012) evaluated the returned homogenised series. They found that all relative automatic homogenisation methods improve the homogeneity of temperature data series (Domonkos et al., 2012, Van der Meulen and Brandsma, 2008). But Venema et al. (2012) noted caveats regarding putting in breaks that

might have been too large (mean absolute break size of 0.8°C compared to an apparent average of 0.6°C in real data).

DeGaetano et al (2005) found that the SNHT performs best at placing breaks in their correct position. But the ability of all methods to detect breaks declines significantly when the break size falls below 0.6 standard deviations of the variance of the tested series. Williams et al. (2012) undertook benchmarking of the Menne and Williams (2009) method across the much broader US network based upon synthetic series derived from climate model simulations and found somewhat lower skill than claimed in Menne and Williams (2009) which was critically dependent upon assumptions of the propensity and structure of data inhomogeneities. Those synthetic series dominated by smaller breakpoints were, unsurprisingly, the most challenging to the method. Under the assumption that, in general, station operators will try to minimise the impact of any changes at their station, it is not out of the question that most breaks in station series might be small.

All the benchmarking studies discussed above show that the best methods tend to adjust the records towards the correct value but do not, on average, move the records sufficiently far to fully account for the known effects. An alternative approach is to compare against an independent series of known absolute quality. Sadly, such comparator series are a relatively recent innovation and are available at a regional scale only over the USA from the US Climate Reference Network(USCRN). Diamond et al. (2013) and Menne et al. (2010) undertook comparative studies of several US Historical Climatology Network (USHCN) sites to nearby USCRN stations. Some of these sites were classified as poorly sited and some well sited relative to guidance from WMO's Commission for Instrumentation and Methods of Observation. The poor siting of many stations was a direct result of site location compromises that came about when changing from the liquid in glass thermometer (LIG) hosted in Cotton Region Shelters to electronic max/min thermometers (MMTS) that required a continuous power supply and were manufactured with a short electrical lead. Menne et al. (2009) found that once homogenisation with their PHA method was undertaken that the interpolated mean, max, min values averaged to a 0.25° latitude and 0.25° longitude grid for both groups were both directly comparable and in very close agreement with the USCRN. A later study by Hausfather et al. (2016) confirmed this close agreement (Figure 2.14).

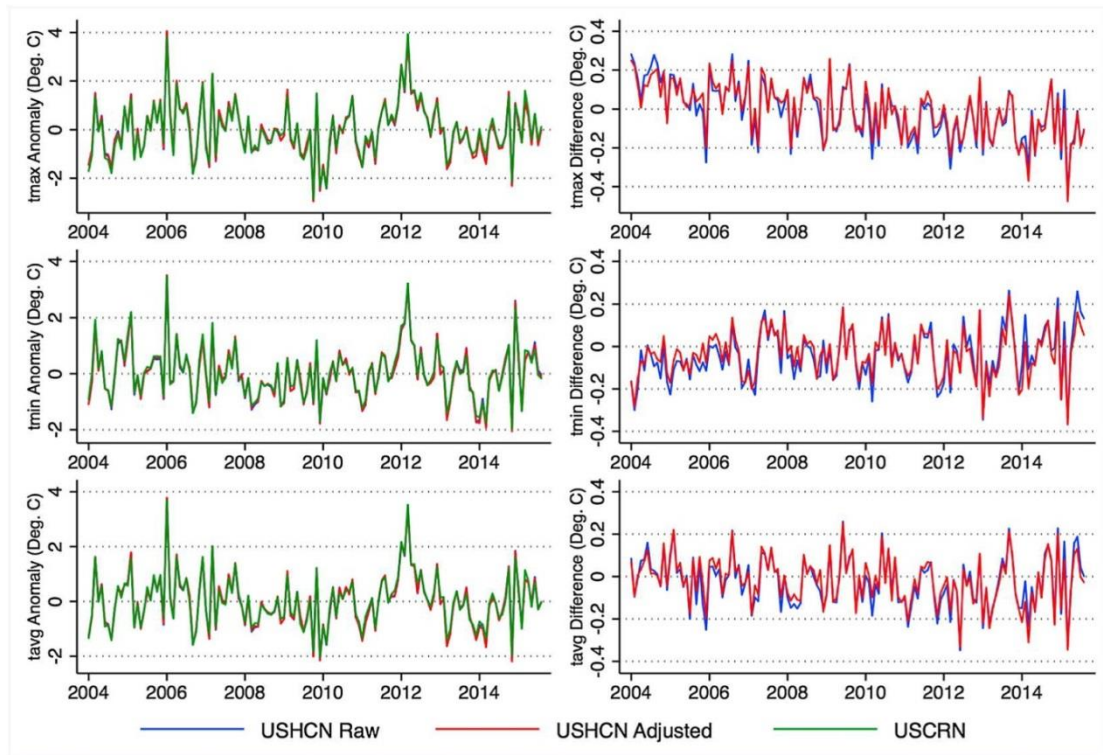


Figure 2.14. The maximum (t_{max}), minimum (t_{min}) and mean (t_{mean}) of the USCRN, USHCN raw, and USHCN adjusted data left column. Right column USHCN raw series minus USCRN (blue) and USHCN adjusted minus USCRN in red (Hausfather et al., 2016).

2.5 Brief Summary of state-of-the-art land surface air temperature datasets

Notable current global datasets of land surface air temperature include CRUTEM now at version 5 (Osborn et al., 2020). GHCN-M now at version 4 (Menne et al., 2018), NASA GISS (Lenssen et al., 2019), the Berkeley Earth dataset (Rohde et al., 2013a) and the Chinese Global Land Surface Air Temperature data set (CMA-LSAT) (Xu et al., 2017). Key characteristics of these products are summarised in Table 2.1 and a comparison of several of the products is given in Figure 2.15. This highlights the degree of correspondence of these products which agree strongly on year-to-year variations. On centennial timescales the NOAA and NASA products, both based upon GHCNMv4, estimate greater warming than CRUTEM and Berkeley Earth over the common periods of record.

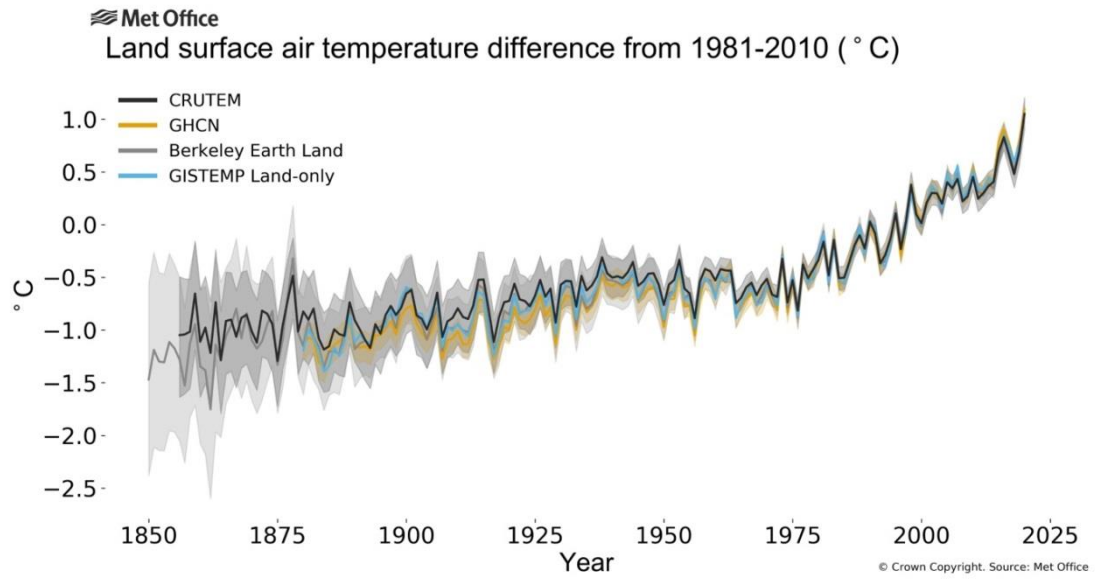


Figure 2.15. Comparison of several LSAT datasets normalised to 1981-2010 (<https://www.metoffice.gov.uk/hadobs/monitoring/temperature.html>)

Dataset	Start of Record	Spatial resolution	Climatology Period	Number of Sources used	Number of records used	Homogenisation method	Interpolation	Reference
GHCNMv4	1850	5° by 5°	1961-1990	1 ISTI Version 1.1.0	27,231	Pairwise Homogenisation Adjustment. Breaks detected by the SNHT and adjustments made by pairwise comparison to neighbouring stations	None	(Menne et al., 2018)
GIS Temp	1880	2° by 2°	1951-1980	1 GHCNM v4	Approximately 26800	As in GHCNMv4 with some additional post-processing adjustment based on nightlight technique & some additional Quality Control	Weighted interpolation & average out to 1200 Km, based on latitude band and 80 subgrids	(Lenssen et al., 2019)
CRUTEMv5	1850	5° by 5°	1961-1990	104	10649 (Initial dataset) 8,000 after QC etc.	Homogenised by the institutions and groups providing the data.	None	(Osborn et al., 2020)
Berkeley	1753	1° by 1°	1951-1980	14	44,455 initial dataset with 36,866 remaining after removal of short or otherwise useable data series	Berkeley separate time series into separated independent segments at identified breaks and treat these segments as an independent dataset. They call this process "Scalping"	Gaussian process regression/kriging	(Rohde et al., 2013)
CMST_LSAT	1900	5° by 5°	1900-2017	14	13,687	50% of the data was homogenised by the institution providing the data. The remainder was homogenised by Penalized Maximal t-test	None	(Yun et al., 2019)

Table 2.1 A summary of the key characteristics of the five modern global gridded datasets highlighting similarities and distinctions between them.

In building GHCNMv4, (Menne et al., 2018) used the records from the ISTI database (Rennie et al., 2014) that contained more than 10 years of observations. They discovered and eliminated a degree of duplication that slipped through the ISTI process (Jared Rennie, Pers Comm). Quality checks include checks for outliers and spatial consistency used in previous versions (Lawrimore et al., 2011). Homogenisation was carried out using the Pairwise Homogenisation Adjustment algorithm (Menne and Williams, 2009). This algorithm uses the SNHT to identify breakpoints (Alexandersson, 1986) and identified about 70,000 shifts in total across the network that were then adjusted using apparently homogeneous neighbour segments. GHCNMv4 is presented without interpolation into data sparse regions, although in the NOAA Globtemp product interpolation is undertaken (Zhang et al., 2019).

GISTemp (Lenssen et al., 2019) uses the GHCNMv4 dataset. GISTemp continues to follow Hansen et al's (1999) protocol of only carrying out minor additional homogenisation steps. Most of these additional adjustments that were made are based on the "Nightlights" procedure, from satellite imaging. To adjust these urban stations, they were paired with rural neighbours, If fewer than three suitable neighbours were present to compare against, the urban station is removed (Hansen et al., 1999). The global impact of this adjustment was found to be less than 0.01°C (Hansen et al., 2010, Hansen and Lebedeff, 1987). Gridbox averages are calculated for all possible grid boxes using data from stations within 1200 km, linearly weighted. These averages are estimated for 160 boxes of equal area and combined into latitude zones before global gridding is completed.

CRUTEM 5 (Osborn et al., 2020) is the latest version of CRUTEM, the land surface component of HadCRUT. It continues to adopt their tradition of relying upon data that is homogenised by the providing parties. The view is taken that local knowledge and access to metadata enable better homogenisation than using a consistent set of statistical algorithms (Jones et al., 1985b, Jones et al., 1986, Osborn et al., 2020). This current version has increased the number of stations considered to 10649 from the 5583 of CRUTEM4, with a commensurate increase to 8000 stations with sufficient data for gridding, up from 4842 in CRUTEM4 (Osborn et al., 2020). In CRUTEM5 two gridded forms are made available. The original method that undertakes no interpolation and an alternative method that allows for stations in the

“higher latitude regions” to contribute to a grid cell value even if they are outside that grid. Osborn et al (2020) do not explicitly define high latitudes in this context, but previous versions of CRUTEM defined high latitudes as 70° to 90° N and 50° to 90°S. This is justified by the narrowing of the grid boxes as latitude increases.

The Berkeley Earth dataset extends back furthest in time, reaching back to the 1700s. It also makes use of far more stations, 44,455 apparent stations in all. Berkeley merged 14 data sources, the largest being GHCN daily and GHCN monthly version 3 (Rao et al., 2018). Rohde et al. (2013a) divide stations at breakpoints that were identified using a pairwise comparison with neighbours into individual segments. They called this process “Scalpel” and these different segments are treated separately as if they were individual stations. This gave rise to over 179,928 independent station fragments. Any fragment containing one year or less of data was removed. Berkeley then divides the globe into 15,984 regions. Station temperatures are interpolated into these boxes using kriging interpolation. These values are then re-gridded into 1° by 1° grid boxes. For each month Berkeley calculates an estimated monthly global anomaly based on kriging interpolation of the station fragments (Rohde et al., 2013a).

CMA-LSAT follows the CRUTEM philosophy on homogenisation. In all 14 different sources were used to build the dataset including CRUTEMv4.6, GHCNMv3 and Berkeley Earth. Other sources included three regional data sources and eight national sources. Testing for duplication follows a similar process to that used in the construction of the ISTI databank. Homogenised data acquired from National Meteorological Services made up about 50% of data sources and these took priority and were integrated without any additional changes. The remaining sources of lower priority acquired to improve coverage of data over China and neighbouring countries were not homogenised beforehand by the data providers (Xu et al., 2017). This meant that 50% of the data sources used had to undergo homogenisation prior to integration. The homogenisation process applied was a Penalized Maximal t-test where the critical penalty factor was empirically constructed using ratios (Wang et al., 2007). A total of 9765 stations dataset were used in the final gridding process, 8300 of these from the northern hemisphere and 1465 from the southern hemisphere. In a deviation from the CRUTEM methodology, stations with fewer than 10 years of data were included (Xu et al., 2017).

2.6 Uncertainty characterisation

2.6.1 Theory

At the highest level uncertainty in global LSAT estimates can be split into two broad categories: Parametric uncertainties and structural uncertainties. Parametric uncertainties are those uncertainties that arise from inherently uncertain choices within the methodological framework adopted by each group. Whereas structural uncertainties are highly uncertain and can only be imperfectly estimated from the range of available estimates which themselves are a very finite sample of the full range of potential scientifically defensible methodological approaches to performing homogenisation (Thorne et al., 2005a).

Considering first structural uncertainty. In reality, there are fewer than the implied 5 degrees of freedom. All five datasets share some degree of commonality in terms of data sources and some share either some (CRUTEM and CLSAT) or almost all (GHCNMv4 and GISTEMP) of the homogenisation choices in common (Section 2.5). Despite these caveats Figure 2.16 clearly highlights distinctions between the products even when globally aggregated. The principal aim of the present thesis is to improve the sampling of structural uncertainty by introducing and assessing a methodologically independent approach to homogeneity assessments.

Quantification of parametric uncertainties in land surface temperature datasets is predominately divided into three types: station uncertainty, sampling uncertainty and bias uncertainty (Brohan et al., 2006, Lenssen et al., 2019). The first work on parametric uncertainty was carried out by Brohan et al. (2006) on CRUTEM3 and later built on by Morice et al. (2012) to produce an ensemble approach. Most estimates assume that these terms are independent and thus can be combined in quadrature.

Station uncertainty includes systematic and random components, including all handling of the data at station level, transcription, record adjustment and post-processing, including homogenisation uncertainty. Following Brohan et al. (2006) station uncertainty can be expressed as:

$$A_{actuals} = T_{obs} - T_N + \epsilon_N + \epsilon_{obs} + C_H + \epsilon_H + \epsilon_{RC} \quad \text{Eqn 2.1}$$

Where: $A_{actuals}$ is the true station monthly mean anomaly.

T_{obs} is the observed station value.

T_N is the estimated station normal.

ϵ_N is the uncertainty associated with calculating this climatology

ϵ_{obs} is the uncertainty associated with measurement.

C_H is the adjustment added in the homogenisation of $T_{actuals}$

ϵ_H the uncertainty associated with the calculated adjustment.

ϵ_{RC} is the uncertainty associated with miscalculating and /or misreporting the monthly mean temperature.

Because all the station uncertainties are independent and specific to each station, the station uncertainty components can be combined in quadrature when multiple stations contribute to a grid box or areal average. Uncertainties arising from individual stations thus become increasingly negligible with increasing spatio-temporal aggregation.

Sampling uncertainty within a grid box is the difference between the true grid box mean value and the estimated value arising from the finite sampling of the grid box by available observations. It is a function of the number of stations, the climate variability in the grid box, and the position of the stations within the grid box.

Following Jones et al. (1997) sampling uncertainty can be expressed as:

$$SE^2 = \frac{\bar{\sigma}_i^2 \bar{r}(1 - \bar{r})}{1 + (n - 1)\bar{r}} \quad \text{Eqn 2.2}$$

Where $\bar{\sigma}_i^2$ is the mean station standard deviation.

n is the number of stations in the grid box

\bar{r} is the average inter-site correlation.

Gridbox sampling uncertainty thus rapidly diminishes as n – the number of stations contributing – increases.

There are also many land regions that are habitually unsampled. Some data products undertake interpolation to infill these regions, whereas others leave them as missing (Section 2.5). Either approach adds uncertainty in the estimate of the true globally complete mean which must be accounted for. Those datasets that do not attempt to interpolate typically quantify this uncertainty via recourse to experimentation using spatially complete fields from either climate models or reanalyses products (Morice et al., 2012, Menne et al., 2018).

Bias uncertainty is restricted to two main sources of uncertainty and accounts for: i) those small biases which may be undetected in homogenisation efforts – the so-called ‘missing middle’; ii) The urban heat island effect already discussed in detail in section 2.3.1 and thermometer exposure effects discussed in section 2.1.3. associated with the evolution of thermometer exposure practice from north wall mounting through to the standard practice of screens or aspirated instrumentation used today. True estimation of this combined uncertainty value would require detailed metadata data for all stations back to a least 1850, which is not available. Folland et al (2001) estimated that the error associated with the urban heat island effect is one sigma of $0.0055\text{ }^{\circ}\text{C}$ per decade (Brohan et al., 2006, Folland et al., 2001). Later work (Parker, 2004, Peterson, 2003, Peterson and Owen, 2005), broadly agrees with this estimate. For instrument exposure changes between 1900 and the present Parker (1994) estimated an uncertainty of $\pm 0.2^{\circ}\text{C}$. Folland et al. (2001) advanced Parker’s estimate and concluded that empirical estimation of uncertainty is 0.2°C for latitudes of 20°S to 20°N before 1930 and decreases linearly to zero in 1950. Outside the tropics, the uncertainty range is 0.1°C before 1900 decreasing also linearly to zero by 1930 (Brohan et al., 2006). These estimates are regionally based and may not be reflective of the true uncertainty for individual stations.

2.6.2 Derivation of uncertainty estimates

A significant innovation in recent years has been that parametric uncertainty estimates are now produced for most global LSAT products. However, these

estimates differ in their construction vis-à-vis which sources of uncertainty are considered and how they are quantified which hampers direct comparability.

Morice et al. (2012) extended the work of Brohan et al (2006) by producing a 100-member ensemble of ‘equi-probable’ CRUTEM4 estimates by taking realisations of the homogenisation error, normalization error and the bias associated with urbanisation and exposure uncertainties identified by Brohan et al. (2006). To generate each ensemble member for each grid box, Morice et al. (2012) first added a suite of small breaks ($\sigma=0.4C$) to reflect the Brohan et al (2006) estimate that undetected small breaks occur on average every forty years. Climatological uncertainty was then estimated from a standard distribution with a mean of zero and a standard deviation that depends on the number of years that a station had sufficient data over 1961-1990. Observation error is not included as it is random and it tends to quickly cancel with spatio-temporal aggregation (Willett et al., 2014).

Urbanisation and exposure uncertainties were added to the grid box anomalies because studies of these errors are based on regional impacts and not an individual station. The urbanisation error was assumed to be $0.0^{\circ} C$ before 1900, increasing linearly after 1900 and sampled from a gaussian distribution with a mean of zero and a standard deviation of $0.0055^{\circ}C$ per decade (Parker, 2011, Jones et al., 2008, Fujibe, 2009). The exposure error was estimated as per Brohan et al. (2006) and therefore for each individual ensemble member each month grid box anomaly is calculated as follows in Jones et al. (2012) :

$$A^{Land} = \left(\frac{1}{K} \sum_{n=1}^K T_{a[n]} \right) - \epsilon_u - \epsilon_e \quad \text{Eqn 2.3}$$

Where A^{Land} is the grid box temperature anomalies compute from K station anomalies within each grid box.

$T_{a[n]}$ Is the true station climatology adjusted anomaly

ϵ_u is the urbanisation uncertainty.

ϵ_e is the exposure uncertainty.

The entire ensemble generation process is as set out in Figure 2.16.

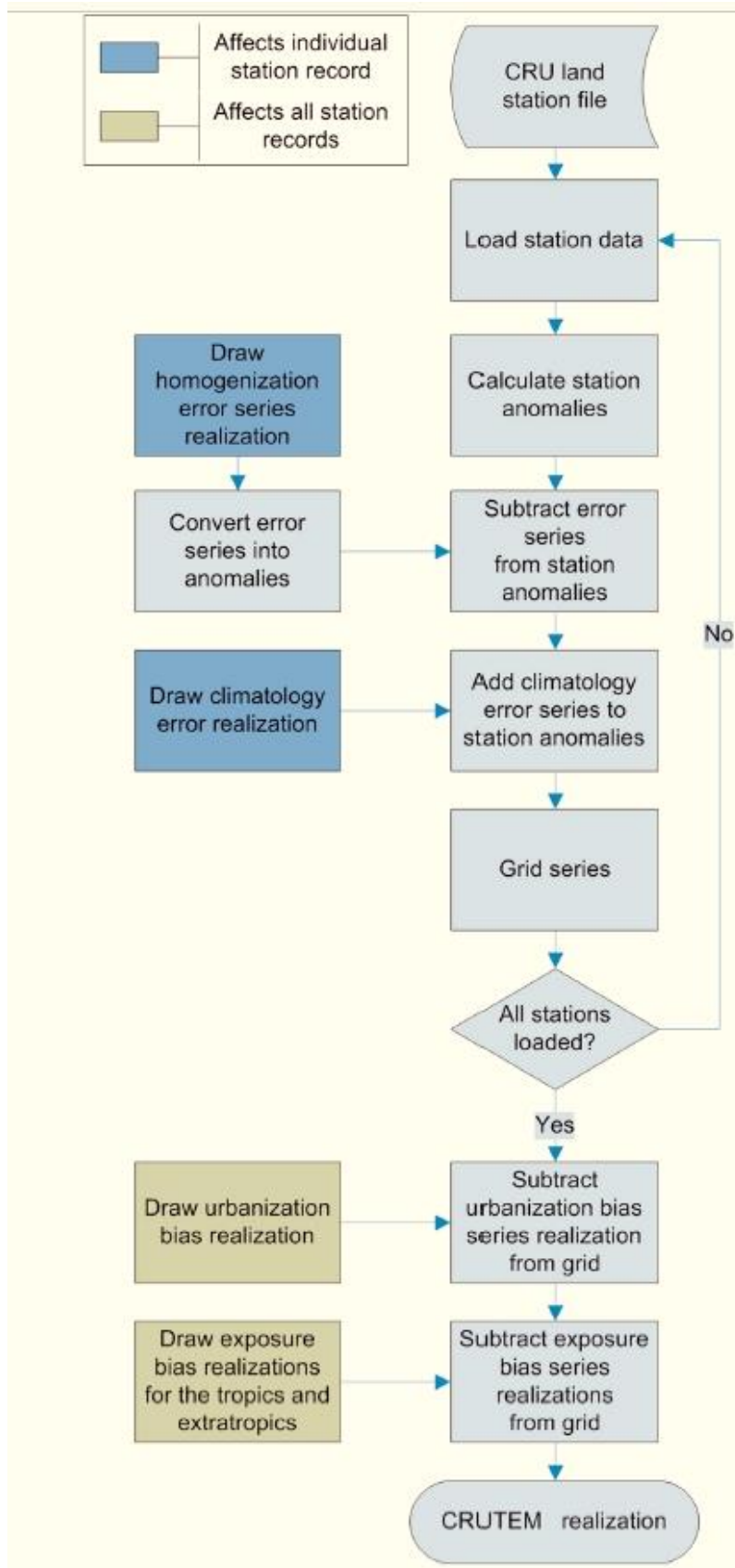


Figure 2.16 Flow of the ensemble generation for CRUTEM dataset ensemble taken from Morice et al (2012)

GHCMNv4 provides an ensemble of homogenisation estimates arising from 100 reasonable variations of the PHA algorithm settings (Menne et al., 2018).

Reasonable settings were selected based upon the benchmarking exercise of Williams et al. (2012). The uncertainty associated with incomplete homogenisation that is caused by the undetected small breaks of less than 0.2°C (the so-called ‘missing middle’) was taken care of by the addition of randomly seeded systematic offsets on average once every 50 years from a normal distribution with a mean of -0.01°C (to match the slight non-zero mean of detected breaks) and a standard deviation of 0.2°C that best infills the missing-middle effect.

Four additional uncertainty sources were then quantified:

(1) Station anomaly uncertainty that arises as a result of the interpolation exercise in individual grid boxes, when estimates were calculated for stations for which insufficient observations existed for the base period of 1961-90 by using neighbour stations’ data to infill in that period (Vose et al., 2014).

(2) Instrument exposure uncertainty accounted for by a random number drawn from a normal distribution applied on a latitude band basis following Brohan et al. (2006) and Morice et al. (2012).

(3) Gridbox sampling uncertainty arising from the finite sampling of the true grid box average following Brohan et al. (2006).

(4) Coverage uncertainty due to incomplete sampling, which is particularly an issue in the earlier years, estimated by comparing available area average anomalies with complete fields calculated from the spatially complete NCEP-NCAR reanalysis (Kalnay et al., 1995).

For GISTEMP, (Lenssen et al., 2019) collapse the uncertainties into three types:

(1) station uncertainty that includes systematic and random components, including all handling of the data at station level, transcription, record adjustment and post-processing, including homogenisation uncertainty. For homogenisation uncertainty, Lenssen et al. (2019) rely fully on the uncertainty model and estimates unmodified from Menne et al. (2018) discussed above.

(2) Bias uncertainty including the urban heat island effect estimated globally as 0.01 C° (Hansen et al., 2010).

(3) Sampling uncertainty, an overarching term to include calculation of global and region means from incomplete spatial and temporal records. In GISTemp this uncertainty is important due to the interpolation out to 1200 km. To calculate sampling uncertainty, Lenssen et al. (2019) compared the GISTemp anomaly estimates to MERRA (Modern-Era Retrospective Analysis for Research and Application) reanalysis as the prime comparator substantiated by ERA-5 (European Centre for medium range Forecast Reanalysis) and JRA-55 (Japanese 55 year Reanalysis).

Berkeley quantified two main sources of uncertainty: (1) Statistical/Data-driven uncertainty (errors in the data not reflecting the true values) (2) Spatial incompleteness uncertainty. To estimate the statistical uncertainty, Berkeley carried out two tests. The first test was to randomly separate all the data into five subsamples, calculate the anomalies, re-average the five subsamples and recalculate the anomalies using an arbitrary base period. They then compared the result of the five subsamples to the original estimate. The second test was to apply a jackknife statistics test (Tukey, 1958, Quenouille, 1949, Miller, 1974) to the GHCMNv3 dataset (Lawrimore et al., 2011). They randomly divided GHCMNv3 into 8 subsamples. Each contained 7/8 of the full GHCMNv3. All 8 subsamples were put through the Berkeley averaging process (Rohde et al., 2013b). The standard error was then calculated among these samples. Spatial uncertainty was determined empirically by calculating 1960 to 2000 average values for well sampled areas such as Europe and the US using data from stations that only existed in the past and then calculating the error arising by using the reduced network.

As is readily evident from the above discussions, where dataset creators have quantified uncertainty estimates a variety of approaches have been taken. However, most uncertainty models except for the Berkeley model are heavily influenced by the assumptions underlying Brohan et al. (2006) and Morice et al. (2012). Until recently the norm was to provide limited or no estimation of uncertainties, but now all but one product includes a substantive effort at uncertainty quantification. Available products differ in: what sources of uncertainty are quantified; how they are

quantified; and how they are presented to users. In addition, the availability of 5 datasets of varying degrees of methodological independence enables some exploration of structural uncertainty (Thorne et al., 2005a). Nevertheless, all products substantively overlap in their source data used and all use some form of neighbour based homogenisation techniques.

2.7 Reanalysis products

Reanalysis products consist of globally complete estimates of the atmospheric state back in time (Thorne and Vose, 2010, Bosilovich et al., 2012). They combine selected observations from past weather records using data assimilation techniques and modern NWP models to produce a physically consistent estimate of the atmospheric state through time which is predicated upon the model physics and the use of the available observational constraints.

Data assimilation and NWP systems are continuously improved and updated. The direct use of their contemporaneous analyses in a retrospective analysis of past climate would thus introduce inhomogeneities into the record, obscuring the true underlying climate trends. For this reason, each reanalysis effort uses a frozen NWP and data assimilation version to ingest and reprocess historical observations consistently (Slivinski, 2018, Kalnay et al., 1995).

Reanalyses products are an attractive option for many climate-related studies and applications as they provide spatiotemporally complete gridded data estimates of the state of the climate in a convenient format. Reanalysis and observations are not the same and must not be confused (Parker, 2016). However, observations provide an essential constraint to reanalysis. The quality of the observations impacts on the quality of the reanalysis products and the effect of the model biases depends upon the observational constraints which have changed dramatically through time (Bosilovich et al., 2012).

Currently, the reanalysis community are engaged in the production of a range of products. Up until recently, reanalysis products were domain specific, with separate reanalyses for land and ocean. Recent efforts have been made to produce coupled reanalysis products including CERA-20C (Laloyaux et al., 2018). Atmospheric

reanalyses are more mature than their ocean equivalents. The remainder of this section considers atmospheric reanalysis developments.

2.7.1 Full-input reanalysis products

Full-input atmospheric reanalysis products make use of the full range of time-varying observations available including those from:

- Surface synoptic stations;
- Radiosonde ascents;
- Aircraft; and
- Satellites.

The availability of observations of each type has varied dramatically through time. Globally representative radiosonde profile data are only available since the International Geophysical Year (1958). Early satellite data exist in the 1970s but operational meteorological satellites were only introduced in 1979 and have evolved substantively since with major step changes in capabilities in 1998 and then again around 2005. Aircraft observations, at least at scale, are more recent still. Furthermore, the types of instruments and measurements available have changed dramatically through time.

NCEP-NCAR (Kalnay et al., 1995) was the first reanalysis, released in 1995 with coverage from 1948 to the present. It was produced using a fixed version of the NCEP (National Center for Forecast Prediction) forecast model. It is a coarse resolution reanalysis product by today's standard with a lid at 3hPa. It was updated in 2001 using an improved forecast model as version R1 (Kistler et al., 2001) and again in 2002 as version R2.

ERA-40 (Uppala et al., 2005) was released by the European Centre for Medium Range Weather Forecasting (ECMWF) in 2003. It is a 45 year reanalysis product from 1957 to 2002. Prior to ERA-40 shorter reanalysis products were released by ECMWF such as ERA-15. ERA-40 like NCEP-NCAR assimilated data from a wide range of platforms but did not assimilate satellite data before 1973. It was much improved in quality and resolution from the earlier products. But it still had several

shortcomings including excessive precipitation over the tropics and too strong a Brewer Dobson circulation (Uppala et al., 2005).

ERA-Interim (Dee et al., 2011b) was released by ECMWF in 2008 as an upgrade to ERA-40. It used a newer version of the ECMWF IFS model and corrected several errors discovered in ERA-40. It was the first ECMWF Reanalysis to use 4D-VAR data assimilation which enabled exploitation of considerable additional data outside the assimilation time. Its coverage is from 1979 to 2019 when its production was ceased following the adoption of ERA5. ERA-Interim correspondence to observations are much improved particularly in the Southern Hemisphere compared to those of ERA-40.

ERA5 (Hersbach et al., 2020) is the successor to ERA-Interim and the most recent product from ECMWF. Currently available from 1979 to the present. A backward extension to 1950 has been completed and has been recently released in interim form, with plans to push back further to 1930 (Hersbach, pers comm). ERA5 benefits from ingestion of a broader range of input data sources and the use of a newer version of the IFS forecast model. It is available at a much finer horizontal resolution at 31 km than any preceding reanalysis product.

The Japanese Reanalysis JRA-25 (Onogi et al., 2007) was released by The Japan Meteorological Agency in 2006. It originally covered the twenty five years from 1979 to 2004 but was extended by ten years through to 2014. Its main data source is the ERA-40 input data but augmented with a number of predominantly Asian sources and, as a result, it improves the estimates over Asia. It benefits from lessons learned with ERA-40. The Japan Meteorological Agency released an update as JRA-55 in 2015 (Kobayashi et al., 2014). It employed a 4-dimensional variational assimilation scheme (4D-VAR) and extends coverage from 1958 to the present. The JRA-55 reanalysis has been used to validate the uncertainty estimation for GHCNMv4 and GISTEMP.

NASA's Modern Era Retrospective Analysis for Research (MERRA) was first released in 2009 (Rienecker et al., 2011) with the view of making use of NASA's earth orbiting systems. It was upgraded in 2015 (Gelaro et al., 2017) with MERRA-2. MERRA-1 was unable to ingest certain data types and this was a major factor in the upgrade.

To summarise, full-input reanalysis products have evolved and improved considerably through time. Each generation of products has benefitted from lessons learnt from the prior production as well as benefitting from improvements in the underlying forecast model and data assimilation schemes. Reanalyses are now available at better spatio-temporal resolution than ever before and represent some of the most heavily used and cited products in all of geophysical sciences.

2.7.2 Sparse-input reanalysis products

Sparse-input (often termed 20th Century) reanalysis products are a relatively new addition to the suite of reanalysis products (Compo et al., 2016). Pioneered by NOAA and the University of Colorado, they have now been produced also by ECMWF (Poli et al., 2016, Laloyaux et al., 2018, Slivinski et al., 2019) and are under preparation elsewhere (Compo, pers. comm.). Sparse input reanalysis products extend back to the 19th Century. Most specify fields of homogenised sea surface temperatures and sea ice concentration as a lower boundary condition (Titchner and Rayner, 2014, Rayner et al., 2005) and ingest solely surface in-situ observations of pressure to provide a dynamical constraint. This allows them to extend much further back in time than full-input reanalysis products. Because they do not ingest surface temperatures from meteorological observations over land they are formally and fully independent of land surface air temperature observations and any time averages derived from them.

20CRv2 (Compo et al., 2011) was the first sparse-input reanalysis provided publicly. It uses a version of the NOAA NCEP forecast system model modified to be able to run just utilising the lower-boundary conditions from SST fields and sea-ice and the surface pressure observations. 20CRv2 reaches back to 1871 and 20CRv2c subsequently extended this to 1851. Both versions have a relatively coarse 2° by 2° horizontal resolution generated using an Ensemble Kalman Filter (EnKF) algorithm. 20CRv2 was the first reanalysis product to create an ensemble of estimates consistent with the stipulated observational constraints. Both 20CRv2 and 20CRv2c produced a 56-member ensemble which was made available along with the ensemble mean fields.

The new NOAA-CIRES-DOE Twentieth Century Reanalysis version 3 (20CRv3) product is a substantial improvement upon 20CRv2c that benefits from an upgraded EnKF data assimilation algorithm, a new variational quality control algorithm, a new bias correction for marine observations before 1871, and an updated bias correction algorithm for all station data over land (Slivinski et al., 2019). It also benefits from a newer version of the NOAA NCEP model and improvements made to version 4.7 of the International Surface Pressure Databank (Cram et al., 2015). Its horizontal resolution is reduced to 0.7° from the 2° resolution of 20CRv2 and 20CRv2c (Slivinski et al., 2019). This serves to improve the assimilation of observations as well as representation of orography and land/sea margins.

20CRv3 specifically addressed some errors that came to light post the release of 20CRv2c including the biases reported by Ferguson and Villarini (2012) and misspecification of sea ice that produced a warm surface temperature bias in some regions. 20CRv3 includes an ensemble of 80 members which are also more dispersive thus providing for a better estimation of the true uncertainty in historical weather (Gil Compo, pers. comm.).

The ECMWF ERA-20C reanalysis, produced under the EU funded ERA-CLIM1 project, provides a deterministic estimate (single analysis with no uncertainty) on a 1° by 1° grid from 1900 to 2010 using a 4D-Var data assimilation approach with variational bias correction (Poli et al., 2016). It made use of the then current ECMWF IFS model version with some modifications to the specification of the model boundaries and forcing data. The observation constraint was comprised of atmospheric surface pressure observations from the International Surface Pressure Databank version 3.2.6 (Cram et al., 2015) and the International Comprehensive Ocean-Atmosphere data set (ICOADS) pressure and wind reports (Woodruff et al., 2011).

The ECMWF CERA-20C product, funded by the ERA-CLIM2 Project, is a coupled reanalysis product with a 1° by 1° resolution extending from 1900 to 2010 with a ten member ensemble (Laloyaux et al., 2018). It is the first time ECMWF made available an ensemble with a reanalysis product. CERA-20C has much in common with ERA-20C. It is built around the same coupled atmosphere-ocean model used in

ECMWF ensemble forecasts and utilises the same atmospheric pressure observations. Because CERA 20C is a coupled land-ocean reanalysis, a new assimilation system was developed that is a variation of the land and ocean assimilation system used previously in ERA-20 and ORAS4 (Ocean Reanalysis) to simultaneously ingest atmospheric and ocean observation using parameters common to both over a 24 hour period. However, atmospheric and oceanic observations are processed separately before combining the results of each to create an analysis field. The scarcity and relatively poor quality of observations in the early 20th Century, particularly in the Southern Hemisphere, increase the uncertainty of the ocean component in the earlier period but this improves significantly into the second half of the 20th Century with the deployment of free-floating buoys and other platforms.

It is the longevity of sparse-input reanalyses combined with their independence from station temperature observations which makes them a target for potential use in homogenisation in the present thesis. Several investigators have previously compared sparse-input reanalyses-to meteorological observations of land surface air temperature (Jones et al., 2012, Wang et al., 2018, Parker, 2016, Ferguson and Villarini, 2012). These prior studies collectively imply close correspondence, at least over certain regions and periods, but with potential caveats. Compo et al. (2013) argued that observed global warming is not an artefact of deficiencies in station temperature observation by independently inferring global average temperature changes from the sparse-input reanalysis and comparing to the then existing versions of several global datasets such as GHCNM and CRUTEM. Parker (2011) investigated the correspondence between CRUTEMv3 and 20CRv2 for the period of 1979 to 2008 and found that CRUTEMv3 warmed 0.05°C per decade more than 20CRv2 over this sub-period and that over the full period where the two datasets had corresponding estimates that there was a high degree of correspondence between the two datasets on a monthly basis. Ferguson and Villarini (2012) carried out a regional analysis over a test area of 15° by 15° in the central United States. They detected a shift in land surface air temperatures in 20CRv2, but could not find a corresponding shift in CRUTEMv3. They contended that the shift was most likely the result of an observation shock in the surface pressure observation constraint in and around 1940 to 1950 (Figure 2.17) when the number of observations increased significantly (Ferguson and Villarini, 2012).

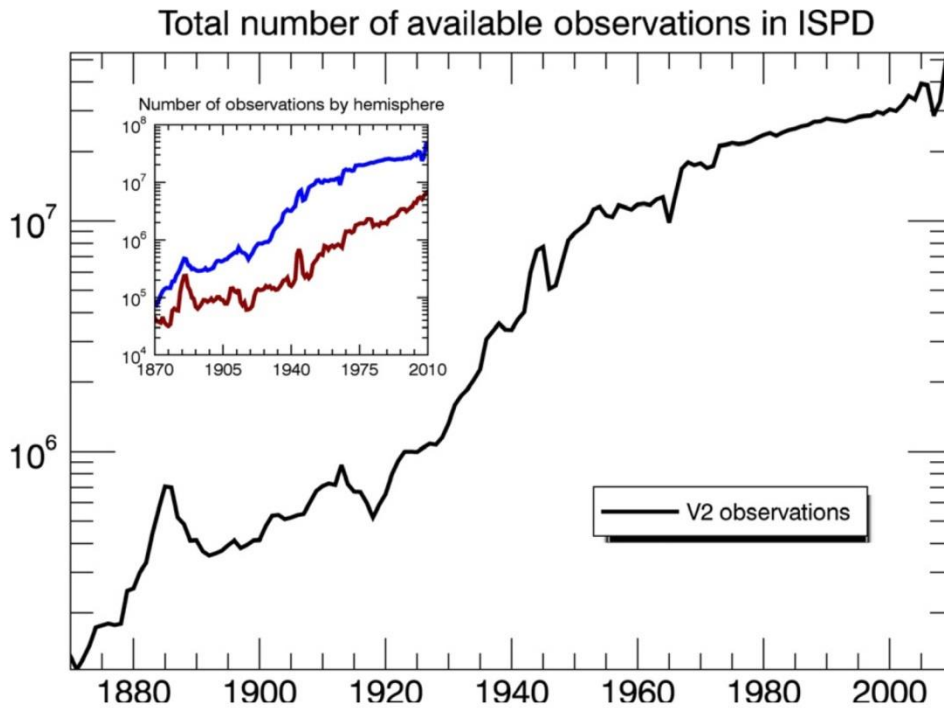


Fig 2.17 Time series of the number of pressure observations per year in version 2 of the International Surface Pressure Databank (ISPDv2) from 1870 to 2010. Note the logarithmic scale along the y-axis. Inset panel: time series during the same period showing the number of observations in the Northern Hemisphere (blue curve) and Southern Hemisphere (red curve), Caption & image taken from Cram et al. (2015)

As with traditional full-input reanalysis products, successive generations of sparse-input reanalysis products show improved quality as the community learn from previous efforts and as data assimilation techniques and model skill improve (Slivinski et al., 2019). Sparse-input reanalyses depend on the quality of the reconstruction of Sea Surface Temperatures as well as Sea Ice extent. These SST fields were carefully developed but may nevertheless contain remaining inhomogeneities, particularly near the ice edge before ~1950. However, it is reasonable to assume that these inhomogeneities are smaller than, and that any SST inhomogeneities are independent of, inhomogeneities occurring in land temperature records.

Sparse-input reanalysis products rely on surface pressure observations which come from the International Surface Pressure Databank (ISPD) of synoptic surface pressure observations. This databank extent goes back to 1768 although coverage degrades substantially before the mid-20th Century (Cram et al., 2015). Over 1900 to 2010 the number of surface pressure observations assimilated per month into the

ECMWF sparse-input reanalysis products increased from 30,000 to over 3 million (Poli et al., 2016). While this increase is overall beneficial it may introduce data shocks that could add time-varying biases into the reanalysis output (Ferguson and Villarini, 2012). The scarcity of surface pressure records for the tropics and the southern hemisphere before around 1935 limited the utility of NOAA's 20CRv2c before around the 1940s (Compo et al., 2006). Subsequent improvements in data availability have ameliorated this in 20CRv3 (Slivinski et al., 2019).

2.8 Radiosonde adjustments using full-input reanalyses

Radiosonde observations in common with most observation types are predominately used for weather forecasting and not for climate assessment. They contain many data issues which are exacerbated by their single-use whereby once a sonde is released it is very rarely recovered, along with frequent changes in instrumentation and lack of global uniformity in practices. Early efforts to homogenise radiosondes undertook a broad variety of approaches across a range of subsets of the total global network (Dai et al., 2011, Parker et al., 1997, Lanzante and Klein, 2003, Thorne et al., 2005b). These efforts were generally not automated and used some combination of time series analysis, available metadata and expert judgement to identify breakpoints and adjust the time series.

Haimberger (2006) introduced an automated homogenisation software package based on "innovations" - the difference between observations and the background forecast generated during the assimilation process in the preparation of the ECMWF's ERA-40 reanalysis - as a reference series interpolated to the radiosonde location. Breakpoints were identified in this innovations series using the SNHT test (Alexandersson and Moberg, 1996), but modified to overcome the particularities of radiosonde observations. The variant of the SNHT employed by RAOBCORE (Radiosonde Observation Correction Using Reanalysis) makes use of the innovations time series to estimate the probability of a break. This probability is combined with prior knowledge from metadata. RAOBCORE then goes on to derive an estimate of the adjustment from the innovation timeseries statistics.

This approach, however, is not fully independent, as the radiosonde observations were used as input data to the reanalysis. They were also used to bias correct early

satellite observations, and so the background field may itself contain inhomogeneities that are similar in nature to the radiosonde data issues and require adjustment. Thus in early versions of the algorithm, it was necessary to remove a spatio-temporally varying estimate of the bias artefacts prior to application of adjustments. After this modification to the innovation series, the breaks in an individual time series were adjusted.

To address concerns around independence/circularity, the “Radiosonde Innovation Composite Homogenisation” (RICH) approach was developed. RICH uses the break detection methodology of RAOBCORE, but for the adjustments either compares the target series with a reference series of observations made up from a weighted composite of neighbouring sites observations (RICH_{-obs}) or, alternatively, neighbouring station innovation statistics (RICH_{-tau}). For both approaches only apparently homogeneous segments of neighbours are utilised. Relative performance of these approaches depends upon circumstances including the density of neighbouring observations of their overall quality (Haimberger et al., 2012).

RAOBCORE and RICH have been widely utilised in scientific assessment activities and scientific analyses. Prior to 1979 radiosonde observations were the main source of observations above the surface. IPCC AR5 (Hartmann et al., 2013) relied heavily on RAOBCORE and RICH estimates. RAOBCORE and RICH have also contributed to almost a decade’s worth of iterations of the BAMS annual state of the climate report series global chapters (<https://www.ametsoc.org/index.cfm/ams/publications/bulletin-of-the-american-meteorological-society-bams/state-of-the-climate/>).

Versions of RAOBCORE and, more latterly, RICH have been used as input to several modern reanalysis products. JRA-55 (Kobayashi et al., 2015) used RAOBCORE version 1.4 (Haimberger et al., 2008) until the end of 2006 and then RAOBCORE version 1.5 (Haimberger et al., 2012) thereafter. MERRA1 (Rienecker et al., 2011) and MERRA 2 (Gelaro et al., 2017) also used RAOBCORE version 1.4 as input series until 2005. Dee et al. (2011b) used RAOBCORE_T_1.3 (Haimberger et al., 2008) for assimilation into ERA-Interim. Hersbach et al. (2020) employed the RICH adjusted data for assimilation into ERA-5 but defaulted to RAOBCORE estimates in regions with an absence of suitable neighbours.

RAOBCORE and RICH have also been widely used in scientific analyses. Lackner et al. (2011) investigated the use of Radio Occultation (RO) upper air data to determine if a climate change signal could be detected between 1985 and 2010 within the 50° N to 50° S region and concluded that a climate change signal is detectable using RO data within a 6-16 years period. They used RAOBCOREv1.4 and RICH and various reanalysis products in their determination. Liang et al. (2018) compared Arctic upper air temperatures observations derived by RO from the Constellation Observing System for Meteorology, Ionosphere and Climate and Formosa Satellite Mission 3 (COSMIC) between the 925 and 200 hPa pressure levels to radiosonde data homogenised using RAOBCORE and RICH and found a correlation of 0.96 or better. For the tropics, Mitchell et al. (2013) used RICH data to study the reported discrepancy between observations and coupled climate models in the tropical troposphere between 1979 -2008 and found that “within observational uncertainty, the 5–95 percentile range that temperature trends from coupled-ocean and atmosphere-only models are consistent with the analysed observations at all but the upper most tropospheric level (150 hPa). Ladstadter et al. (2011) assessed the difference in lower stratospheric temperature records from radiosonde and RO between 2001 and 2009 and found good agreement with RAOBCORE/RICH adjusted radiosonde estimates.

2.9 Summary

Historical observations of meteorology extend back to at least the 18th Century and quasi-globally from at some point in the mid-to-late 19th Century. Change has been ubiquitous in these records and their historical management has been highly fractured. The International Surface Temperature Initiative databank, currently containing more than 35,000 individual station records, is a global effort to assemble all available meteorological records into one databank with an emphasis on provenance. It provides new opportunities to re-examine land surface air temperature records. Currently, state of the art pairwise comparison techniques constitute the preferred method of homogenisation. The need for neighbour stations to act as a comparator raises issues that potentially become acute in data sparse regions and epochs. The available gridded surface temperature products constitute an ensemble

of opportunity, but the similarity in methods and stations used means that the true degrees of methodological independence is lower than implied. Critical in this regard is that all presently published methods rely on some form of neighbour-based homogenisation. Surface-only sparse-input reanalysis products constitute a formally independent estimate of the surface air temperatures. Following the successful application of conventional full-input reanalysis to homogenise radiosonde records it is the contention of the present thesis that a similar analysis may be possible for surface temperatures using the modern suite of 20th Century reanalysis products.

Chapter 3 Preparation of the ISTI databank holdings

3.1 Introduction

The International Surface Temperature Initiative's (ISTI) version 1.1.0 global land surface air temperature databank holdings consist of more than 35,000 individual station records. Many of these are of short-duration, but a small number of stations have records dating back to the 17th and 18th Centuries. The databank team undertook an effort to secure and store data arising from a broad range of sources, ranging from small collections of only a handful of stations from very specific regions to global collations of several thousand sites. The current release consists of a merge of monthly records from over 70 of these underlying data sources. Each source is ranked based on its provenance with more traceable sources given preference (Rennie et al., 2014, Rennie, 2015).

Many stations have been shared repeatedly such that they exist redundantly in many of the sources used to create the merged series. To confound matters yet further, data from distinct sources may differ in coordinate precision, station naming, data precision and rounding practices, the application of quality control and even in a small number of cases homogenisation. Furthermore, some individual sources may have themselves performed merges either of their own underlying sources and/or to create longer composite station series, which will have been invisible to the ISTI databank creators.

ISTI was far from the first attempt to form a global database. It was, however, perhaps the first attempt with truly global buy-in. The very first efforts pre-dated the availability of modern computers and were limited to at most hundreds of stations (Le Treut et al., 2007). With the advances in computational power in the 1980s, a renewed interest in the collection and curation of comprehensive collections of observations from across the globe emerged. This resulted in the construction of the databases that underlaid the CRUTEM1 and GHCNv1 datasets (Jones et al., 1982, Jones et al., 1985a, Jones et al., 1986, Vose et al., 1992, Hansen and Lebedeff, 1987).

Until the late 2000s, there was little interest in revisiting the early work given that the estimation of global averages at annual timescales requires only of the order 180

well-spaced series (Thorne et al., 2018). But advances in the daily data holdings (Menne et al., 2012), changes in data policies, increased use of neighbour-based homogenisation procedures (Menne and Williams, 2009) and criticism from those sceptical of climate science (Curry, 2011) led to renewed interest in the creation of improved holdings. Such improved holdings would enable greater analysis of changes and impacts regionally and locally which is key for impacts and adaptation decision making and the provision of climate services.

The ISTI concept consists of six stages of development commencing with stage Zero, which is the rescue of the old image and hard copies of records onto a digital database through to Stage five, the release of a family of fully homogenised products using distinct methods to sample the structural uncertainty (Figure 3.1), (Thorne et al., 2005). The ISTI databank public release is at stage 3 in this process and consists of a merged set of basic ('raw') data holdings prior to the application of any quality control or homogenisation procedures.

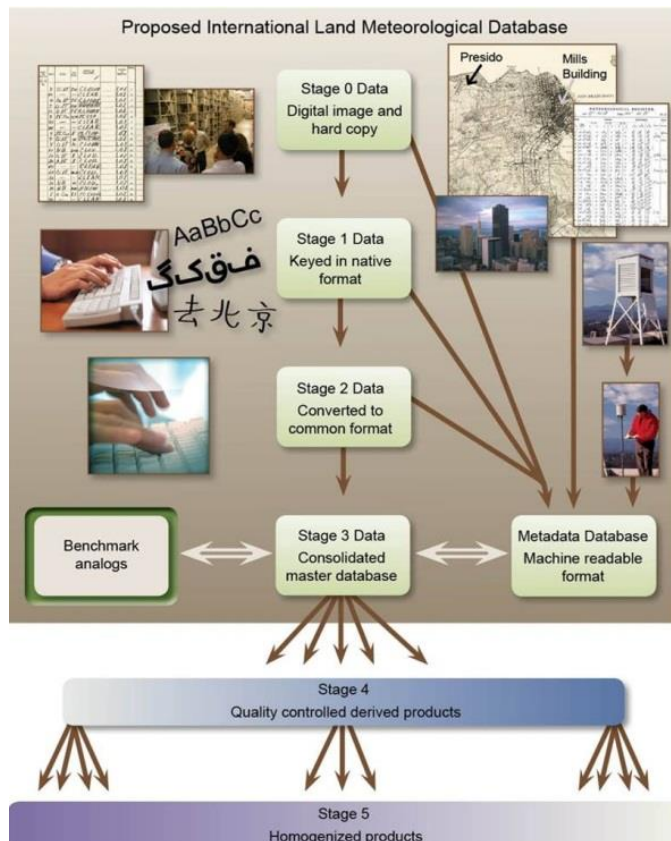


Figure 3.1. The proposed ISTI dataset construction process and stages of development taken from (Rennie et al., 2014)

The remainder of this chapter is structured as follows. In Section 3.2, a summary of the ISTI version 1.0.0 merging process and updates in version 1.1.0 are given. The removal of short period records from v1.1.0 which would be of limited utility to the present thesis is detailed in Section 3.3. Section 3.4 discusses the selection of 29 case study stations and their neighbours for an initial assessment of results both here and in subsequent chapters. Section 3.5 summarises the investigation of various duplication issues found within the ISTI databank and their resolution. Section 3.6 undertakes a discussion on the chapter findings and Section 3.7 concludes.

3.2 Summary of ISTI version 1.0.0 merging process and updates in version 1.1.0

Full details of the ISTI databank construction are given in Rennie et al. (2014) and in associated documentation referenced therein. This section provides solely the necessary details to understand in an appropriate context the substantive post-processing analysis undertaken herein and to support the interpretation of the analysis undertaken in Chapters 4 and 5. A reader interested in more detail is referred to (Rennie et al., 2014) for further particulars of the approach.

The ISTI databank is a merge of numerous underlying data sources. Before merging was performed, the sources were organised into a hierarchy whereby those sources with greater data provenance, extensive metadata and originating from national holdings took higher priority over lesser documented / more integrative sources. Daily-derived sources, from which monthly averages could be calculated in a consistent and traceable manner were given priority over monthly sources. The GHCND holdings (Durre et al., 2010, Menne et al., 2012) were given the highest priority, thus ensuring consistency between GHCND and the ISTI databank for stations where both exist. The magnitude of the task of merging several hundred thousand candidate series from 70 underlying sources necessitated an automated approach to merge decisions.

Firstly stations were checked against common data and metadata quality issues by applying a set of criteria checks to all stations which included:-

1. A decade-by-decade check on the variance to check for issues. If decade-to-decade variations are too large the station is blacklisted.
2. Questionable correlations with nearby neighbours (if available).
3. Geolocation coordinates that place the station over water according to a high resolution digital elevation map

Any issues found were rectified if possible by e.g. ascertaining a land-based location via coordinate correction, and if so the station proceeds to the full metadata comparison check. If that was not possible the station was blacklisted.

Once prechecking was cleared, the process moved onto a check for metadata similarity. This involves a comparison of the candidate station with each station already present in the databank at that stage of integration of sources and is made up of four basic tests:-

1. Based on latitude and longitude the distance between the stations is determined using great arc distances and fitted to an exponential decay function that decays to zero at 100km distance. The metric ranges between zero (no match) and 1 (perfect match).
2. Altitude is treated similarly to distance with the exponential function decaying to zero at 500 meters separation. Not all stations contain altitude data and if absent this test is not applied.
3. The third test is to check if data recording started in and around the same time with an exponential decay function applied if the start dates are within 10 years of each other.
4. The final metadata check is to determine if the station names are similar in any way (similar names different spelling). For this purpose, the Jaccard Index (JI) is applied. This is based upon an alphabetical intersection of the two sample names (target & candidate names) divided by their union. The weakness in the JI is that it looks to see if the same letters exist in both names and disregards their position and deviations in the same letter combination will produce a JI of 1 (e.g. Tokyo and Kyoto).

Based on a numerical combination of the above, metadata similarity (metadata_metric) was calculated using a weighting for each component as follows:

$$\text{metadata_metric} = \frac{(9 * \text{Dist}) + (1 * \text{Height}) + (2.5 * \text{year}_{T_{\max}}) + (2.5 * \text{year}_{T_{\min}}) + (5 * \text{JI})}{20} \quad \text{Eqn 3.1}$$

Where: Dist is the prior probability based upon distance; Height is the prior probability based upon altitude; year_{T_{max}} and year_{T_{min}} are probabilities based upon start year intersections and JI is the Jaccard Index. All terms on the RHS of Eqn. 3.1 are bounded between 0 and 1 such that the resulting metadata_metric is also bounded between 0 and 1. The weights to given terms derive from expert judgement as to the relative import of each aspect in determining the probability of a match.

This metadata analysis gave rise to three potential decisions as to how to proceed further:

1. Include a source as a new series if the metadata is sufficiently dissimilar from all other stations already present in the merged holdings.
2. Carry forward to consideration for a merger with one or more existing series that match the metadata sufficiently.
3. To withhold it from further consideration owing to potential metadata issues.

Specifically for the second option above, all candidate/target stations that exceed 0.5 for metadata metric proceeded to the full data comparison test if there are at least 60 months overlapping data. For the data comparison an Index of Agreement (IA) was calculated based on the following formula by (Willmott 1981):

$$IA = 1.0 - \frac{\sum_{i=1}^n |T_i - C_i|}{\sum_{i=1}^n |C_i - \bar{T}| + |T_i - \bar{T}|} \quad \text{Eqn 3.2}$$

Where IA is the Index of Agreement, T_i and C_i are monthly values for the target and candidate station respectfully and \bar{T} is the mean of the target station. The IA were calculated for the T_{max} and T_{min} separately and values are bounded between 0 to 1

by construction. Because the duration of overlap can distort the calculated IA value with longer overlapping periods biasing upward the IA, a lookup table for H1 (station match) and H2 (station uniqueness) was generated and a cumulative distribution function calculated based upon the overlap period for both station uniqueness and station sameness. The longer the overlap between the target and the candidate the nearer to one the IA score must achieve to be considered a match.

If no or insufficient data overlap occurred and no data comparison was made, the final decision to merge was based solely on the metadata metric. In this case, the qualifying metadata metric score increases from 0.5 to 0.9. If the highest candidate receives a score greater than 0.9 then the candidate is merged with the target. Otherwise, it is withheld.

For stations with sufficient data overlap, once the H1 and H2 values were calculated the posterior of similarity and uniqueness were calculated as follows:

$$\text{Posterior of similarity} = \frac{\text{Metadata metric} * H1_{t_{\max}} * H1_{t_{\min}}}{3} \quad \text{Eqn 3.3}$$

$$\text{Posterior of uniqueness} = (1 - \text{Metadata metric}) + H2_{T_{\max}} + H2_{T_{\min}} \quad \text{Eqn 3.4}$$

If any station reached a posterior probability of similarity of 0.5 or greater then the candidate station was merged with the target station with the highest such value. If none of the stations exceeded 0.5 for this metric, but one of the posterior of uniqueness exceeded the threshold of 1.3 then that candidate station was assessed to be unique and was added to the target dataset. If stations meet neither criteria they were withheld.

When merging is deemed appropriate, data from the candidate station (arising from the present source deck) was only merged with the target station (arising from the present merged holdings) at timestamps in the target station that have missing data such that higher priority source decks provide the data by preference. In version 1.0.0 the data to be merged referred to as the “Gap Threshold” must be a string of at

least 60 months (5 years) long. This assures that higher priority sources contribute the most possible data to the merged station series following completion of the process with lower priority sources being used to backfill data in version 1.0.0. It also serves to minimise the possibility of adding a break by inadvertently infilling with data from a different station.

The above algorithm only considered the Tmax and Tmin. Once that process was completed a Tavg series was calculated by averaging the Tmax and Tmin in all cases. The process was then repeated for those sources containing solely Tavg with some minor modifications to the above formulas as shown in equation 3.5 and 3.6:

$$\text{Posterior of similarity} = \frac{P_{\text{metadata}} * H1_{Tavg}}{2} \quad \text{Eqn 3.5}$$

$$\text{Posterior of uniqueness} = (1 - P_{\text{metadata}}) + H2_{Tavg} \quad \text{Eqn 3.6}$$

A full detailed narrative of the above summary may be found in (Rennie et al., 2014) - and updates at (Rennie, 2015).

3.2.1 Updates to v1.1.0

An update to the original process was released as version 1.1.0 in 2015. Significant changes were made in the treatment of several sources as follows:

1. Incorporating updates to the primary GHCND source that included the addition of 1,400 more stations.
2. One underlying source, “Russsource” was found to be a compilation of twenty-seven individual sources of widely differing provenance and quality and a decision was made to separate this source into its individual components.
3. The removal of CRUTEM4 as a source because its inclusion was producing an unacceptable frequency of false unique station identifications owing to its use of highly processed and homogenised series.

Further details are given in a NOAA technical note (Rennie, 2015).

The merge algorithm methodology did not undergo any changes between the two versions but two threshold values were modified to maximise the amount of data in the version 1.1.0 release (Table 3.1).

Name	Description	Version 1.0.0 Threshold Value	Version 1.1.0 Threshold Value
Metadata metric threshold	This metric takes into account, distance, height difference, and the Jaccard Index metric	0.50	0.75
Gap Threshold	Gap period in months that must exist when merging a candidate station with a target station	60	12

Table 3.1 Changes in thresholds between ISTI release v1.0.0 and release v1.1.0 (Rennie, 2015)

Release V1.1.0 used in this thesis has over 80% global land area coverage when aggregated to 5° by 5° grid boxes since 1960 (Figure 3.2). This significantly reduces back in time into the 19th Century when available records are concentrated predominantly in the United States of America and Western Europe (Figure 3.3). It is clear, though, that many records potentially exist in paper or image form which could improve the situation in future (Allan et al., 2011, Brönnimann et al., 2019). Efforts are ongoing via work by NOAA NCEI and the Copernicus Climate Change Service to improve the stewardship of data holdings (Thorne et al., 2018), and a broad variety of efforts are underway to rescue historical data holdings. So, there is considerable hope for further improvements (see Chapter 2 for further discussion).

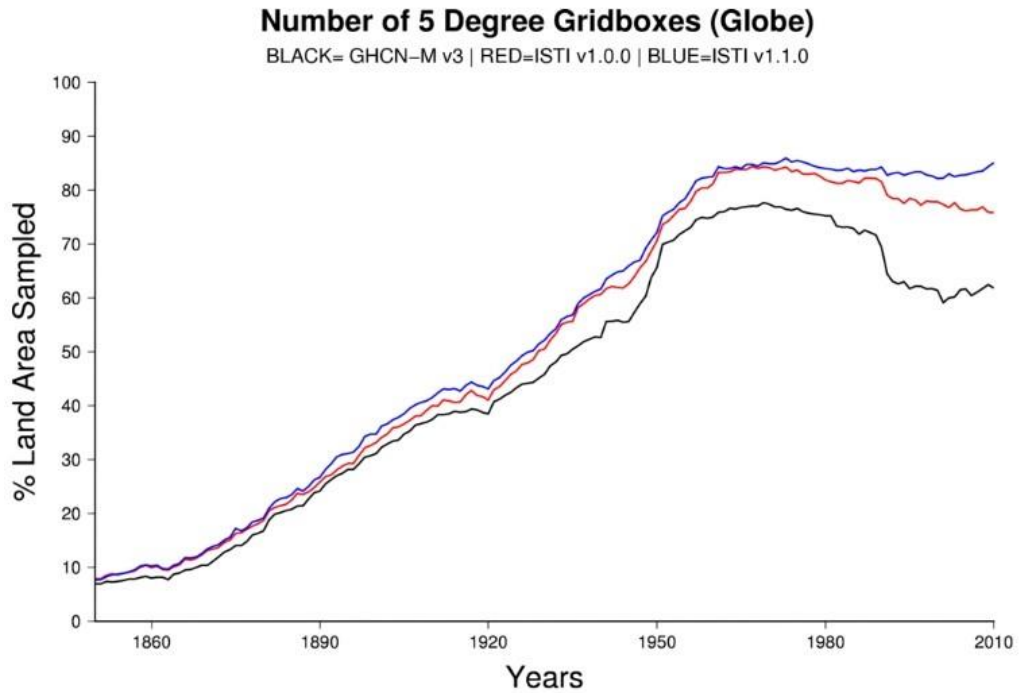


Figure 3.2. Timeseries of percentage of 5° by 5° grid boxes that contain land which has at least one station present plotted against year. The black curve represents coverage in the precursor GHCNv3 product. The red curve is ISTI v1.0.0 and the blue curve is ISTI v1.1.0. Taken from Rennie (2015).

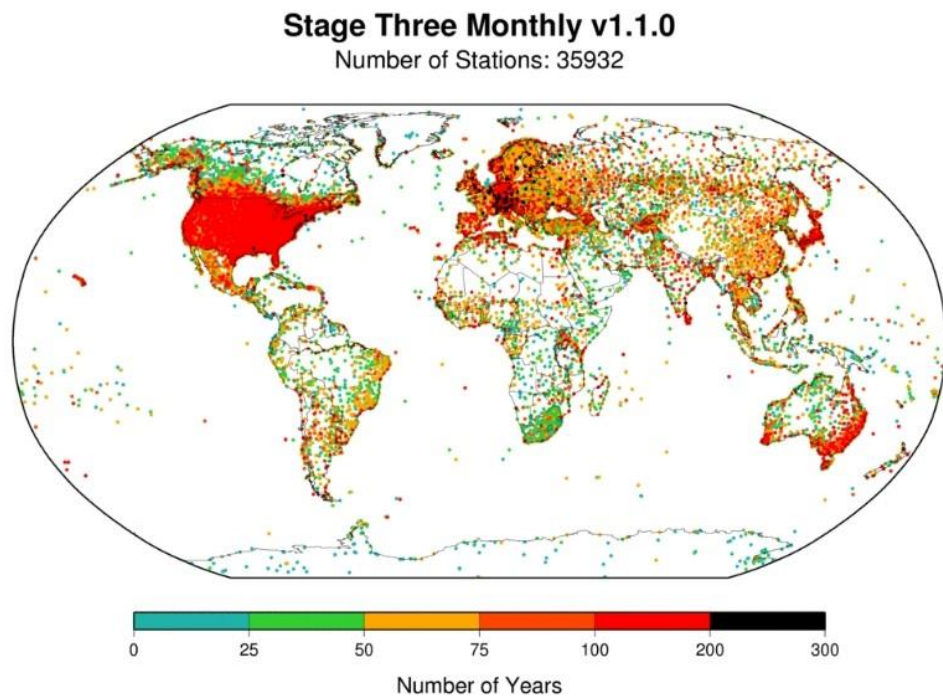


Figure 3.3 Global distribution and period of record (colours) of stage 3 monthly stations. Note the concentration of long record stations (reds and blacks) in N.America and Western Europe. Long records overplot shorter period records. Taken from Rennie (2015).

3.3 Removal of short period records from V1.1.0

As per WMO Guidelines on homogenisation, as a first step, those stations with fewer than 120 months of observations were removed (Bessemoulin et al., 2018). The total number of stations removed for this reason was 7,491. Most of these stations are located in North America, predominantly the USA, although a significant number exist in Europe, particularly in Scandinavia (Figure 3.4). These regions are well represented within the remaining 28,428 stations (c.f. Figure 3.3), and therefore their removal should not have a substantive impact. Most of these removed stations commence post 1940 (Table 3.2). There are, however, among these removals 495 station records commencing before 1900, twenty five of which commence pre-1880 (Figure 3.5). Brönnimann et al. (2019) have created a catalogue of pre-1850 known records that suggests that the issue of short segment records will become ubiquitous in records in this early period. This is prior to the instigation of national meteorological services and programs of sustained monitoring, and thus records tended to be tied to individual amateur observers and accordingly of relatively short duration.

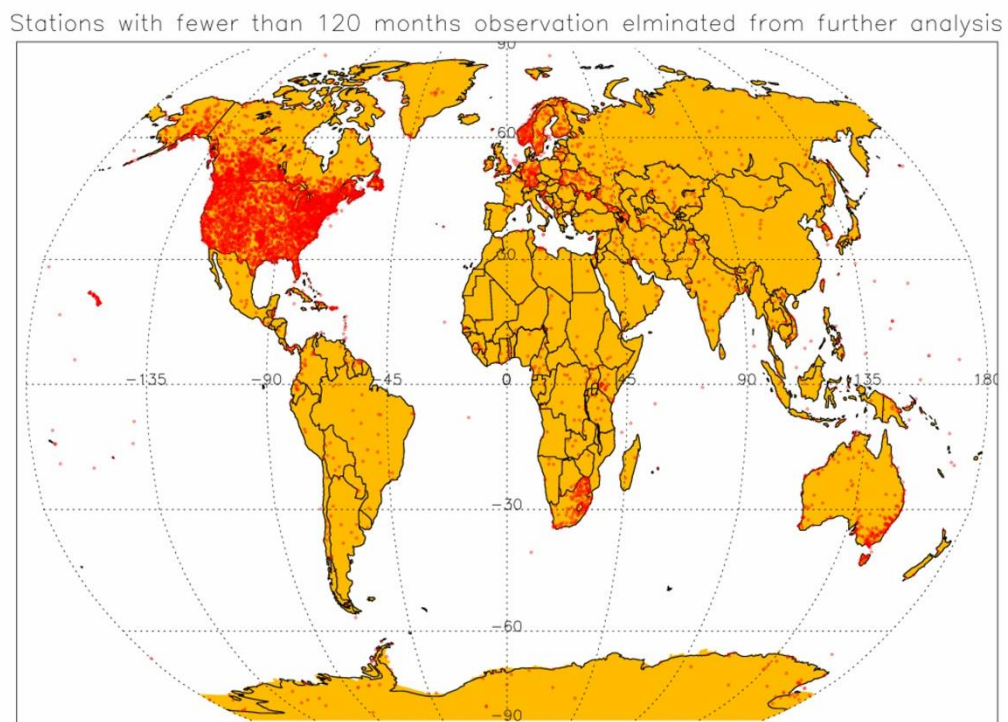


Figure 3.4. Map of all removed stations (red dots) from V1.1.0 based upon removing all records with fewer than 120 monthly average obs. Most removed stations are from North America.

Pre 1900 stations with fewer than 120 observations

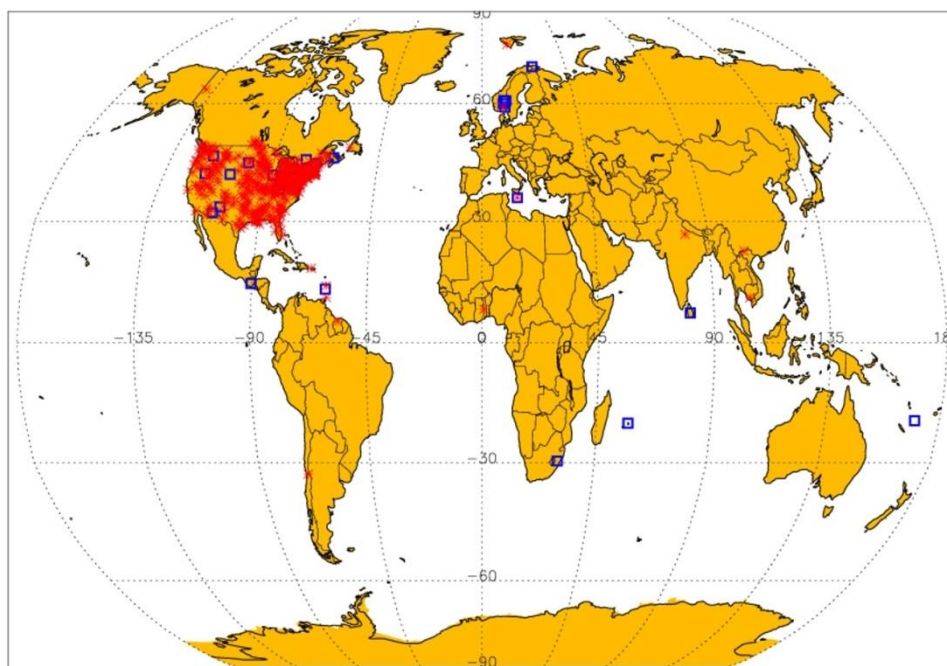


Figure 3.5 The locations of 495 stations with fewer than 120 month observations in the ISTI, release v1.1.0 that commenced observations between 1880 and 1900 are shown in red. The 25 stations that commenced observations before 1880 are shown in the blue squares.

Period	Number of Stations removed	Total monthly observations removed	Minimum Observations removed	Mean Observations removed
Pre 1880	25	1899	34	76
1881-1900	503	32665	24	65
1901-1920	636	39813	24	63
1921-1940	325	22622	24	70
1941-1960	1243	82640	24	66
1961-1980	1693	110480	24	65
Post 1980	3066	217138	24	71
Tot. Summary	7491	507257	24	68

Table 3.2. Summary of the stations with fewer than 120 monthly averages that were removed from further analysis by period, the total removed monthly averages, the minimum station length removed and the mean station length removed (the maximum in all cases is the 119 limit applied). The ISTI databank does not contain any stations with fewer than 24 monthly values.

3.4 Selection of case study stations and their neighbours for initial assessment

To develop and test the approaches undertaken within this thesis, it was necessary to select a small number of case-study stations from climatologically distinct regions that could be considered in some detail. Twenty nine stations were selected from the set of stations remaining after removal of short-period of record locations. Their selection attempted to ensure a representative sampling using sites from urban locations, rural locations, desert, forest/jungle, coastal, crop/grassland, densely and sparsely sampled regions. Steps were taken to also ensure that tropical, mid-latitude and near-polar regions were approximately equally represented in the selection. The case study stations and a subset of key characteristics are summarised in Table 3.3. Their locations are shown in Figure 3.6. Such a finite subset of locations provides a snapshot of the possible issues sufficient to understand potential challenges, but it should be caveated cannot plausibly catch all issues present in the entire holdings.

Station Number	Station name	Country	No. Obs	Start	End	Lat	Lon	Alt	Regions	Environment
AR000087828	TRELEW_AERO	Argentina	1367	01/1901	12/2014	-43.2	-65.266	43	Mid Lat	Desert
ASXLT209646	HOBARTTASMANWAS_949700	Australia	1732	01/1851	10/2013	-42.9	147.5	54	Mid Lat	Complex
ASXLT263670	PERTHAUSWAS_946080	Australia	1028	01/1917	09/2013	-32	115.9	60	S. Tropic	Complex
AYM00089314	THERESA	Antarctic	201	12/1994	12/2015	-84.6	-115.817	1463	Polar	Ice
AYXLT563342	ERIN	Antarctic	146	10/1996	12/2013	-84.6	-128.82	1006	Polar	Ice
CA003031400	CARWAY	Canada	1025	09/1914	10/2011	49	-113.383	1354	Mid Lat	Crop/Grass
CHM00058362	SHANGHAI	China	1703	01/1873	12/2014	31.4	121.467	4	S. Tropic	Complex
CI000085469	ISLA_DE_PASCUA	Chile(Easter Is)	862	12/1941	12/2014	-27.167	-109.433	69	S. Tropic	Crop/Grass
CIXLT967829	SANTIAGOWAS_855770	Chile	1765	01/1861	09/2013	-33.5	-70.7	520	Mid.Lat	Complex
FIE00142226	HELSINKI_KUMPULA	Finland	1961	01/1851	12/2014	60.2028	24.9642	24	Polar	Complex
FI000091652	UDU_POINT_AWS	Fiji	664	01/1951	12/2014	-16.133	-179.983	63	Tropic	Costal
GMM00010628	GEISENHEIM	Germany	1566	07/1884	12/2014	49.983	7.95	123	Mid Lat	Crop/Grass
INM00043057	BOMBAY_COLABA	India	1628	01/1851	12/2014	18.9	72.8167	11	Tropic	Costal
ITE00115588	PADOVA	Italy	1888	01/1851	09/2010	45.3983	11.8803	12	Mid Lat	Complex
JA000047817	NAGASAKI	Japan	1681	01/1851	12/2014	32.732	129.866	35	S. Tropic	Complex
LH000026730	VILNIUS	Lithuania	1935	01/1851	12/2014	54.6331	25.1	156	Mid Lat	Crop/Grass
MT000016597	LUQA	Malta	1862	11/1865	12/2014	35.85	14.4831	91	Mid Lat	Complex
MZXL405557	LOURENCO_MARQUES	Mozambique	1031	01/1865	12/2014	-26	32.6	64	S. Tropic	Costal
NLM00006260	DE_BILT_1	Netherlands	1968	01/1858	12/2014	52.1014	6.1867	2	Mid Lat	Costal
NOE00134898	TROMSOLANGNES	Norway	1908	01/1856	12/2014	69.6767	18.9131	8	Polar	Costal
PKXLT983863	QUETTASHEIKH_MANDA	Pakistan	1138	01/1887	12/1970	30.18	66.95	1803	S. Tropic	Desert
RSM00023662	TOLKA	Russia	800	05/1947	12/2014	63.98	82.08	31	Mid Lat	Forest/Jungle
RSM00028722	UFA	Russia	1486	01/1891	12/2014	54.7167	55.8831	104	Mid let	Forest/Jungle
SPE00120143	HUELVA_RONDA_DEL_ESTE	Spain	1335	01/1903	11/2014	37.28	-6.9	19	Mid Lat	Complex
SWE00136129	VARTAN	Sweden	1950	01/1851	12/2014	59.35	18.1	20	Mid Lat	Complex
TZXLT095229	DAR_ES_SALAAM_TANZANIA_E	Tanzania	803	01/1851	09/2013	-6.5	39.29999	58	Tropic	Costal
USC00300047	ALBANY	USA	1826	01/1862	12/2014	42.6461	-73.7472	13	Mid Lat	Complex
USC00500252	AMCHITKA	USA	136	02/1843	10/1992	51.3833	179.2833	69	Mid Lat	Complex
ZI000067975	MASVINGO	Zimbabwe	1019	01/1924	12/2014	-20.067	30.867	1095	Tropic	Crop/Grass

Table 3.3 A summary of the initial subset of 29 case study stations used herein including their identifier, name, country, geolocation, regional characteristics and local environment. For this analysis observations between January 1851 and December 2014 are used. Station records may start and end outside the selected dates

Once the 29 pilot candidate stations were selected, it was necessary to select neighbouring stations to compare the potential of sparse-input reanalysis products for

homogenisation against more established methods. There are several different ways of selecting neighbouring stations to act or contribute to a reference series for pairwise homogenisation in the literature (Wang et al., 2018, Mamara et al., 2012, Menne and Williams, 2009). These methods often select stations based upon correlation, spatial representativeness, or both. For simplicity, we have selected the 25 nearest neighbours to compare relative performance. This is obviously less optimised than many state-of-the-art techniques. On the flip side, it mitigates against the selection of a neighbour station-set from the available pool of such series that may inadvertently contain data issues of similar structure to those present in the candidate station. There is an argument that such optimised search criteria may, in certain circumstances, be inadvisable for this reason. The locations of this nearest-neighbours selected using this simple approach are shown in Figure 3.6.

For completeness, we also explored the selection of twenty five neighbours with at least a 50% data overlap (i.e. we removed from the neighbour set any stations with less overlap and continued to expand the search radius accordingly until 25 stations matching these criteria were identified). Unsurprisingly, this selection method came at a cost of decreasing the correlation and increasing the standard deviation of the difference series as the selected neighbours are, on average, more distant. This set of alternative neighbours can be seen in the lower panel of Figure 3.6 and the comparison of both methods is in Table 3.4.

In well sampled regions, except for some very long running stations, the resulting increase in distance to the selected neighbours is not significant and the cost to correlation and standard deviation of differences is marginal. In the less well sampled regions with increased separation, this selection results in decreased correlation and an elevated standard deviation of the difference series.

Station	25 Nearest Neighbours with at least 50% of time series overlap			25 Nearest Neighbours by distance			Difference in sigma	Difference in r	Difference in distance (KM)
	Median Neighbour	Median sigma	Median r	Median Neighbour	Median sigma	Median r			
AR000087828	768	1.178	0.506	496	1.103	0.651	0.075	-0.145	272
ASXL209646	538	0.824	0.615	49	0.419	0.881	0.405	-0.266	489
ASXL263670	105	0.567	0.864	26	0.520	0.889	0.047	-0.025	79
AYM00089314	1285	2.616	0.587	1276	2.505	0.596	0.112	-0.009	9
AYXL563342	1149	2.383	0.668	1145	2.225	0.668	0.158	0.000	4
CA003031400	91	1.043	0.936	35	1.019	0.933	0.025	-0.003	56
CHM00058362	633	1.041	0.674	264	0.638	0.869	0.403	-0.195	369
CI000085469	3645	1.415	0.013	3534	1.520	0.037	-0.105	-0.024	111
CIXL967829	759	1.329	0.396	245	1.148	0.448	0.181	-0.052	514
FIE00142226	399	1.116	0.894	43	0.599	0.973	0.518	-0.079	356
FJ000091652	751	0.663	0.663	351	0.644	0.484	0.019	0.178	400
GMM00010628	75	0.459	0.969	44	0.464	0.966	-0.006	0.003	31
INM00043057	826	0.943	0.428	434	0.850	0.536	0.093	-0.108	392
ITE00115588	125	0.782	0.870	85	0.793	0.861	-0.011	0.009	40
JA000047817	169	0.554	0.897	89	0.420	0.938	0.134	-0.041	80
LH000026730	406	1.060	0.901	167	0.730	0.951	0.330	-0.050	239
MT000016597	823	1.223	0.584	272	0.814	0.800	0.409	-0.217	551
MZXL405557	729	1.486	0.275	197	1.877	0.225	-0.392	0.050	532
NLM00006260	222	0.821	0.900	61	0.508	0.956	0.313	-0.055	161
NOE00134898	533	1.660	0.752	71	1.185	0.794	0.475	-0.042	462
PKXL983863	858	1.634	0.438	273	1.518	0.598	0.116	-0.160	585
RSM00023662	477	1.768	0.883	439	1.707	0.892	0.061	-0.009	38
RSM00028722	351	1.124	0.916	199	0.939	0.941	0.185	-0.026	152
SPE00120143	250	0.804	0.824	131	0.626	0.878	0.178	-0.054	119
TZXL095229	734	1.214	0.163	277	0.902	0.258	0.312	-0.095	457
USC00500252	1325	1.796	0.530	1170	2.088	0.613	-0.292	-0.083	155
ZI000067975	653	1.063	1.063	265	0.870	0.694	0.192	0.369	388

Table 3.4 Comparison of standard deviation and correlation (based upon the selection of neighbours based on two criteria (1) by at least 50% time series overlap and then (2) by 25 nearest by distance. The former will expand the search region and include, on average, stations further away. Differences in sigma, correlation and distance between the two selection methods are shown in the final three columns.

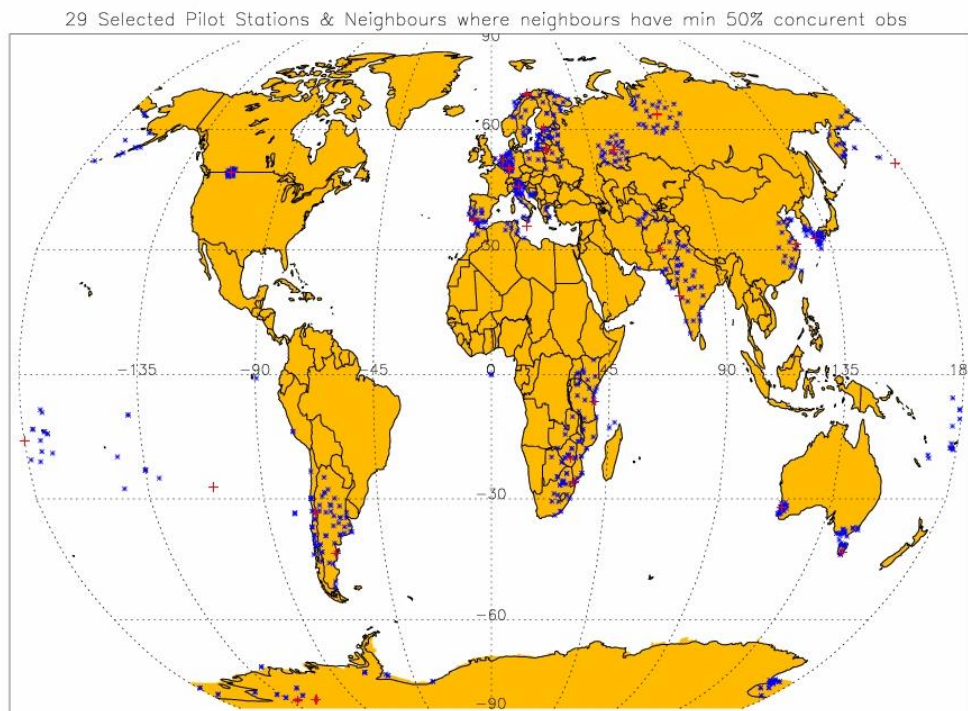
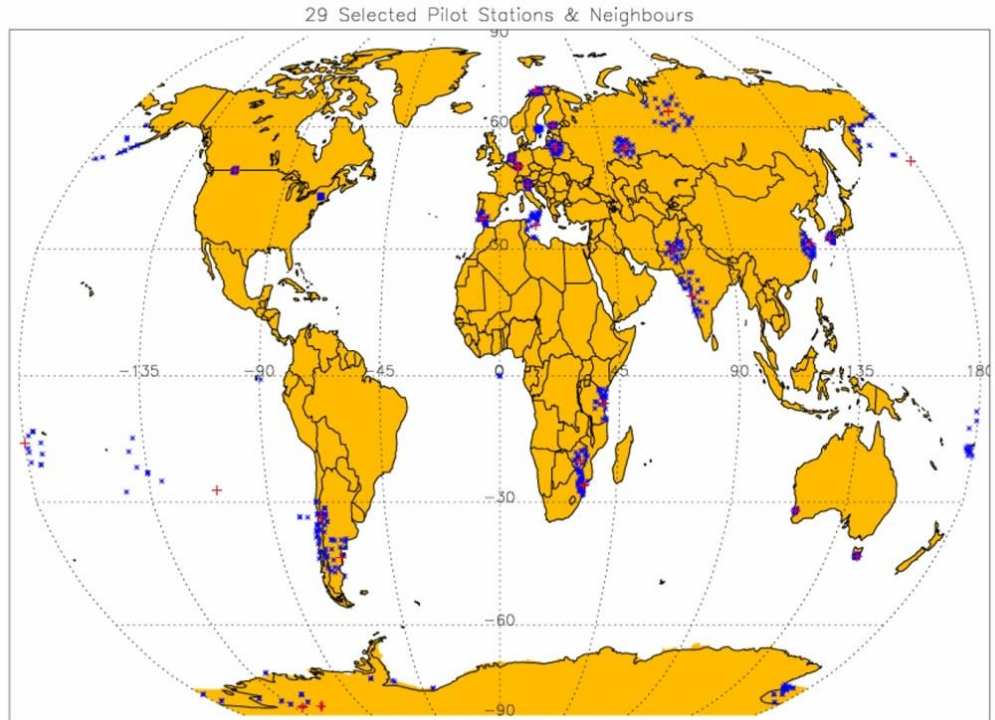


Figure 3.6 The 29 case study stations (marked with red crosses) and their 25 neighbours (blue asterisks) (top panel) and (lower panel) with a stipulation of 50% data record overlap being required. Spread in neighbour locations with a 50% overlap requirement is particularly marked for case study stations outside Europe and North America. Some case study stations share neighbours.

A further examination of the 29 case study stations was then carried out by plotting the difference series between the case study series and each of their 25 nearest neighbours. An example is shown in Figure 3.7 for station Udu Point in the Fiji Islands. Most of the 29

stations had series that looked very similar to this example series and showed differences that looked akin to the expected combination of random and systematic differences that would be expected for nearby stations that, on monthly timescales, should be highly correlated.

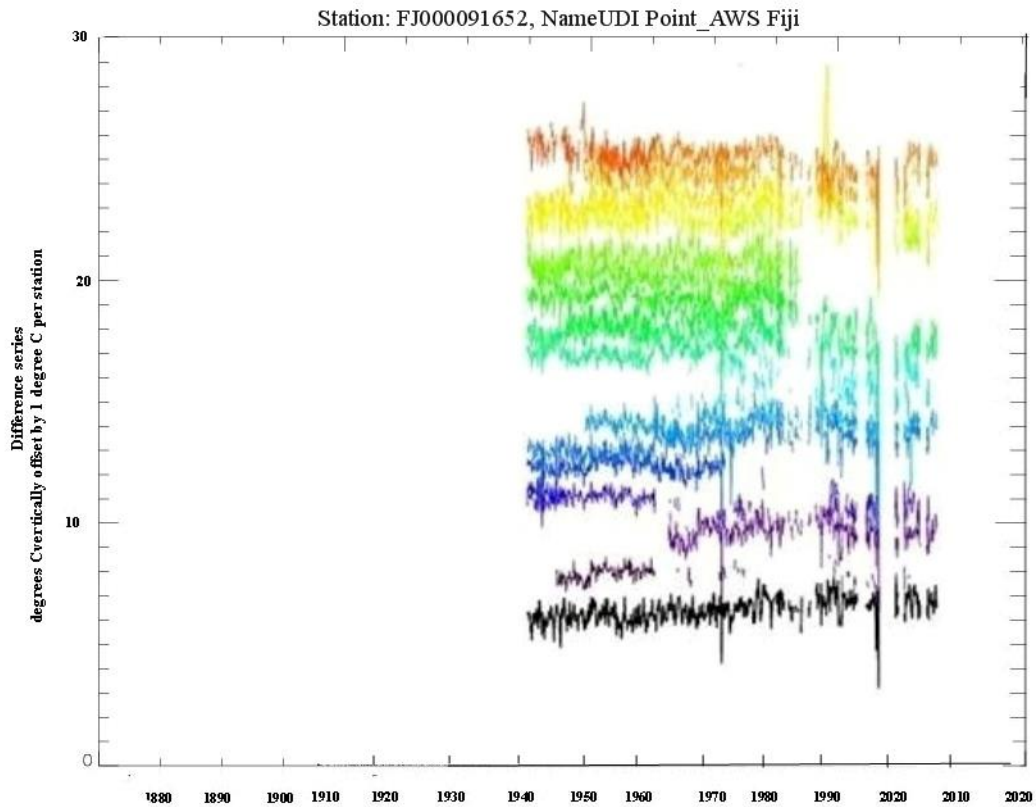


Figure 3.7. The differences in anomalies between the candidate station and each of its nearest 25 neighbours. Each difference is offset from by 1°C vertical intervals for clarity. If the candidate and neighbour station series do not intersect there is no candidate–neighbour pairs difference shown.

However, upon examining the station Helsinki_Kumpula Finland (FI00142226) (Figure 3.8), it immediately became obvious that for this station there were significant issues of time series ghosting whereby a single long-series had been propagated into many geographically proximal records. An alarmingly large number of the neighbour difference series have a propensity to repeated strings of zeroes or repeating annual cycles. It is clear that in this case several variants of one true long-term series have been inadvertently incorporated into the databank release multiple times. Firstly there are several direct duplicates which yield strings of zero differences. Secondly, there are a number of adjusted duplicates whereby the differences repeat annually in at least two distinct manners, indicative of the presence of at least two homogenised versions of the same underlying series. The

question arose as to whether this was a one-off occurrence or whether this phenomenon was present throughout the ISTI databank. The 29 case study stations were firstly considered in more detail for more instances of this phenomenon. While Helsinki Kumpula was by far the most severe case in this very small sample of the ISTI databank, it was not the only occurrence. Single duplication events were also discovered in four further case study stations (Table 3.5). An example of these findings is shown in Figure 3.9. These findings suggested that an analysis of the entire databank holdings would be required to determine how serious this problem was.

Station	Lat	Lon	Location
FIE00142226	60.2028	24.9642	Helsinki Kumpula Finland
TZXLT095229	-6.5	39.3	Dar es Salaam Tanzania
NOE00134898	69.6767	8.9131	Tromsø Norway
JA000047817	32.732	129.866	Nagasaki Japan
ASXLT2633670	-32	115.9	Perth Australia

Table 3.5 Summary of initial case study series found to suffer from exact series replication between the case study station and one or more of the 25 nearest ISTI databank neighbours.

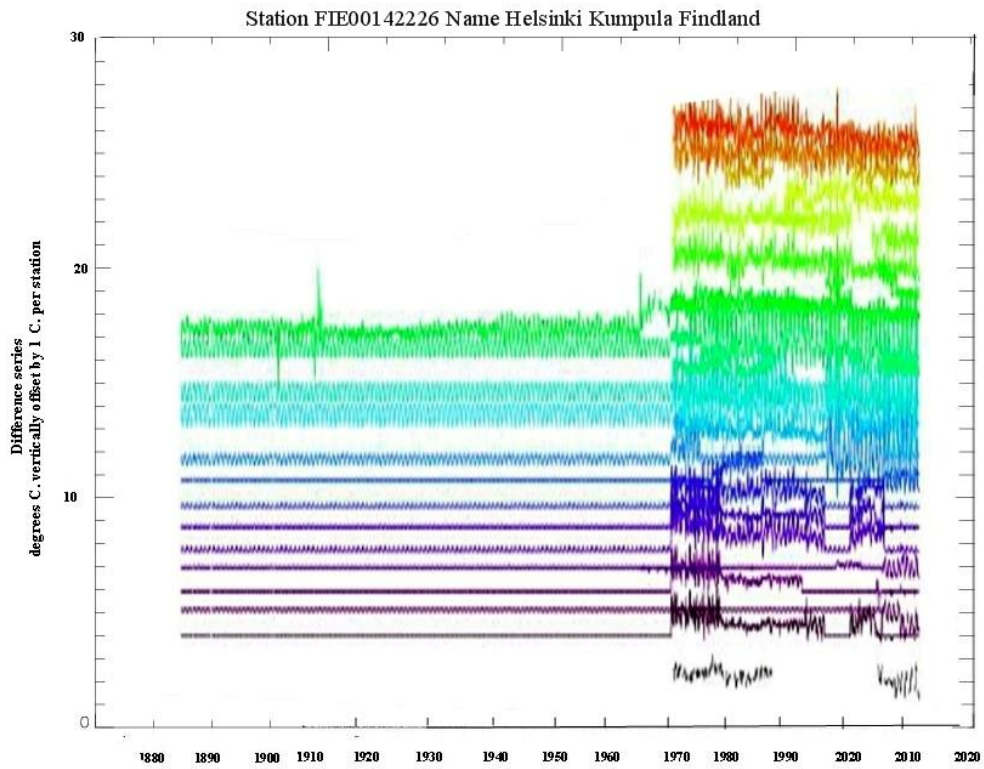


Figure 3.8 Plot of Helsinki Kumpula difference series to its 25 nearest neighbours (coloured and each successively offset vertically by 1 degree for clarity). Note the strong variations in behaviour particularly marked prior to 1970.

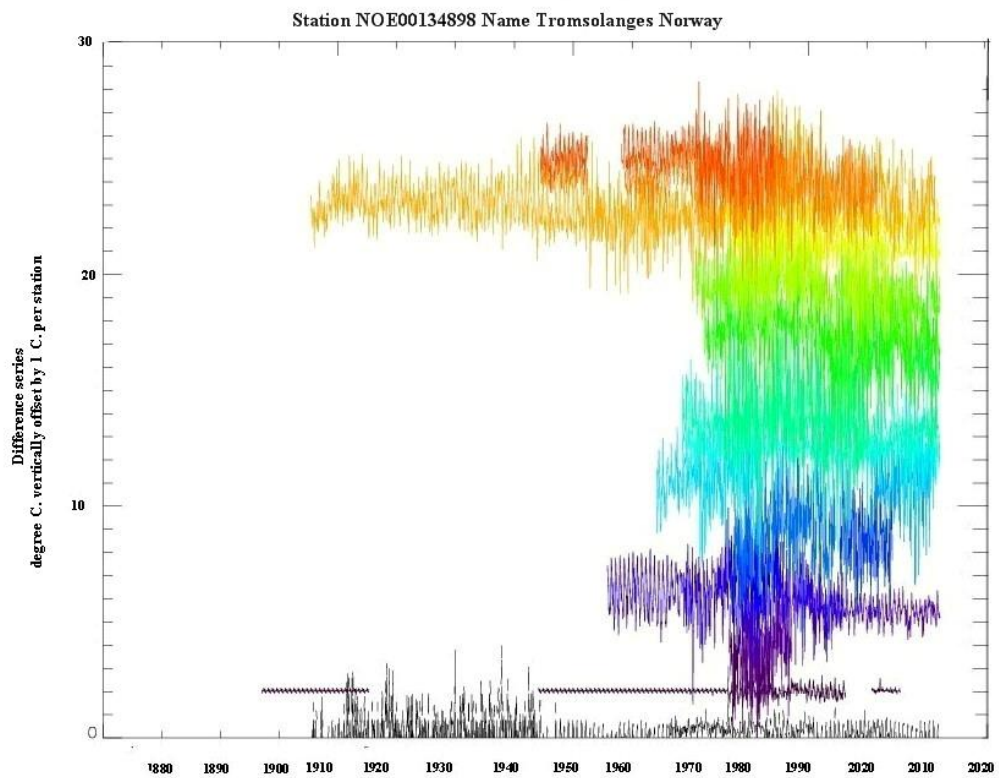


Figure 3.9 Station at Tromsølanges Norway which shows a single duplication event with annually repeating offsets indicative of a homogenised series.

3.5 Investigation of duplication of station data throughout the ISTI databank

Following the discovery of the potential for duplication of long-term series within the station records Jared Rennie of CICS-NC was contacted. He advised that when NCEI was constructing GHCNMv4 they identified a degree of duplication in the ISTI databank and formed a blacklist of such stations (Rennie, pers. comm) which was subsequently shared. The NCEI blacklist was applied. However, it was noted that their blacklisting did not address all the issues discovered in the 29 case study station series and in particular did not contain any neighbours of station FIE000142226, Helsinki Kumpula, Finland, suggesting that further analysis was required.

It was expected that the ghosting of series into nearby neighbours will pertain almost exclusively to data rich regions because stations sufficiently distant apart will not have been candidates for a merger in the ISTI databank and are similarly unlikely to have been merged by upstream sources. The issue is likely to be most prevalent for those long-term meteorological series which have been shared widely over the decades, leading to their presence in multiple source data decks that underly the ISTI databank. In different sources, these may have been merged by the source, and/or quality controlled/homogenised which may confound the automated ISTI databank algorithm described in Section 3.2.

Following the application of NOAA NCEI's blacklist, an analysis was undertaken to identify such cases. Each station was compared with its twenty five closest neighbours. The analysis assessed the prevalence for strings of zero differences or repeating annual differences. If 10% of the differences between the candidate monthly values and a given neighbour were either zero or constituted an exactly repeating annual cycle then that pair was flagged and plotted for further consideration. The plot in each such case constituted the difference series, a trace of the data sources that make up each series (which could aid in assessing the probable cause), the contributing station codes, and the number of valid monthly values in the candidate and neighbour files.

This process examined a total of 710,700 pairs of series and based upon the automated similarity criteria check flagged and output 4091 plots for further examination. These were then analysed manually to decide if the extent of duplication was serious enough to warrant the withholding of one or more of the series concerned.

Four distinct groups of cases resulted from this analysis. The first, facile case, already identified was the single duplication event whereby a single series occurred just twice. The second type was double duplications, that is where duplication occurred across three stations. The third type involved multiple duplications that involved duplication across four or more stations. The final type was complex duplication where data sharing occurred across many stations in a multitude of connections and interconnections whereby two or more station series had been ghosted into other series in a complex manner requiring the set of cases to be disentangled collectively. Table 3.6 provides a summary of the occurrences of each of these clusters of differing complexity and their method of resolution.

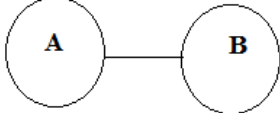
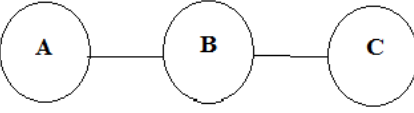
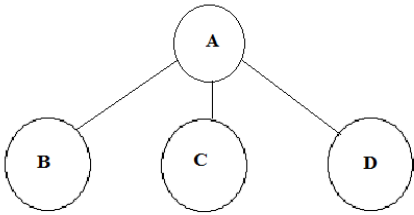
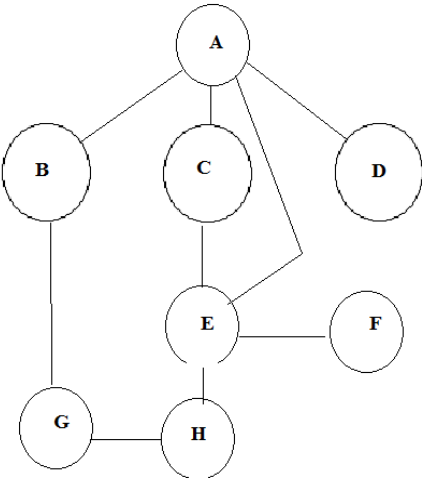
Type of duplicate	Typical Example of Duplication type	Number of cases detected	Stations where individual sources removed in resolving (partial station removal)	Merged	Stations removed in totality in resolving
Single		1143	437	69	207
Double		154	79	0	78
Multiple		118	37	0	35
Complex		75	26	0	31

Table 3.6 Summary of duplication detection and type of duplication of data detected in the ISTI dataset version 1.1.0 along with a summary of the resolution.

3.5.1 Simple paired duplicates

Most of the simple paired series were assessed either to not arise sufficient suspicion or to have sufficiently obvious duplication to warrant further action with very few ambiguous cases. Figure 3.10 shows station USW00003927 at Dallas Fort Worth and station USW00013961 Fort Worth Meacham, 33.24 km apart. In this example, no further action was warranted.

Frequently station series are made up from compiled data taken at different times and on occasion at different locations. These different strands (sources) are often merged to form a single time series. In the construction of the ISTI dataset, each individual data source that forms a time series is recorded in the data file. For example in figure 3.10 station USW00003927 is made up of a single source, source 0. Whereas station USW00013961 is a merge of two sources, source 0 and source 64004300.

Source 0 could have been removed from Fort Worth Meacham to deal with the c.decade of suspicious data at the start of the series, but this would have resulted in the removal of several hundred observations later in the series which did not arouse suspicion. Given that only a subset of the source in Fort Worth Meacham is questionable in this case that points to the issue, if real, arising in the upstream source and not from an erroneous merge decision in the ISTI databank. Decisions such as these were heavily influenced by a reluctance to remove valuable data.

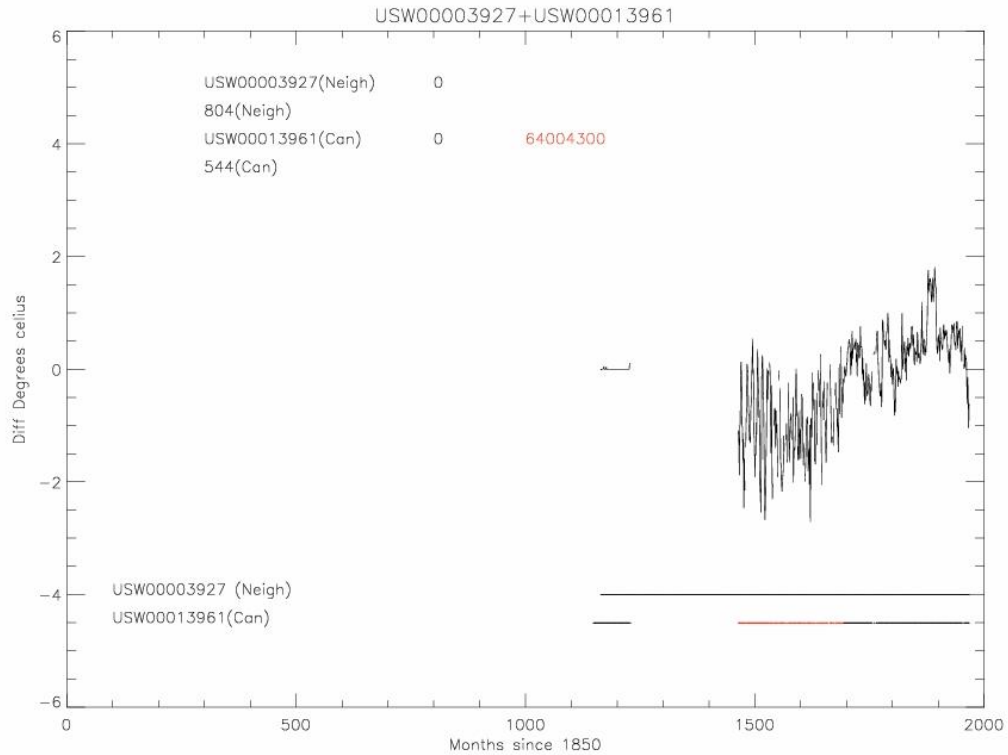


Figure 3.10 An example of a series that was assessed as not having sufficient similarity to warrant further action. The candidate minus neighbour series is the black trace in the center. The lower traces show in colour-code which segments of each station arise from each source and the identifier used in that source (if given and distinct from the station identifier itself).

Figure 3.11 is an example where one source was removed from a merged station series in the ISTI databank to eliminate the duplication. Station AR000087692, located at Balcarce Regional Airport at 37.933°S and 57.583°W, in the province of Buenos Aires, Argentina and station ARXLT185239 is located at Vila Gesell 38.0°S and 57.1°W also in the province of Buenos Aires, Argentina. The sites are 43 km apart. In this case, a source-based overlap unambiguously points to a poor merge decision in the ISTI databank construction process. Removal of this source (identifier 59013210) from the neighbour removes in entirety the duplication apparent in the plot leaving several decades of good observations in both series.

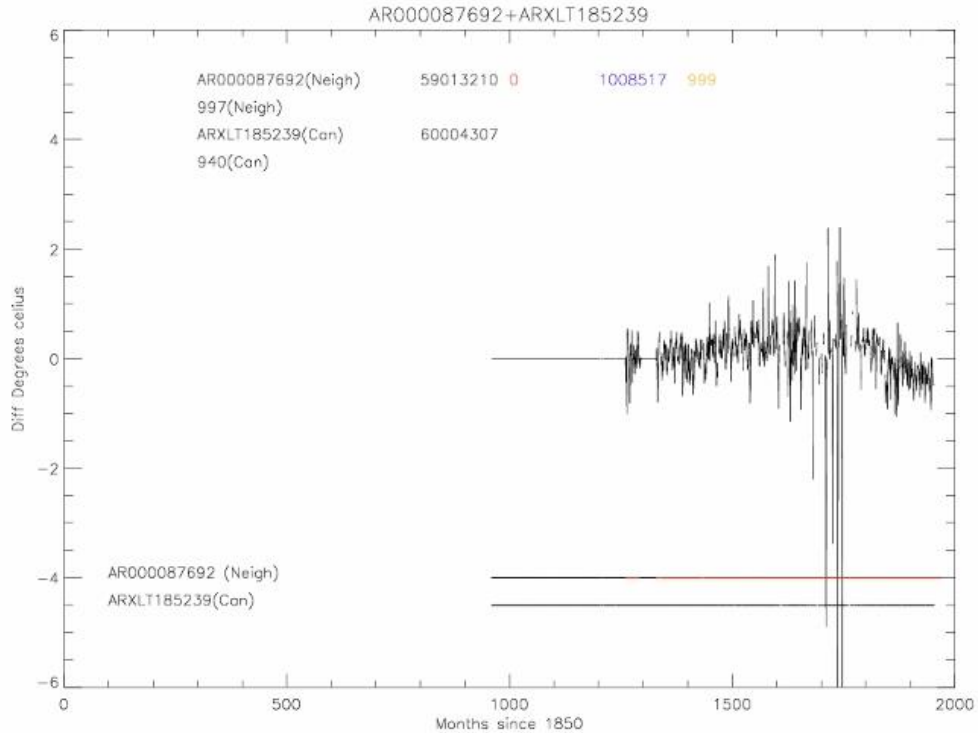


Figure 3.11 Example from Argentina. As the station at Villa Gesell is made up from a single source and the station at Balcarce airport is a combination of three sources, with source 59013210 overlapping with the string of zeros, source 59013210 was removed in resolving the issue.

An extreme example of exact data match that results from a poor decision in the ISTI databank merge can be seen in Figure 3.12. Station ARXLT052250, Santa Cruz Purero, Argentina at 50.2°S and 68.4°W and Station ARXLT508951, Santa Cruz, Argentina at 50.016701°S and 68.566704°W (ISTI databank coordinates reproduced exactly) are 23.6 km distance apart. Yet their data during overlap is an exact match. Station ARXLT50891 Santa Cruz was removed to resolve the problem. Such errors are to a degree unavoidable given the statistical nature of the algorithm (Section 3.2). In retrospect, the method could likely be improved through the addition of a data duplication check, and this will be considered in ongoing work by NOAA NCEI and C3S (Thorne et al., 2017). In total 516 stations had a single or double contributory source removed as a result of such paired station overlap comparisons in addition to the removal of 354 stations in entirety in helping to resolve the single and double duplication issues. Only a minority (69) of these were due to ISTI merging decisions with the remainder arising from upstream data source merges.

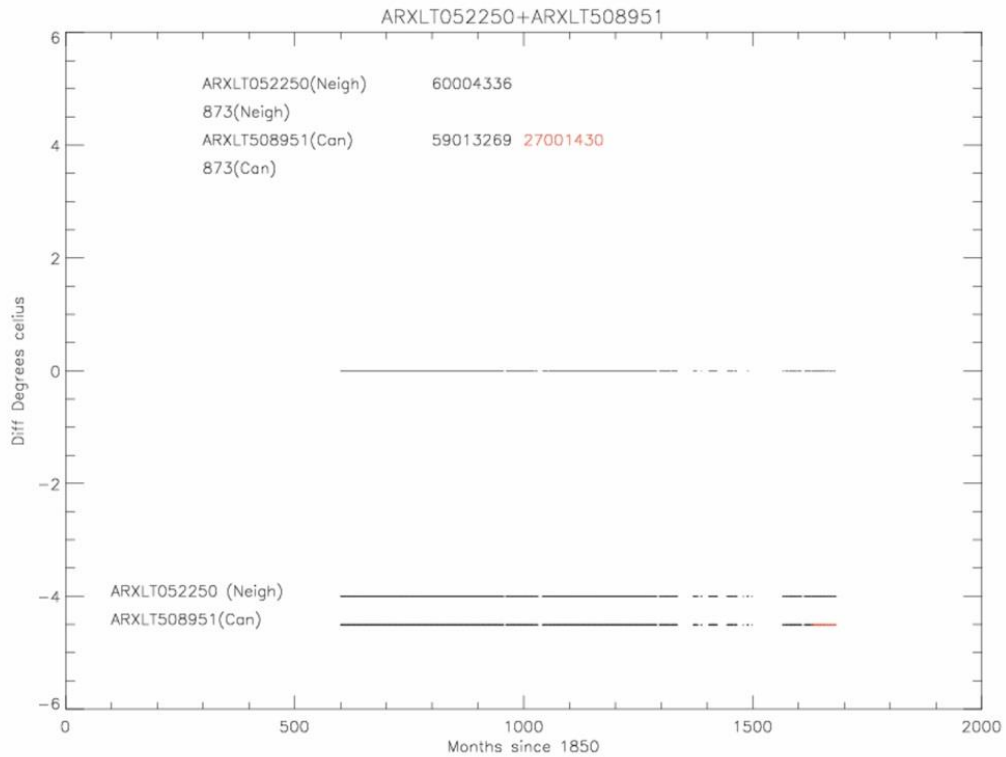


Figure 3.12 An example of exact data match that must result from a poor decision in the ISTI databank merge algorithm as these arise from the non-GHCND source deck.

3.5.2 Complex Cases

To disentangle the complex and multiple station cases required the comparison of multiple paired station plots to ascertain similarities and diagnose the underlying causes. Often these highlighted the ghosting of a single long time series into multiple nearby series (Figure 3.13) leading to a decision to retain the most complete series of the set. But it also highlighted clusters of the type A-B, A-C, A-B-C and even more complex interconnections as indicated in table 3.5 requiring the deletion of more than one overlapping series in any single file.

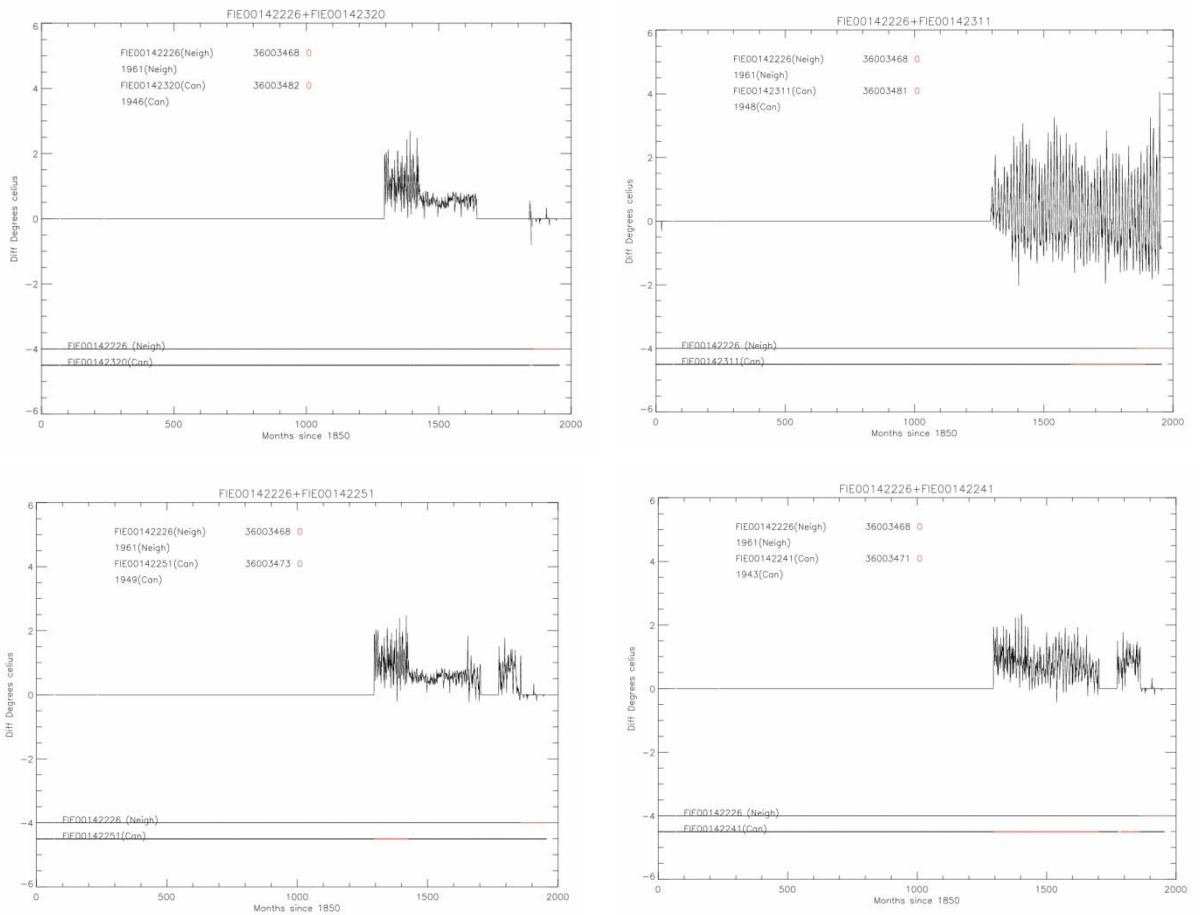


Figure 3.13 Example where a single station's observations are duplicated with many neighbouring stations in a complex arrangement. In this case, station, Helsinki Kumpula, shares data with twelve other stations. Only 4 cases are shown here.

With complex cases, such as Helsinki Kumpula (the file that first flagged the issue), there are multiple cases where the data existed in subsets of other series. In these circumstances, an extensive analysis was required to untangle the complex interrelationship as can be seen in Figure 3.14. This particular issue required the deletion of whole files as the extent of the overlaps involved meant the removal of a single source was not sufficient to resolve the problem and the removal of more than one source from many series often meant that a very small amount of data remained (Table 3.7). Overall, six stations were removed in totality and three had sources removed to address the issues found at this location.

Full Station Removed	Stations With Source Removed
FIE00142320	FIE00142331
FIE00142311	FIE00142326
FIE00142251	FIE00142096
FIE00142111	
FIE00142241	
FIE00142235	

Table 3.7 Summary of those files removed in full or in part to resolve Helsinki Kumpula identified overlaps.

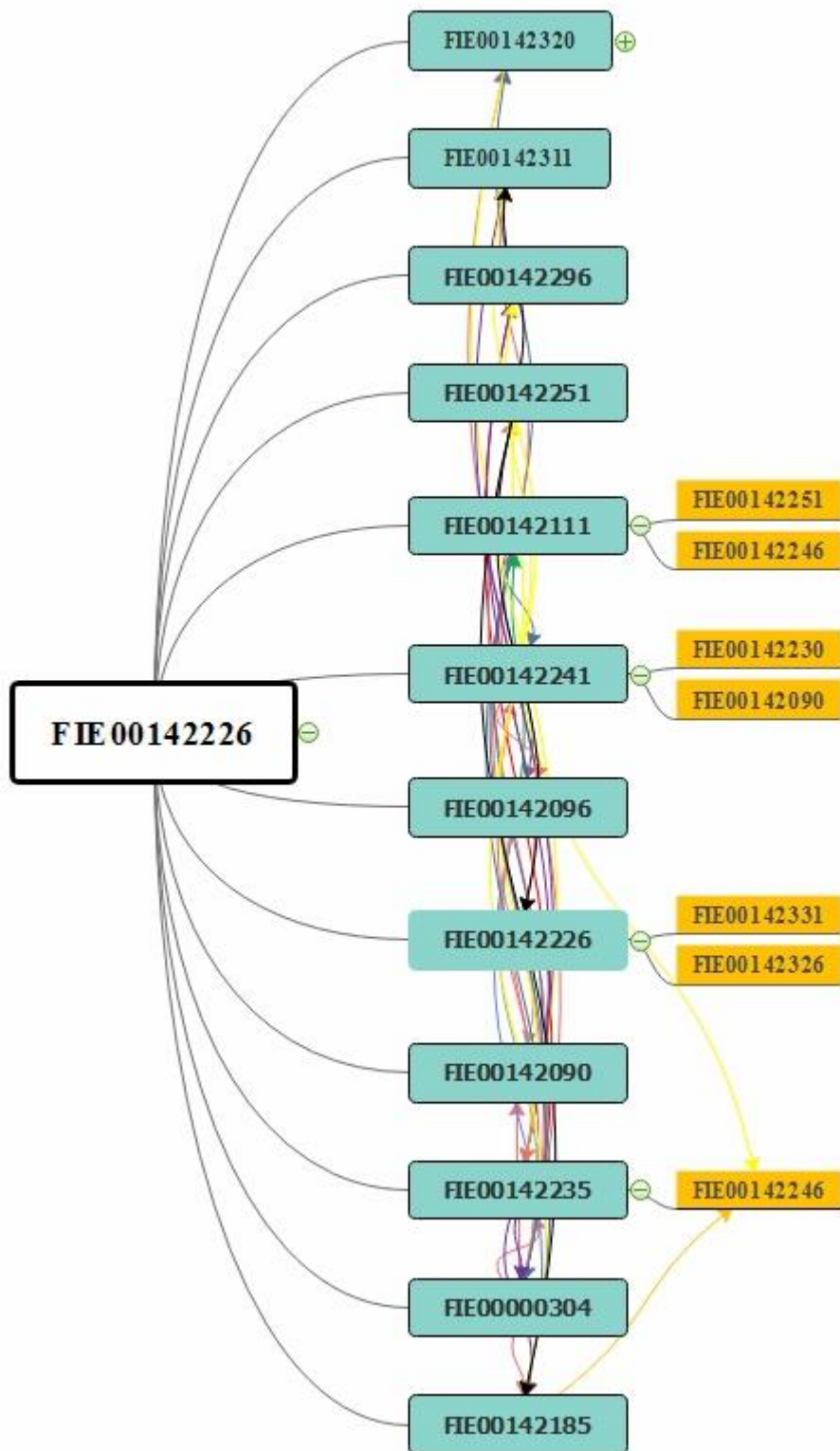


Figure 3.14 This figure illustrates the interconnection and duplication between Helsinki Kumpula Finland, station FIE0042226 and twelve of its neighbours, several of which also share commonalities with additional stations shown on the right. Helsinki is by far the most complicated case with multiple stations being ghosted into one another in full or in part and several stations containing segments arising from multiple other stations in ways that would require expert local knowledge to satisfactorily resolve.

3.5.3 Resolution of cases of stations with identical coordinates

A regional analysis uncovered two stations in Canada that had identical latitude and longitude coordinates, but with different identification codes and variations in the name. Such a result is possible, particularly where the coordinate resolution is coarse (e.g. a 0.1 degree resolution coordinate has a c.10 Km radius in which the true location may exist). Equally, it is known that in several regions of the world the WMO identification system has been changed at various points in time e.g. Canada being given a new block in the 1980s. Some countries have also tended to recycle their use of identifiers or to rename stations without moving them. Such changes may have confounded the automated ISTI databank algorithm. Out of an abundance of caution, it was decided to not permit two series to have exactly matching locations in subsequent analysis steps.

The exact location match issue is most prevalent in Canada and the USA, where the ISTI databank is most dense, accounting between them for in excess of 75% of the identified cases. Table 3.8 is an example of the metadata for such a station pair at Powell River, Canada with Figure 3.15 highlighting when data is available in each series. In this case of Powell River, the overlap between observations was almost complete with station CA001046392 containing more observations than CA001046391 owing to starting earlier. However, there are two short gaps with missing data in CA001046391 that perhaps could have been infilled with a spliced section from CA00104392. Because the missing data was only short sections and to avoid the possibility of introducing inhomogeneities unnecessarily, it was decided in this case to delete CA001046392 completely.

Station Pair	Name	Latitude	Longitude	Altitude(meters)	No. Obs
CA001046391	Powell River A	49.833302°	-124.5°	130	681
CA001046392	Powell River	49.833302°	-124.5°	125	528

Table 3.8 The Powell River example of duplication in Canada

Stations CA001046392 and CA001046391

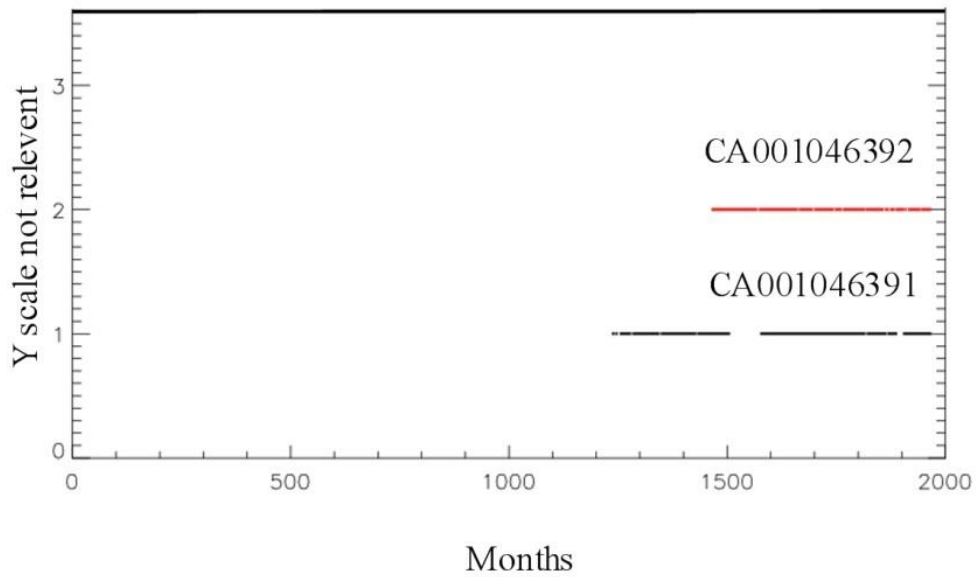


Figure 3.15 The data sources that made up Powell River duplication. Each trace shows data availability. In this case station CA001046392 was removed in resolving.

A second example is Metoryuk in the USA (Table 3.9) where the station name changed considerably. Figure 3.16. shows no overlap between the observations and merging was deemed appropriate in this case as a result. Here, because the name had changed so dramatically with only one letter being similar it is assumed that the automated ISTI databank merge had not deemed the metadata sufficiently similar and had erroneously decided that the data were unique. This may have been further compounded by the time series gap between the two series.

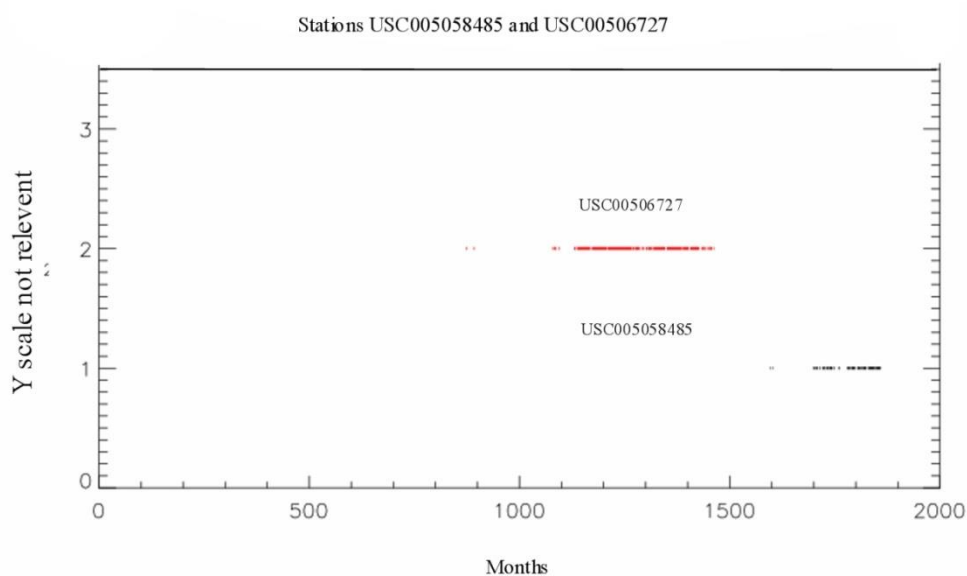


Figure 3.16. As Figure 3.15 but for Metoryuk/Nunivak example where there is no overlap between the two series and a merge was performed.

Station Pair	Name	Latitude	Longitude	Altitude(meters)	No. Obs
USC00505845	Mekoryuk	60.3833°	-166.199°	13.69	129
USC00506727	Nunivak	60.3833°	-166.2°	15.8	332

Table 3.9 The Metoryuk/Nunivak example, deemed to be the same site with two different names.

In total after consideration of all such cases, a total of 166 stations were removed and 58 were merged leading to a reduction in the station count of 224. (Table 3.10).

Resolution	Number of cases
Deleted because of full or near full data overlap	166
Station series merged into other stations and then removed	58

Table 3.10 Station location exact match resolution summary

This is a conservative approach that may remove true parallel measurements between e.g. a long-running manual and replacement AWS sensor set-ups. Hence these decisions, while appropriate for the current work, may not be appropriate for all potential applications and expert local knowledge may help to disentangle a number of these issues. Canada represented 165 of these cases, the USA 41 and the rest of the world only 18.

3.6 Discussion

The ISTI databank merged in excess of 70 underlying data sources using an automated approach to create a single, uniformly formatted, database suitable for climate applications. Despite the careful work of Rennie et al. (2014) to merge many underlying sources and to remove duplication, a number of errors remain in the database that had to be resolved herein. In addition, short segment records with fewer than a decade's worth of observations have been screened out. NCEI discovered duplication and other apparent data issues when constructing GHCNMv4 (Rennie, pers. comm.) and this blacklisting was applied to this work in the first instance. Subsequent inspection of 29 selected case study stations highlighted ghosting of either entire series or segments into neighbour series. Consideration of all stations and their 25 nearest neighbours led to the removal of many station series and deletion of segments in several more. Finally, several stations were found to be identically located and out of an abundance of caution only one series was retained either through a merge if the series did not meaningfully overlap or through retaining the longest of the series in other cases. All findings herein have been communicated back to NOAA NCEI.

Table 3.11 summarises the impacts of all post-processing applied to the ISTI databank, while Figure 3.17 provides a geographical summary of the locations impacted. By far the largest volume of station removal is uncontroversially arising from the removal of those stations containing less than a decade's worth of observations. The next largest removal arises from series ghosting and is concentrated in Germany and Scandinavia, although with occurrences in every permanently inhabited continent. The final issue is exact coordinate matches which are predominantly a North American issue. The total effect of all blacklisting and removals is to reduce the station count from an initial 35,919 stations to a final count of 27,639 a reduction of 23.1%, of which fully 20.9% was due to the removal of short duration station records.

Processing Step	Station Removed in Entirety	Station Modified
Removal of Short Period Records	7491	
NCEI Blacklist	268	
Station Duplication	351	579
Exact Location Match	166	58

Table 3.11 A summary of the processing steps undertaken in the pre-processing of the ISTI databank performed herein

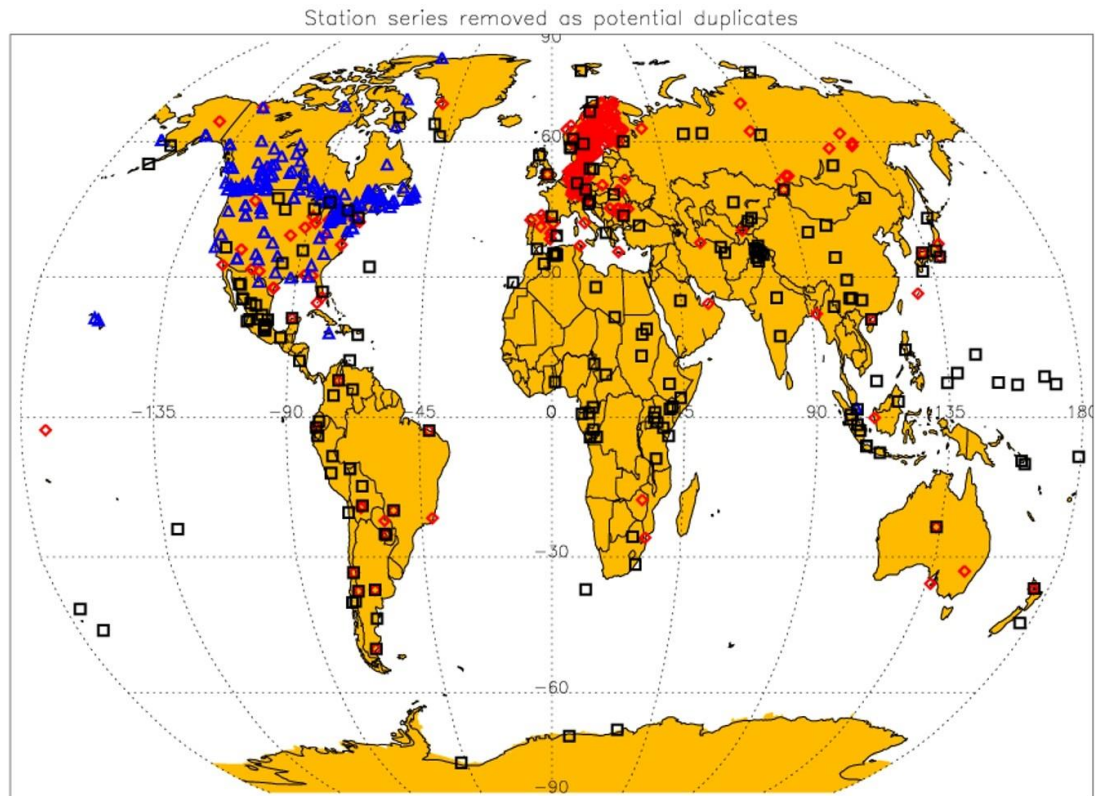


Figure 3.17 Summary of the locations of removed files due to: the NCEI blacklisting (black squares); accounting for station series ghosting (red diamonds) and exact location matches (blue triangles). This map does not include files deleted because of fewer than 120 months of observations. For that see Figure 3.4

For finding station ghosting, the search criteria used herein was limited to stations with over 10 years of records and to the 25 nearest neighbours. Given the manual nature of the process employed it was not possible in the limited time available to look more broadly for cases. In the US and Europe, the 25 nearest neighbours can be in very close proximity and additional issues of ghosting will likely remain. Furthermore, short period records may well be more susceptible to this issue. A future more comprehensive screening might look for cases irrespective of distance or record duration and may well arise a number of further data issues not picked up on

here. As an example, for the De Bilt case study station, one further ghosting occurrence was picked up when expanded to find the 25 nearest neighbours with at least 50% overlap. This particular case was also removed.

Irrespective of the selection of potential cases of station series ghosting, following the detection of potential cases manual inspection was required. This manual inspection introduces an irreducible degree of subjectivity. In this instance, and for the purposes of the current thesis, a degree of leeway was given in cases where a lot of potentially usable data could be removed in remedying apparent data issues. In future work, more surgical approaches than the two options of removal of entire sources or entire stations employed herein may be advisable.

The presence of stations in the ISTI databank with exactly the same coordinates is more curious given the weighting given to coordinates, altitude and name in the metadata metrics in the ISTI merge. Several of these may well be real parallel observations undertaken across the conversion between e.g. manual and automatic sensors. At a minimum, the data must be distinct as these cases were not picked up in the data ghosting exercise which preceded the analysis of geographically coincident measurement series. Such data are potentially hugely valuable to many applications. To resolve such issues would require institutional/local knowledge that we, sadly, do not have. The retention of a single longest series or merged series is a pragmatic choice that minimises the potential risk for a global long-term trend characterisation of inadvertently double-counting stations and over-weighting individual point estimates.

Given the number of data issues discovered in the present chapter, even after the application of the GHCNMv4 blacklist, it is inevitable that GHCNMv4 as originally published contains at least a substantial subset of the duplication issues we have uncovered. The presence of exact duplicates and/or annually repeating duplicates will potentially confound the PHA method employed by NCEI. That algorithm searches for differences in pairwise behaviour (Chapter 2) and clearly when neighbours are an exact match or annually repeating match the statistical assumptions underlying the algorithm will be substantively violated. When multiple pairs contain similar spurious behaviour: i) this is harder to detect; and ii) that behaviour is more likely to be inadvertently projected onto changepoints associated

with neighbouring stations. Their retention will also add undue weight to single series over other series in the grid box average series. Impacts will be largest at the grid box level in those regions where the issue is substantial. However, overall conclusions reached by GHCNMv4 regarding e.g. global temperature trends are unlikely to be adversely affected given the distribution of the issues found and the number of stations that remain unaffected. Jared Rennie has confirmed that our blacklisting has now been applied to GHCNMv4 operational updates which means that these issues should no longer pertain to the monitoring version of the product.

3.7 Conclusion

The ISTI dataset as used in the present thesis has been substantively analysed for potential data issues including the prevalence for short series (<1 decade) and record duplication/overlap. The problem of duplication is strongly regionalised to richly sampled areas of the globe. The processing has led to the removal (sometimes by merging) of a total of 684 station segments and 425 station series in entirety, beyond the NCEI originated blacklist for GHCNMv4 (which itself removed 268 stations), from further consideration. This exercise although extensive did not necessarily weed out all cases of duplications and future users of the ISTI databank or similar holdings should use due caution. All the decisions have been communicated back to NOAA NCEI and are reported in the supplementary information to Gillespie et al. (2020). It should be noted that the ISTI databank shall in the medium-term be superseded by efforts to create an integrated set of holdings across variables and timescales from synoptic to monthly (Thorne et al., 2017).

Chapter 4 Assessment of the utility of 20th Century Products as a reference series for homogenisation of land surface temperatures

4.1 Introduction

Scientists have been collecting and analysing global land surface air temperature records for a very long time. As discussed in Section 2.2, Callendar put together the first truly global collection of temperature estimates in 1938 and concluded that carbon dioxide from the burning of fossil fuels was partly responsible for the warming of the climate that he had assessed over the previous fifty years. Since then datasets have become progressively more complete, and the methods used in their creation have become more advanced (Section 2.5). Notable current global and regional works include, but are not limited to, CRUTEM now at version 5 (Osborn et al., 2020), GHCN-M now at version 4 (Menne et al., 2018), E-OBS (Cornes et al., 2018), and the Berkeley Earth dataset (Rohde et al., 2013b). Although these datasets are in broad agreement in terms of global mean behaviour, even at these scales potentially important divergences occur prior to the mid-20th Century (Blunden and Arndt, 2019).

Despite numerous advances, the creation of long-term climate data records remains a challenging proposition. Meteorological observations were generally taken to observe and predict local and regional weather and not to monitor long-term climate change. Change in the records has been ubiquitous and has often been beneficial in that these changes often resulted from improvements in methods or equipment or other changes that enhance the accuracy of the recording. The original ‘raw’ data are very often biased as a result of a wide range of factors well-reviewed in the literature. These include station moves, urbanisation effects, instrument changes, land cover changes, and observation practice changes amongst others (Peterson et al., 1998, Trewin, 2010, Parker, 1994, Changnon and Kunkel, 2005). The degree to which these biases do not represent the true climate evolution complicates attempts to quantify climate variability and change unless adequately identified and adjusted for (Willett et al., 2014).

Compounding this, long-term data are only available for certain locations, with relatively few meteorological measurements having been performed quasi-continuously since the 19th century (Bronnimann et al., 2013, Rennie et al., 2014). These locations are not distributed equitably across the global land surface and are concentrated in Eurasia, North America, and parts of Australasia. It is, therefore, a challenge to infer truly global estimates of long-term change.

Recently, the International Surface Temperature Initiative (ISTI) has undertaken an open and transparent effort to recover, combine and create a database of ‘original’ (raw) monthly land surface air temperatures from historical observational records, with an emphasis on provenance and completeness (Rennie et al., 2014). This database in its current iteration contains more than 35,000 individual station records (although many are short-period records). It is the most extensive global collection of instrumental land surface air temperature series produced thus far. It increases by approximately 3-fold the number of station series that were available to researchers prior to its assembly, with improved spatial completeness back to at least 1850 (Rennie et al., 2014). To date, it has been homogenised to create both a new version of the Global Historical Climatology Network Monthly product – GHCNMv4 (Menne et al., 2018), and an estimate of Diurnal Temperature Range changes (Thorne et al., 2016). Both of these have utilised the operational version of NOAA NCEI’s Pairwise Homogenisation Algorithm (Menne and Williams, 2009) to create bias-adjusted station series. To better quantify the uncertainty in homogenised data products arising from the ISTI databank, it is imperative that a broader range of methodological approaches be explored to probe the structural uncertainty in surface temperature records derived from these holdings (Thorne et al., 2005b).

Such novel approaches could include using climate reanalysis products. Over recent decades these products have been generated starting with the NCEP/NCAR reanalysis (Kalnay et al., 1995, Kistler et al., 2001), with several groups developing full-input reanalysis products with the most recent versions being the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 (Hersbach et al., 2020) the Japanese Meteorological Agency’s JRA-55 (Kobayashi et al., 2015), and NASA’s Modern- ERA Retrospective Analysis for Research (MERRA-2) (Gelaro et al., 2017). ECMWF reanalyses have been successfully used, instead of neighbour-based approaches, to homogenise radiosonde temperatures (Haimberger et al., 2012).

More recently, surface-only sparse-input reanalysis products that extend back to the 19th Century have been produced (Poli et al., 2016, Laloyaux et al., 2018, Slivinski et al., 2019, Compo et al., 2011). Most specify fields of homogenised sea surface temperatures and sea ice concentration as a lower boundary condition (Titchner and Rayner, 2014, Rayner et al., 2005). All assimilate only surface pressure or surface winds and pressure as a dynamical constraint to reconstruct the full atmospheric state over the globe. They are thus formally and fully independent of land surface air temperature observations and any time averages derived from them. A number of precursor comparisons of these products to meteorological observations of land surface air temperature (Jones et al., 2012, Wang et al., 2018, Parker, 2016, Ferguson and Villarini, 2012, Compo et al., 2013) imply close correspondence, at least over certain regions and periods, but with potential caveats. For example, Ferguson and Villarini (2012) highlighted good correspondence from the mid-20th Century onwards but with the potential for a spurious break over the Central United States around the mid-20th Century in the NOAA-20CR version they analysed. As with traditional full-input reanalysis products, successive generations of sparse-input reanalysis products show improved quality as we learn from previous efforts and as data assimilation techniques and model skill improves (Slivinski et al., 2019).

This chapter sets out to assess whether using the latest generation of sparse-input reanalysis products may plausibly constitute an alternative approach to homogenise the ‘raw’ ISTI monthly databank holdings. This could provide a valuable methodologically-independent estimate of the necessary adjustments to these fundamental data holdings. The present analysis is a necessary precursor to such a homogenization effort by evaluating critically whether the primary building block of the new method, sparse-input reanalysis fields, can provide suitable comparator-series for the homogenisation of land surface air temperature series

Having outlined the context, the remainder of the chapter is structured as follows. Section 4.2 considers the options of constructing a comparator series for homogenisation and outlines the role that sparse-input reanalysis products could play. Section 4.3 details the reanalyses products used and the interpolation method employed to arrive at a reanalysis-based comparator record. Section 4.4 examines the potential that using the reanalysis ensemble members may have over the ensemble mean and compares reanalyses to pairwise approaches at the station level.

Section 4.5 compares the potential for homogenisation applications of sparse-input reanalysis and pairwise approaches at various aggregations. Section 4.6 provides a discussion of the findings, and conclusions are given in Section 4.7.

4.2. Possible approaches to constructing comparator series

Homogenisation of station time series to remove non-climatic influences from the record is essential to estimate the underlying climate record. The goal is to remove artificial non-stationarities in a series (“breaks”) while retaining any real trends (Menne et al., 2009, Venema et al., 2012). Homogenisation of a candidate station record thus requires some form of comparator series. Acquiring a suitable and robust series can be a challenge. The series must contain a reasonable approximation to the real geophysical variations experienced at the candidate station to avoid misappropriating real climate variability and trends arising from data artefacts. Fundamentally, a comparator series needs to be as highly correlated with the target series, and with as low noise (small variation in the differences), as possible. The higher the correlation and lower the noise the smaller the breaks in the candidate series that can be detected and the lower the propensity to falsely identify breaks (Menne et al., 2009, Williams et al., 2012).

In a perfect world scenario, consulting the station’s comprehensive metadata would be the solution to breakpoint detection, identifying where shifts or discontinuities may be expected (Trewin, 2010). In such a scenario, whenever an instrument had been changed or a station moved there would have been a period of parallel measurements undertaken and these series would also be available. Furthermore, all sites would have been well maintained and all siting would follow stipulated criteria that ensure representativeness. There would also exist a backbone of high-quality traceable reference stations (Thorne et al., 2018). Sadly, in the real world, very often metadata are incomplete or missing, parallel measurements are rarely made and even more rarely openly shared, many sites are sub-optimal, and there exists, at least historically, no absolutely traceable reference network. Thus those interested in creating data records must confront

the challenge of working with data series that are poorly documented and highly likely to contain unresolved issues arising at unknown times.

In some cases in instances where suitable neighbours are absent, researchers have used sections of the record of the station under examination that they have high confidence in to homogenise suspect sections of the same record (Peterson et al., 1998, Mamara et al., 2012). But most techniques now routinely employ the use of several nearby stations in the same region (Peterson and Easterling, 1994). Early neighbour-based techniques used some form of neighbour averaging (or compositing), but a growing recognition that quasi-contemporaneous or large breaks in neighbours might lead to their misattribution has led to most modern techniques using some form of multiple pairwise comparison techniques (Venema et al., 2012). These start by finding all potential breaks by comparing, in turn, each station to every other station within a given set of stations and then proceed via logical elimination to ascertain whether detected breaks most likely exist in a candidate series or in individual neighbours.

For homogenisation, the individual station records must be both of sufficient length and overlap substantially to be able to inform on relative time series characteristics. The ISTI databank consists of station records of varying duration, period of observations, and completeness such that it is very much the exception rather than the norm to have a 1:1 correspondence in data availability between any pair of stations. This means that any particular comparison can typically only elucidate potential data issues in a subset of the candidate station series under consideration. Figures 4.1, 4.2, and 4.3 provide indicative examples of the challenges that pertain in using the nearest 25 neighbouring stations to homogenise the longest station series including step-changes in neighbour density and comparators repeatedly appearing and disappearing.

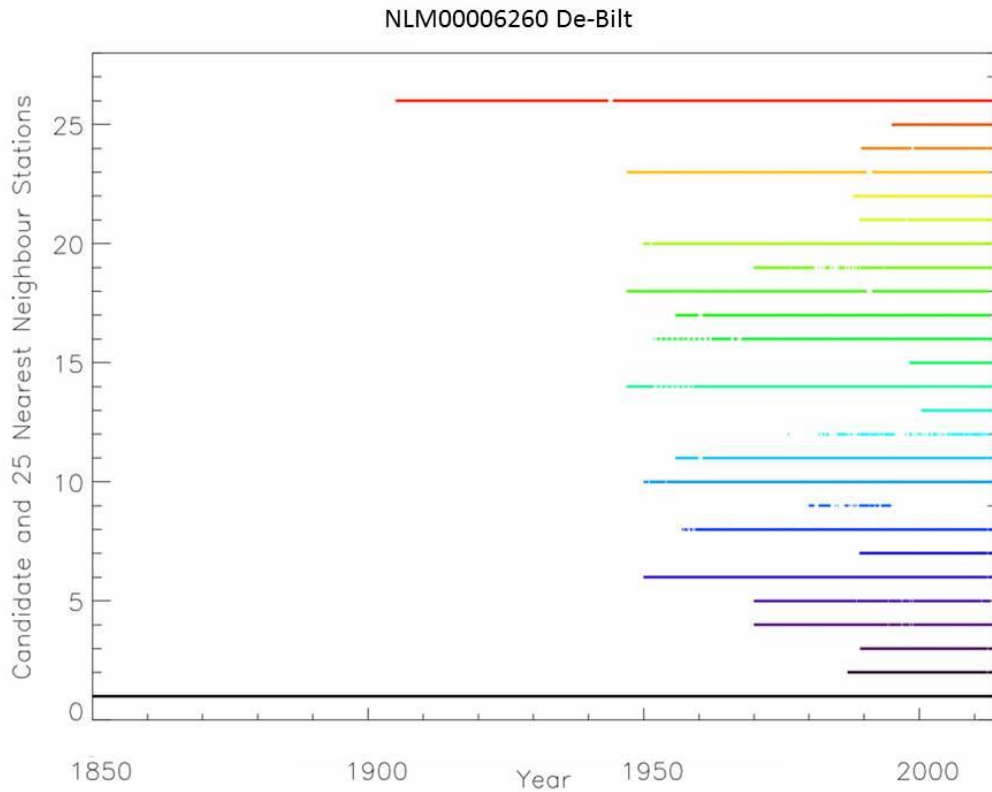


Figure 4.1 Summary of neighbour station data availability for De Bilt since 1850 (the series extends to the 1700s but for the present study the interest is in the period since the 1850s driven by the availability of sparse-input reanalysis products and globally representative observations). This series is a centennial station series with almost continuous availability (bottom black) since the 1850s although prior to 1897 data arises from Utrecht and then several additional sources: http://projects.knmi.nl/klimatologie/daggegevens/antieke_wrn/index.html. Within the ISTI databank data, 1901 to date arises from the KNMI hosted E-OBS. Data prior to 1901 arises from GHCNMv2 collection which appears to arise directly from KNMI. The 25 nearest neighbours (other colours) are shorter with no suitable neighbour amongst them to use for homogenisation in the 1850 to 1900 period. There is one potential neighbour for the period of 1900 to 1945, after which there are several possible neighbours for pairwise homogenisation. Effectively pairwise homogenisation techniques are not possible for the period of 1850 to 1945 without expanding the neighbour search radius due to a lack of suitable neighbours.

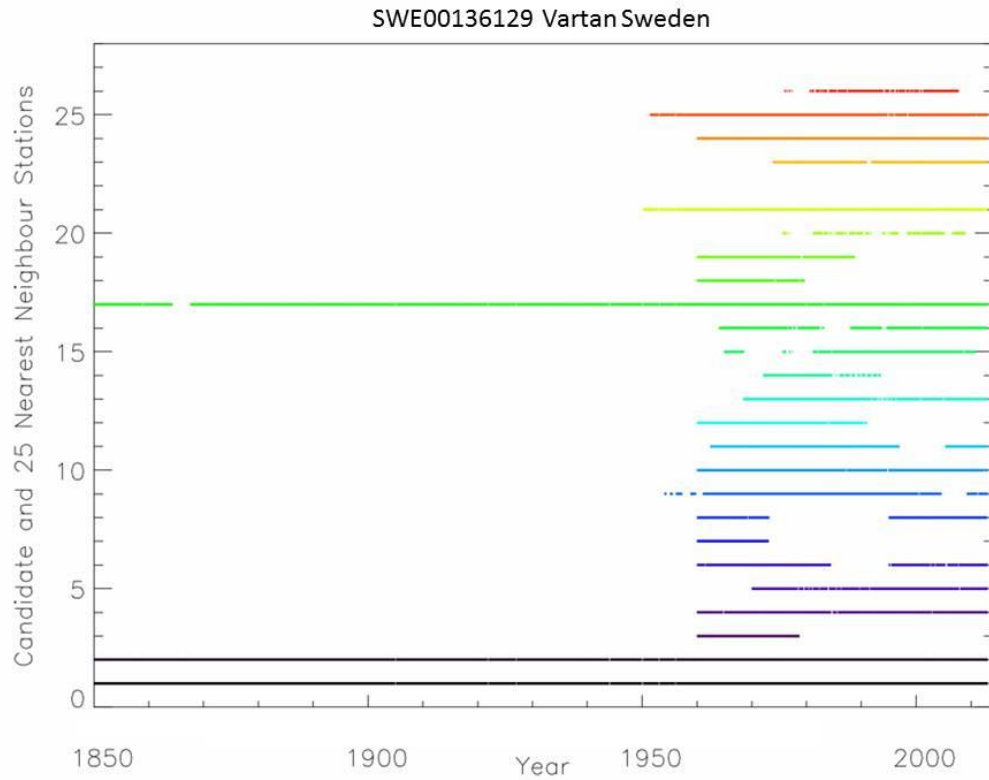


Figure 4.2 As Figure 4.1 for Vartan, Sweden since 1850 (the series again extends back before 1850). A limited set of pairwise comparisons would be possible throughout the series but with a marked step-change in capability around 2/3 of the way through the series when a substantial number of neighbour series become available.

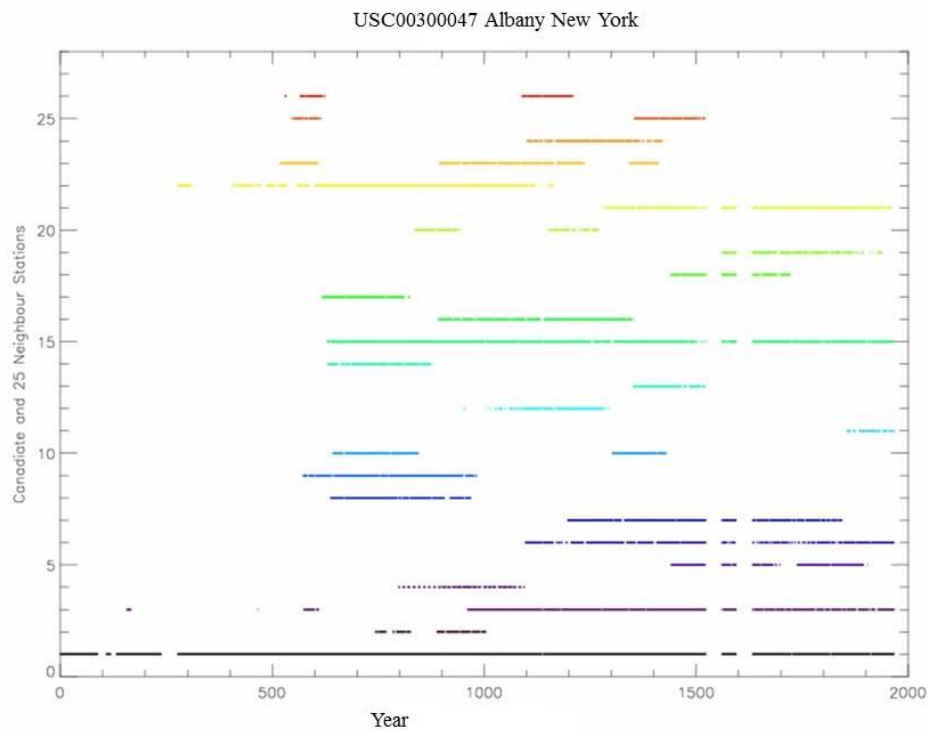


Figure 4.3 As Figure 4.1 but for Albany, New York (the series again extends back earlier than the 25 nearest neighbours (other colours) which are much less complete and frequently drop in and drop out with no suitable neighbours amongst them to use for homogenisation for the entire period.

As a novel alternative, series from reanalysis products offers the potential to circumvent many of these issues. Sparse-input reanalysis products (Poli et al., 2016, Laloyaux et al., 2018, Slivinski et al., 2019) can extend back to the mid-19th Century and include surface temperature estimates consistent with the prior forecast field, assimilated meteorological measurements (which exclude the land surface temperatures), and any specified boundary conditions. Reanalyses will thus always have a corresponding value to every station observation over the common period of record and are substantively independent. The use of full-input modern period reanalysis products that assimilate considerable additional data from radiosondes, aircraft, satellites, etc. to homogenise radiosonde records has proven effective (Haimberger et al., 2012). The question remains whether this is more broadly the case and, specifically, whether the centennial-scale sparse-input reanalysis products can perform a similar function for land surface air temperatures.

4.3 Interpolation of sparse-input reanalysis gridded series to station locations

Observational data were drawn from the post-processed set of station data holdings arising from the ISTI databank release v1.1 (Rennie et al., 2014, Lawrimore et al., 2015). The post-processing applied to these data prior to their application herein is described in Chapter 3.

Sparse-input (20th Century) reanalysis products are relatively recent additions to the family of reanalysis products. Pioneered by NOAA and the University of Colorado, they have now been produced also by ECMWF and are under preparation elsewhere. Recourse is made to four versions of these reanalysis products arising from ECMWF, NOAA and the University of Colorado:

1. The NOAA-CIRES Twentieth Century Reanalysis version 2c (20CRv2c) provides 2° by 2° resolution estimates over 1851-2012 generated with an Ensemble Kalman Filter (EnKF) algorithm. Use is made of both the ensemble mean product and the underlying 56 ensemble members (Giese et al., 2016).
2. The ECMWF ERA-20C reanalysis, produced under the EU funded ERA-CLIM project, provides a deterministic estimate (single analysis with no uncertainty) on a 1° by 1° grid from 1900 to 2010 using a 4D-Var algorithm (Poli et al., 2016).
3. The NOAA-CIRES-DOE Twentieth Century Reanalysis version 3 (20CRv3) is a comprehensive update of previous versions of 20CRv2c. It has an improved resolution of approximately 0.7° by 0.7° covering the period from 1836 to 2015 and an ensemble of 80 members (Slivinski et al., 2019). Solely the ensemble mean is considered herein as the full ensemble of 80 members was released only after the present analysis was completed. 20CRv3 benefits from an upgraded EnKF data assimilation algorithm and an improved NOAA atmospheric model. The observational constraint benefits from an enhanced observational database version 4.7 of the International Surface Pressure Databank (Cram et al., 2015) (Figure 4.4) from data rescue efforts, a new variational quality control algorithm, a new bias correction for marine

observations before 1871, and an updated bias correction algorithm for all station data over land (Slivinski et al., 2019).

4. The ECMWF CERA-20C product is a coupled reanalysis product with a 1° by 1° resolution extending from 1900 to 2010 with a ten member ensemble (Laloyaux et al., 2018).

All of the sparse-input reanalyses used here are available upon a regular grid. To construct a comparator series of monthly 2m air temperature estimates using reanalysis for each target station it is thus necessary to interpolate the gridded estimates to the station locations. Several possible interpolation methods exist of varying complexity. Interpolation by inverse distance weighting (IDW) is a popular method which is computationally efficient and considered to be relatively accurate (Willmott and Robeson, 1995). IDW is strongly recommended where the points to be interpolated are dense enough to capture local variation (Childs, 2004) and reduces any concern about topographic complexity that may generate micro-environments impacting on climatic values (Vicente-Serrano et al., 2003). The IDW equation is as follows (Childs, 2004):

$$V = \frac{\sum_{i=1}^n \left(\frac{v_i}{d_i^2} \right)}{\sum_{i=1}^n \left(\frac{1}{d_i^2} \right)} \quad (\text{Eqn. 4.1})$$

Where n is the sample size, in this case 9, v is the value at each grid point, and d is the distance from the grid point to the station position

Vicente-Serrano et al. (2003) compared twenty-five different interpolation methods for temperature and precipitation in the Ebro Vally Spain, a region selected for its geographic heterogeneity and spatial climate diversity. They found that inverse distance weighing performed well with a coefficient of determination (r^2) of 0.72

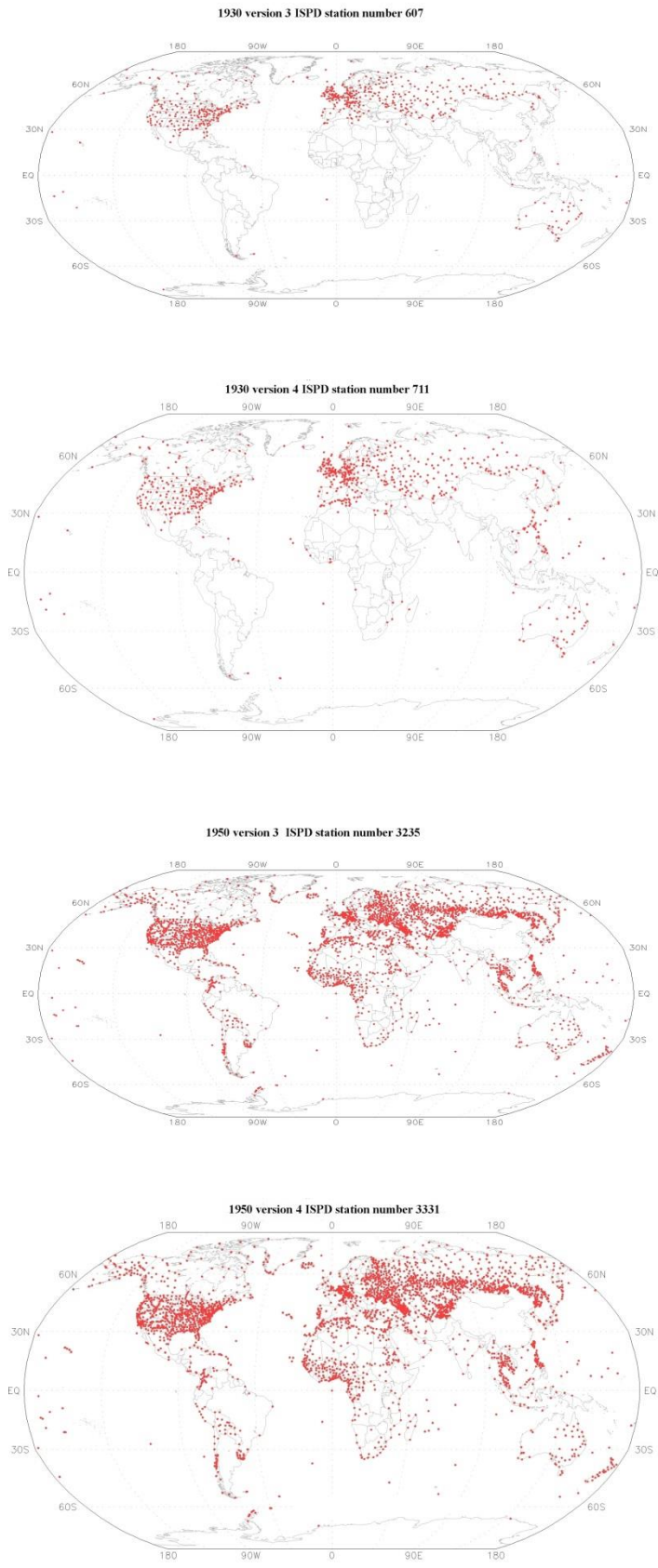


Figure 4.4 Maps showing ISPD V3.7 & V4.7 for 1930 and 1950 (Source <https://psl.noaa.gov/data/ISPD/>)

and was only bettered by a regression model with r^2 slightly better at 0.75 and regression model + residuals with r^2 of 0.74.

For each station in turn, a three by three grid of the nine nearest reanalysis grid points is used to interpolate to provide a station temperature estimate. A station located at the equator would have a maximum diagonal grid distance of 630 kilometres for 20CRv2c decreasing significantly as it nears the poles. The other, finer resolution, reanalysis products (ERA-20C, 20CRv3, and CERA-20C) are under half these distances. Two methods of inverse distance weighting were considered: (1) inverse distance weighting and (2) inverse distance squared weighting. Weighting was applied to the absolute values, that is the recorded observations before any adjustment was applied and the resulting series was then anomalised after matching to target station data availability to ensure a complete mirroring of station observations to the sparse reanalysis estimate for the time period. Using a station period of record climatology (that is not normalised to any thirty year averaging period) at this stage maximises the retained pool of station records. Results were compared individually in anomaly space for a selection of global stations (Chapter 3) and evaluated using Pearson correlations and standard deviations. On an individual station basis, there were only negligible differences (Table 4.1) with, in general, a very marginal performance advantage using inverse squared distance weightings. Given the very marginal differences, subsequent sections and Chapter 5 consider only the inverse distance squared weighting approach.

Station	Name	Country	Inverse distance squared weighted interpolation		Inverse distance weighing interpolation		Differences	
			Sigma	r	Sigma	r	Sigma	r
AR000087828	Trelew_Aero	Argentina	1.096	0.628	1.096	0.628	0.000	0.000
ASXL2209646	Hobartasmanwas_949700	Australia	0.857	0.510	0.857	0.510	0.000	0.000
ASXL2263670	Perthauswas_946080	Australia	0.759	0.749	0.759	0.749	0.000	0.000
AYM00089314	Theresa	Antarctic	1.937	0.715	1.937	0.715	0.000	0.000
AYXL1563342	Erin	Antarctic	1.756	0.751	1.756	0.751	0.000	0.000
CA003031400	Carway	Canada	1.511	0.846	1.530	0.842	-0.019	-0.004
CHM00058362	Shanghai	China	1.072	0.635	1.087	0.626	-0.015	-0.009
CI000085469	Isla_De_Pascua	Chile(Easter IIs)	2.247	0.219	2.290	0.203	-0.043	-0.016
CIXLT967829	Santiagowas_855770	Chile	1.139	0.463	1.181	0.452	-0.043	-0.011
FIE00142226	Helsinki_Kumpula	Finland	1.283	0.860	1.302	0.857	-0.020	-0.003
FJ000091652	Udu_Point_Aws	Fiji	0.598	0.477	0.598	0.477	0.000	0.000
GMM00010628	Geisenheim	Germany	0.828	0.896	0.839	0.894	-0.011	-0.001
INM00043057	Bombay_Colaba	India	0.664	0.574	0.695	0.560	-0.031	-0.015
ITE00115588	Padova	Italy	0.989	0.790	1.005	0.788	-0.016	-0.001
JA000047817	Nagasaki	Japan	0.771	0.787	0.775	0.786	-0.004	-0.001
LH000026730	Vilnius	Lithuania	1.229	0.875	1.251	0.874	-0.022	-0.001
MT000016597	Luqa	Malta	0.682	0.765	0.678	0.769	0.003	0.004
MZXL2405557	Lourenco_Marques	Mozambique	1.197	0.405	1.200	0.404	-0.003	-0.001
NLM00006260	De_Bilt_1	Netherlands	0.929	0.859	0.937	0.858	-0.008	-0.001
NOE00134898	Tromsolangnes	Norway	1.085	0.826	1.106	0.825	-0.022	-0.001
PKXL2983863	Quettasheikh_Manda	Pakistan	1.484	0.516	1.504	0.508	-0.020	-0.008
RSM00023662	Tolka	Russia	1.717	0.895	1.717	0.895	0.000	0.000
RSM00028722	Ufa	Russia	1.432	0.873	1.448	0.872	-0.016	-0.001
SPE00120143	Huelva_Ronda_Del_Este	Spain	0.706	0.855	0.706	0.855	0.000	0.000
SWE00136129	Vartan	Sweden	1.123	0.856	1.136	0.854	-0.013	-0.002
TZXL2095229	Dar_Es_Salaam_Tanzania_Beaf	Tanzania	0.919	0.333	0.912	0.345	0.007	0.012
USC00300047	Albany	USA	1.448	0.756	1.489	0.754	-0.041	-0.002
USC00500252	Amchitka	USA	0.488	0.766	0.488	0.766	0.000	0.000
ZI000067975	Masvingo	Zimbabwe	0.861	0.652	0.861	0.652	0.000	0.000
Average			1.131	0.694	1.143	0.692	-0.012	-0.002

Table 4.1. Comparison of interpolated reanalysis minus station difference series using inverse linear distance and inverse linear squared distance for correlation and Sigma using the 20CRv2c ensemble mean product for interpolation to selected stations (Chapter 3). The difference between the methods on both an individual basis and an aggregate basis for both sigma and correlation are small with a slight overall improvement when using the inverse squared distance approach. Given that this product is the coarsest resolution reanalysis, differences are smaller for other reanalysis products considered (not shown).

4.4 Analysis of relative performance of sparse-input reanalysis ensemble averages and ensemble members

The availability of many of the sparse-input reanalysis products as both an ensemble mean product and individual ensemble members yields questions as to how to appropriately treat these data in a homogenisation-exercise context. Individual ensemble members constitute fields that are each in-turn geophysically consistent with the model physics but may contain substantial random effects. Conversely, the ensemble mean will not in itself constitute an estimate that is entirely consistent with the model physics at any given timestep but may contain reduced random effects compared to any given individual ensemble member and therefore be better correlated and less dispersive from any particular target station series.

To investigate the relative value of individual ensemble members compared to the ensemble mean recourse is made to 20CRv2c. The 20CRv2c product comes as a 56-

member ensemble and the ensemble mean. We examined the correlations and standard deviations of the differences between the 20CRv2c ensemble mean and the station anomalies and compared this to that for the 56 individual ensemble members and the 25 nearest neighbours. Figures 4.5, 4.6 and 4.7 show examples of the individual ensemble members performance against the ensemble mean performance and these 25 neighbours for selected stations (many more stations were inspected manually). The ensemble mean for 20CRv2c for these stations shows preferable statistical properties than the individual ensemble members. In almost all cases the individual ensemble members correlations were lower and standard deviations higher than the ensemble mean. This is consistent with expectations if the ensemble spread is principally random in nature, which is commensurate with the documented ensemble design principles (Compo et al., 2006, Compo et al., 2008). Table 4.2 highlights ensemble spread and ensemble mean performance across all case study stations. In the few cases where an individual ensemble member performed better than the 20CRv2c ensemble mean, the difference between that ensemble member and the 20CRv2c ensemble mean was very small. It was never the case that more than one or two ensemble members out-performed the ensemble mean.

Figures 4.4, 4.5 and 4.6 also show the relative performance of the different sparse-input reanalysis systems and their comparative power to neighbour based approaches. A preferable reference series should have the highest possible correlation and lowest possible standard deviation overall such as to maximise the chances of finding any breaks in the series of non-climatic origin. For the Bombay (India) site (Figure 4.4) 20CRv3 is potentially preferable to pairwise homogenisation. On the other hand, the site at Perth (Australia) (Figure 4.5) in a well sampled region clearly demonstrates the power of pairwise homogenisation when sufficient neighbours are nearby.

Figure 4.6, the site at Santiago Chile is an example where there is a large distinction between neighbour-based and reanalysis based approaches. In general, 20CRv2c shows preferable statistical properties to ERA-20C and 20CRv3 is better again than 20CRv2c. Neighbour series, with some obvious exceptions, show better correlation but have a considerable spread in standard deviations in all cases.

Station_name (29 Pilot Stations)	Station name	Country	St/20CR v2c mean r	Maximum St/Ensemble member r	Minimum St/Ensemble member r	Diff between Max of ensemble r and 20CR v2c mean	Sigma St/20CR v2c Mean	Max sigma St/Ensemble	Min sigma St/Ensemble
AR000087828	TRELEW_AERO	Argentina	0.628	0.618	0.578	0.010	1.096	1.219	1.153
ASXLT209646	HOBARTTASMANWAS	Australia	0.510	0.490	0.460	0.021	0.857	0.903	0.878
ASXLT263670	PERTHAUSWAS_946080	Australia	0.749	0.711	0.673	0.038	0.759	0.893	0.824
AYM00089314	THERESA	Antarctic	0.715	0.774	0.735	-0.059	1.937	1.862	1.711
AYXLT563342	ERIN	Antarctic	0.751	0.770	0.740	-0.019	1.756	1.855	1.719
CA003031400	CARWAY	Canada	0.842	0.842	0.831	0.000	1.530	1.580	1.531
CHM00058362	SHANGHAI	China	0.626	0.610	0.581	0.016	1.087	1.166	1.119
CI000085469	ISLA_DE_PASCUA	Chile(Easter Is)	0.203	0.236	0.096	-0.033	2.290	2.423	2.333
CIXLT967829	SANTIAGOWAS_855770	Chile	0.452	0.427	0.374	0.025	1.181	1.370	1.312
FIE00142226	HELINKI_KUMPULA	Finland	0.857	0.836	0.808	0.021	1.302	1.533	1.411
FJ000091652	UDU_POINT_AWS	Fiji	0.477	0.464	0.412	0.012	0.598	0.632	0.604
GMM00010628	GEISENHEIM	Germany	0.894	0.893	0.886	0.001	0.839	0.873	0.845
INM00043057	BOMBAY_COLABA	India	0.560	0.515	0.455	0.044	0.695	0.844	0.778
ITE00115588	PADOVA	Italy	0.788	0.782	0.768	0.006	1.005	1.067	1.030
JA000047817	NAGASAKI	Japan	0.786	0.772	0.740	0.014	0.775	0.861	0.800
LH000026730	VILNIUS	Lithuania	0.874	0.862	0.836	0.012	1.251	1.466	1.337
MT000016597	LUQA	Malta	0.769	0.761	0.706	0.008	0.678	0.749	0.684
MZXTL405557	LOURENCO_MARQUES	Mozambique	0.404	0.396	0.329	0.008	1.200	1.300	1.236
NLM00006260	DE_BILT_1	Netherlands	0.858	0.856	0.840	0.002	0.937	0.997	0.942
NOE00134898	TROMSOLANGNES	Norway	0.825	0.791	0.763	0.034	1.106	1.362	1.246
PKXLT983863	QUETTASHEIKH_MAND	Pakistan	0.508	0.487	0.421	0.021	1.504	1.722	1.590
RSM00023662	TOLKA	Russia	0.895	0.891	0.885	0.004	1.717	1.781	1.726
RSM00028722	UFA	Russia	0.872	0.869	0.860	0.003	1.448	1.532	1.466
SPE00120143	HUELVA_RONDA_DEL	Spain	0.855	0.848	0.829	0.007	0.706	0.775	0.723
SWE00136129	VARTAN	Sweden	0.854	0.846	0.825	0.008	1.136	1.212	1.152
TZXTL095229	DAR_ES_SALAAM_TAN	Tanzania	0.345	0.352	0.293	-0.007	0.912	0.954	0.924
USC00300047	ALBANY	USA	0.754	0.748	0.736	0.006	1.489	1.563	1.519
USC00500252	AMCHITKA	USA	0.766	0.777	0.729	-0.011	0.488	0.519	0.476
ZI000067975	MASVINGO	Zimbabwe	0.652	0.587	0.494	0.065	0.861	1.153	1.016

Table 4.2. Summary of comparison of correlations and standard deviation of ensemble members to the ensemble mean and a summary of the comparison of the standard deviation of ensemble member's differences to the ensemble mean differences. The summary shows the maximum and minimum values obtained for the correlation across the entire ensemble versus the station anomalies in comparison to the correlation value of the ensemble mean to the station anomalies and the same for the standard deviation of the difference series of station anomalies minus 20CRv2c ensemble-mean interpolated anomalies.

St = Station values

r = Correlation coefficient

Sigma – Standard Deviation

INM000043057 Bombay India

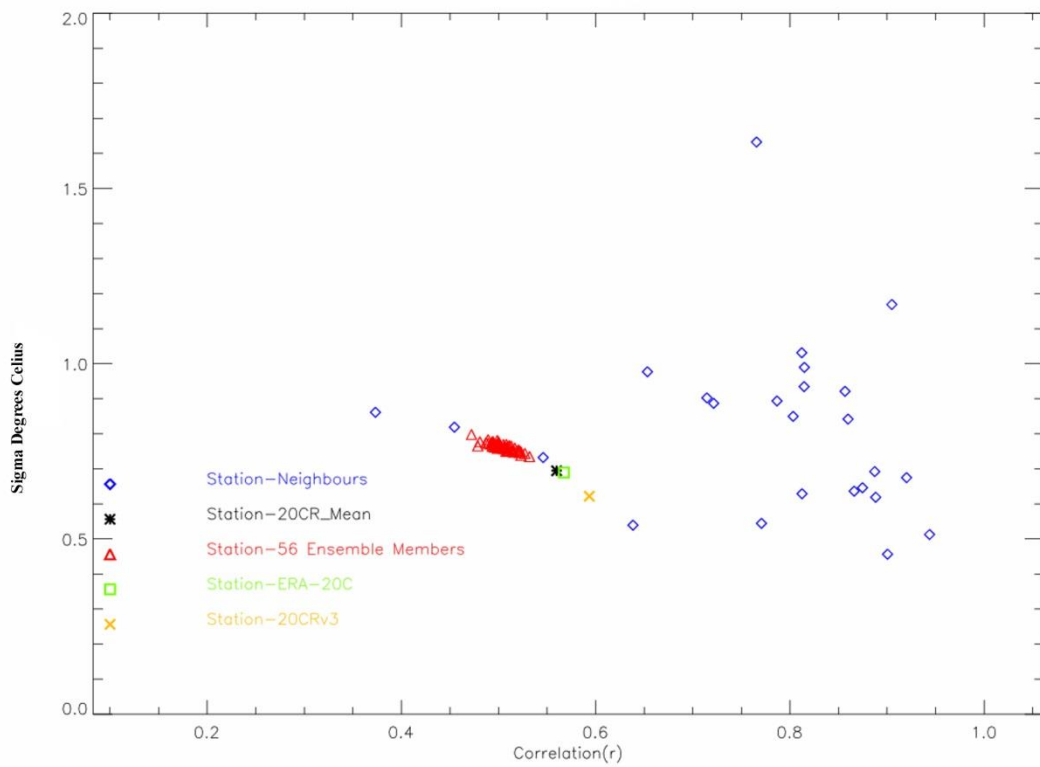


Figure 4.5 An example analysis from Bombay, India (18.9°N, 72.8°E) of correlation (r) and standard deviation (σ) of the different series of the 56 ensemble members of 20CRv2c to determine if the ensemble mean or individual ensemble members are most suitable for further comparison to pairwise homogenisation. This analysis for this station is over the full period of January 1851 to December 2014. There are a total of 1628 observations out of a possible total of 1968 observations. The correlation v standard deviation are plotted for the 25 nearest neighbours, the three reanalysis products and the 56 20CRv2c ensemble members. Values closer to [1,0] would constitute increasingly valuable comparators which are highly correlated with low variability of the difference series.

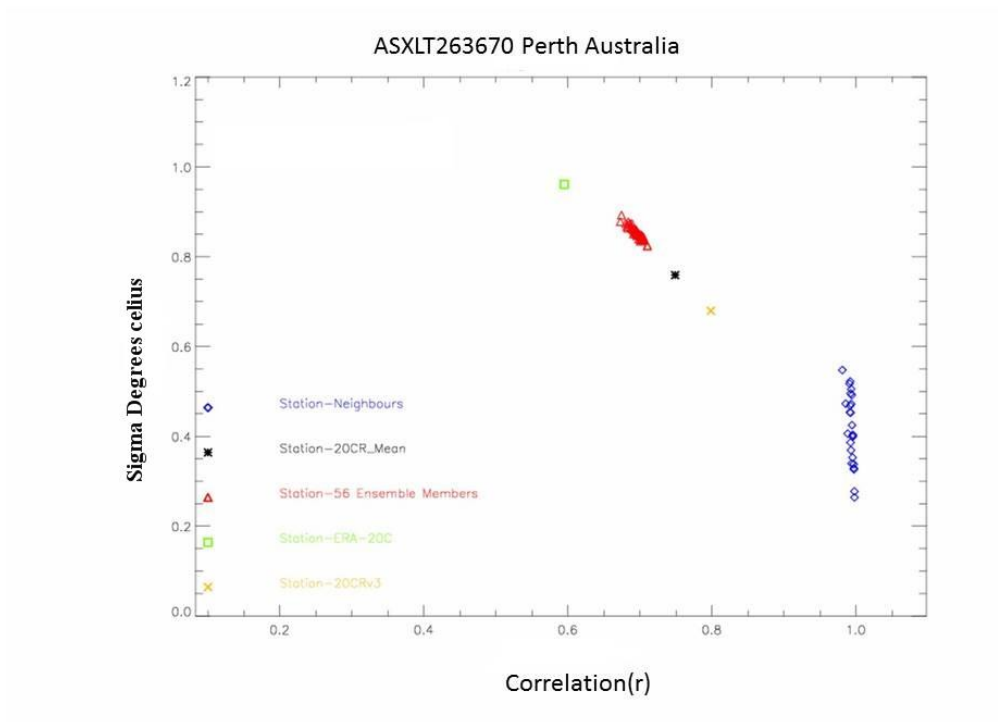


Figure 4.6. As Figure 4.5 but for Perth, Australia (32° S, 115.9° E) This analysis for this station is over the full period of January 1917 to September 2013. There are a total of 1028 observations out of a possible total of 1161 observations.

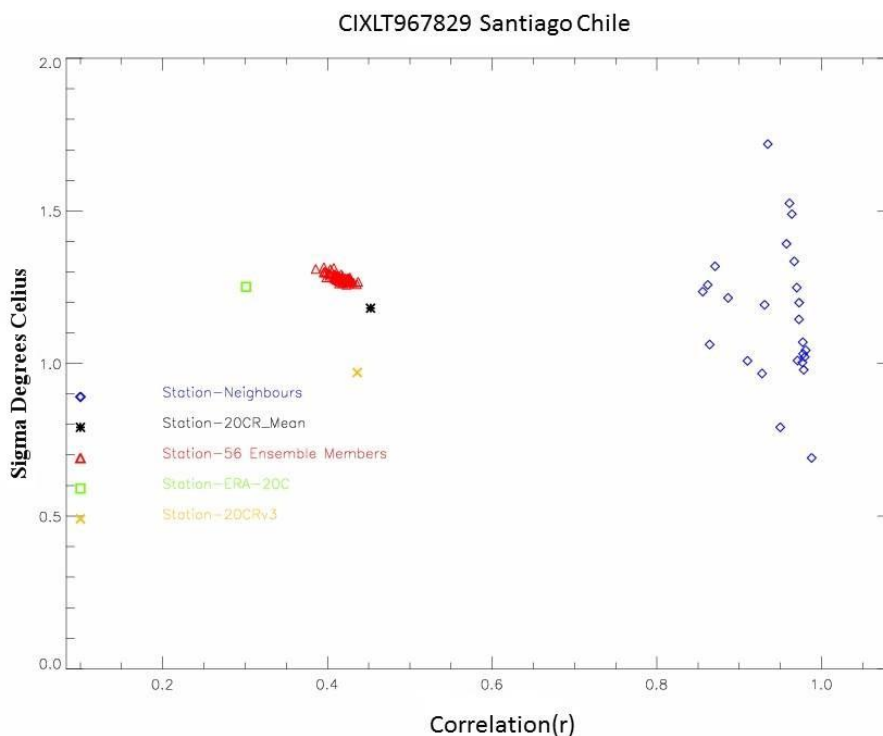


Figure 4.7 As Figure 4.4 but for Santiago, Chile (33.5° S, 70.7° W) This analysis for this station is over the full period of January 1861 to September 2013. There are a total of 1765 observations out of a possible total of 1833 observations.

For the case-study stations analysed, the expectation of the ensemble-mean having preferable statistical properties to underlying ensemble members holds true in the majority of cases. Given these results, it was decided to use solely the ensemble mean series in subsequent analysis. Nevertheless, the improvement is not ubiquitous and there may still be value in using the ensemble members for homogenisation, or indeed other applications, in future work. The 20CRv3 product comes as an 80-member ensemble but at the time of the analysis undertaken here, only the ensemble mean product was available. Initial inspection of the more recently released 20CRv3 ensemble confirms that the 20CRv2C ensemble member vs. ensemble average behaviour found herein holds also for 20CRv3.

Examining the ensemble mean and the ensemble spread for 20CRv3 (McColl and Compo, 2021, Slivinski et al., 2021, Slivinski et al., 2019) the ensemble spread is narrow back to the early 20th century suggesting that 20CRv3 is well constrained and a good estimate as to the state of the climate up to 1900. Before the 20th century not least due to lack of observations both of surface temperature and surface pressure the ensemble spread show a significant degree of variability and as a result, homogenised estimates may be less reliable. By the same token, however and estimate arrived at from PHA homogenisation or any similar pairwise technique is likely to suffer from the same or greater level of uncertainty

4.5 Analysis of suitability for undertaking homogenisation

An evaluation of the applicability of sparse-input reanalysis products to the assessment of homogeneity of individual station series requires an assessment of both individual station correspondence and aggregated spatial differences, under the assumption that after sufficient aggregation station data artefacts, even though individually systematic, become pseudo-random. First, the 29 selected station series are considered in-depth. Then, area-aggregated series are examined using Giorgi regions (Giorgi and Francisco, 2000) to subdivide into regionally-aggregated series. Finally, the relative performance is studied in densely-sampled and sparsely-sampled regions.

4.5.1 Case study stations analysis

Section 3.4 details the criteria for the selection of the 29 case study stations. State-of-the-art homogenisation procedures compare a candidate series for homogenisation in-turn with a number of surrounding neighbours to mitigate against the inhomogeneities that may exist in individual neighbouring stations. To assess the potential performance of sparse input reanalysis datasets against the neighbour-based homogenisation techniques, the correlation and standard deviation of the difference series using the 25 nearest neighbours is compared in Table 4.3 under the assumption that the median neighbour is a reasonable indicator of the overall performance of the neighbours to detect breaks. This avoids negatively biasing the apparent performance of neighbour difference approaches overall if the neighbour set includes a small number of outlier series. Overall, the case study station analysis shown in Table 4.3 highlights that at an individual station level across most of the globe, sparse-input reanalysis-based estimates are now broadly comparable in the key metrics of correlation and standard deviation to neighbour-difference series approaches.

Table 4.3 also demonstrates a marked difference between the performance of the ERA-20C and CERA-20C reanalysis. CERA-20C consistently shows a markedly lower apparent agreement with the case study stations. It is beyond the current analysis to assess rigorously why this may be so, but given the similar version of the ECMWF IFS model versions used it presumably arises from the coupling of the atmospheric model to the ocean reanalysis in some manner. Because of this apparent degradation in performance, ERA-20C was carried forward and the coupled reanalysis of CERA-20C was not used in the remainder of the present analysis.

The advantage that the new sparse input reanalysis products have in terms of contiguous data availability is also of significance and its utility here can be demonstrated in the potential to homogenise long time series extending back into the early nineteenth century. Figure 4.8 shows an example from De Bilt in the Netherlands (the headquarters of KNMI). This series has been well maintained, extends back to prior to the mid-nineteenth century (following some splicing of series prior to the early 20th Century), and is available quasi-continuously through to the present. The difference series to reanalysis (Figure 4.8 top panel) shows a marked

break in the series at around the turn of the twentieth Century relative to both 20CRv2c and 20CRv3. This corresponds to a change in input source in the ISTI databank, although both arise ultimately from KNMI as far as can be ascertained (Rennie, pers. comm.). Detailed KNMI metadata shows that this station was moved 4-5 km from Utrecht to De Bilt in 1897. ERA-20C and CERA 20C do not extend further back further than 1900 and therefore cannot identify this break. Only one of the 25 nearest neighbours extends back to earlier than 1950, meaning that using solely the 25 nearest neighbours (Figure 4.8 middle panel) neighbour-based comparisons are effectively able to elucidate only the latter third of the station series. Even this one series stops prior to the possible 1897 break.

Extending the neighbour search to include stations with >50% overlap (Figure 4.8 bottom panel) permits pairwise comparisons all the way back to at least 1850. These comparisons support the reanalysis-based estimation of a significant data issue arising around the turn of the twentieth century in the ISTI databank version of this series. The comparisons using these more distant neighbours, however, show greater variability than using the 25 geographically nearest neighbours (compare the variability around the mean offsets per station in the middle and bottom panels over their common periods) highlighting the inherent trade-off in pairwise neighbour approaches over selecting proximal versus sufficiently overlapping neighbours.

The results shown in Figure 4.8 are indicative of a broader issue with neighbour-based homogenisation approaches in that contiguous pairwise comparisons for the whole period of record are rare. Across all 29 case study stations, only two stations had neighbours within their closest 25 with paired comparisons exceeding 1800 months in length. The shortest overlapping record was five months between the candidate and a neighbour. This becomes particularly problematic for longer-term analyses as the ISTI databank has relatively few centennial scale station records. In such cases, the current ERA-20C and CERA-20C reanalysis products which start at the beginning of the 20th Century may be of lower utility compared to 20CRv2c and 20CRv3 which extend back to the early to mid-19th Century. The use of reanalysis fields has a clear benefit as there is an estimate for each and every time there is an observation over the reanalysis period of record. However, it is not simply data availability that defines the quality of a comparator series for homogenisation. It also matters how well the comparator is correlated with the target station series and what

are the standard deviation and autocorrelation of their difference series. These properties will collectively determine the likelihood of being able to detect and adjust for breakpoints in the series (Williams et al., 2012).

Individual correlations between the case study stations and their nearest neighbours vary from near 1 to values of less than 0.1 (Table 4.3). Neighbour-based pairwise comparison for the case study stations situated in those areas of the globe that are densely sampled generally exhibit high correlations. For example, for USC00300047 (Albany, USA) the correlation between the candidate and each of the 25 nearest neighbours range from a high of 0.96 to a low of 0.75. However, the distances between the neighbours and the candidate station are small ranging from 11.5 to 43 km. Conversely, the remote case study stations have correlations that are considerably lower, particularly for island stations in the Pacific Ocean. For example, the nearest neighbouring station to Easter Island is over 2000 km away on French Polynesia and is effectively uncorrelated. The station with the best correlation is Juan Fernandez Island at 3000 km distance and with a correlation value of just 0.25.

Similarly, in densely sampled regions, neighbour difference series generally have low standard deviations. In the more sparsely sampled regions, the standard deviations grow markedly for neighbour-based approaches.

Country	Station Code	20CRv2C r	20CRv3 r	ERA-20C r	CERA-20C r	Median Neighbour r	Sigma 20CRv2C	Sigma 20CRv3	Sigma ERA-20C	Sigma CERA-20C	Sigma Median Neighbour	Median number Neighbours Observations
Argentina	AR000087828	0.628179	0.710	0.528	0.559	0.651	1.096	0.894	1.131	1.076	1.103	522
Australia	ASXL209646	0.510256	0.553	0.550	0.349	0.881	0.857	0.828	0.807	1.055	0.419	380
Australia	ASXL263670	0.748808	0.798	0.595	0.606	0.889	0.759	0.680	0.961	0.964	0.520	359
Antarctic	AYM00089314	0.714936	0.823	0.752	0.446	0.602	1.937	1.286	1.471	2.725	2.505	144
Antarctic	AYXL563342	0.751464	0.835	0.774	0.512	0.669	1.756	1.340	1.432	2.595	2.225	155
Canada	CA003031400	0.841768	0.911	0.844	0.765	0.933	1.530	1.186	1.524	1.890	1.019	271
China	CHM00058362	0.625535	0.715	0.824	0.533	0.869	1.087	0.927	0.714	1.211	0.638	743
Easter Island(Chile)	CI000085469	0.203144	0.216	0.190	0.120	0.037	2.290	2.248	2.321	2.457	1.520	415
Chile	CIXLT967829	0.452266	0.436	0.306	0.367	0.448	1.181	0.970	1.207	1.212	1.148	548
Finland	FIE00142226	0.856972	0.858	0.853	0.764	0.973	1.302	1.264	1.251	1.558	0.599	564
Fiji	FJ000091652	0.476834	0.450	0.474	-0.019	0.484	0.598	0.609	0.578	0.969	0.644	286
Germany	GMM00010628	0.894357	0.947	0.908	0.804	0.966	0.839	0.595	0.808	1.251	0.464	652
India	INM00043057	0.559735	0.593	0.580	0.353	0.536	0.695	0.622	0.646	0.840	0.850	746
Italy	ITE00115588	0.788281	0.856	0.874	0.672	0.861	1.005	0.773	0.725	1.267	0.793	1093
Japan	JA000047817	0.785653	0.741	0.853	0.511	0.938	0.775	0.837	0.665	1.181	0.420	824
Lithuania	LH000026730	0.874201	0.924	0.927	0.862	0.951	1.251	0.921	0.928	1.299	0.730	570
Malta	MT000016597	0.769185	0.792	0.788	0.539	0.800	0.678	0.647	0.628	0.956	0.814	632
Egypt	MZXL405557	0.403986	0.468	0.319	0.221	0.225	1.200	1.131	1.256	1.504	1.877	186
Netherlands	NLM00006260	0.857822	0.881	0.952	0.810	0.956	0.937	0.854	0.531	1.084	0.508	512
Norway	NOE00134898	0.824754	0.878	0.882	0.758	0.794	1.106	0.893	0.800	1.271	1.185	545
Pakistan	PKXL983863	0.508271	0.586	0.497	0.604	0.599	1.504	1.345	1.733	1.534	1.518	341
Russia	RSM00023662	0.894767	0.961	0.947	0.891	0.892	1.717	1.049	1.225	1.798	1.707	550
Russia	RSM00028722	0.871945	0.945	0.910	0.824	0.941	1.448	0.877	1.124	1.555	0.939	679
Spain	SPE00120143	0.855083	0.897	0.887	0.663	0.874	0.706	0.562	0.575	1.013	0.626	686
Sweden	SWE00136129	0.853955	0.911	0.912	0.821	0.960	1.136	0.917	0.860	1.017	0.619	633
Tanzania	TZXL095229	0.345024	0.325	0.264	0.114	0.258	0.912	0.917	0.996	1.211	0.902	315
USA	USC00300047	0.754443	0.802	0.777	0.701	0.920	1.489	1.209	1.300	1.265	0.726	464
USA	USC00500252	0.765648	0.722	0.776	0.592	0.613	0.488	0.550	0.456	0.789	2.088	90
Zimbabwe	ZI000067975	0.652156	0.750	0.449	0.402	0.694	0.861	0.771	1.289	1.458	0.870	420

Table 4.3. A summary of the correlations (r) and the standard deviations (sigma, °C) of the anomaly difference series between the station anomalies and the reference which is either a reanalysis data set or the median of the 25 nearest neighbours for: *high-density stations (italic)*: Intermediate stations (regular font): and **stations located in sparsely sample areas (bold)**.

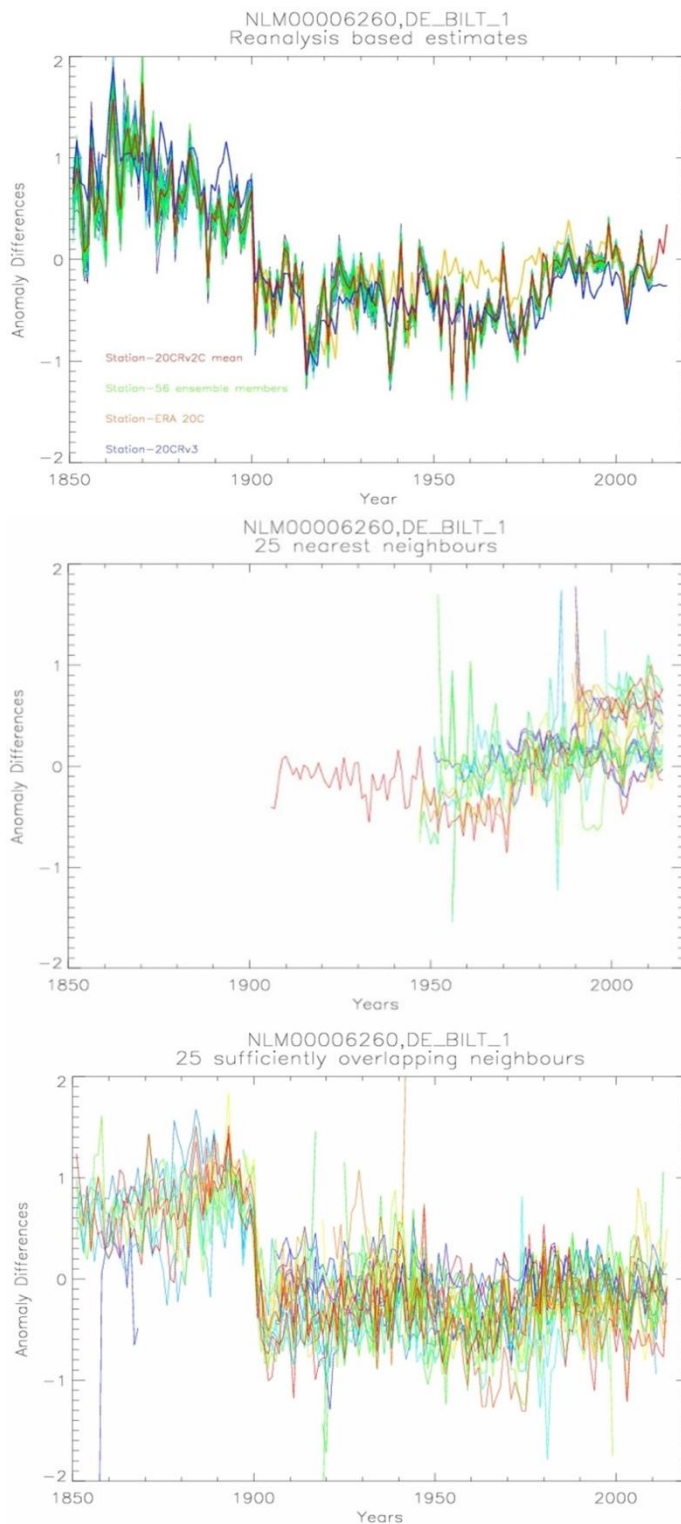


Figure 4.8. Top panel: Anomaly difference series between the long-running De Bilt series in the Netherlands (although note caveats around splicing stations identified in Figure 4.1) for the sub-period of record since 1850 and the sparse-input reanalysis-based estimates. Middle panel: anomaly difference series using De Bilt's 25 nearest neighbours. Bottom panel: anomaly difference series using De Bilt's 25 nearest neighbours with a minimum 50% data overlap. Comparisons are now available for the entire post-1850 portion of the De-Bilt data record, but at a cost to correlation and the standard deviation of the difference series (Table 4.3). In the two lower panels each neighbour difference series is a different colour for illustrative purposes in the middle and lower panels.

4.5.2 Regionally aggregated analyses

While breaks in individual stations will be systematic, when averaged over a sufficient sample size they should become increasingly pseudo-random in nature. Conversely, systematic issues in the reanalyses will tend not to cancel when similarly averaged. Thus an aggregated analysis was performed to elucidate any likely data issues in the sparse-input reanalysis products. This analysis uses the Giorgi regions (Giorgi and Francisco, 2000) and an additional class of Not In Giorgi (NIG) to capture a suite of remote locales. Giorgi regions divide the global land surface area into 21 regions, excluding Antarctica (Figure 4.9). Figure 4.10 illustrates the global distribution of retained ISTI databank stations following the analysis in Chapter 3 into these regions. This illustrates the uneven distribution of meteorological stations, with fully 68.2% of the retained ISTI databank stations (following the analysis detailed in Chapter 3) located in Europe (including the Giorgi Mediterranean region) and the lower 48 states of the USA. Thus 68.2% of the global station network covers only 7.5% of the global land surface.

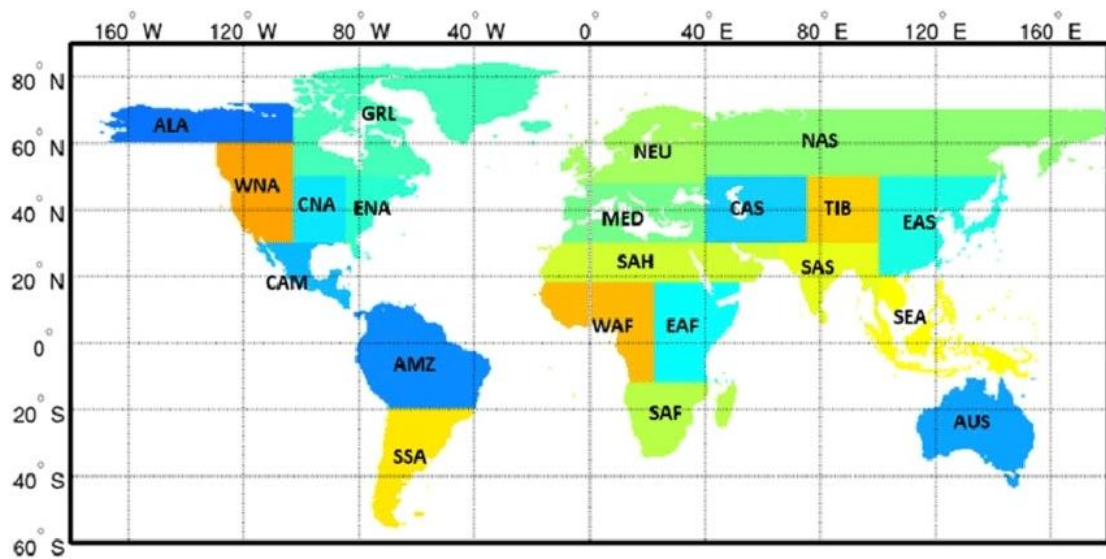


Figure 4.9 Map of regions used for the analysis as defined by Giorgi and Francisco (2000) (Tian et al., 2018).

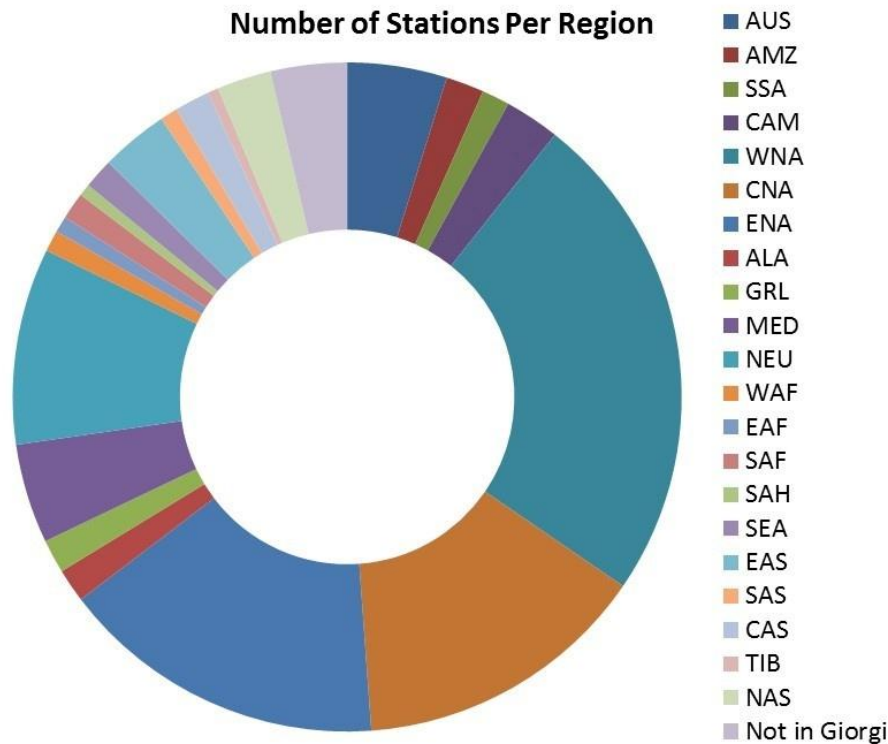


Figure 4.10. The 27,639 long-term stations in the ISTI dataset, following removal of questionable series as detailed in chapter 3, split out into Giorgi region groupings. Note the extra grouping of ‘Not in Giorgi’ which captures the Antarctic, remote islands, and some Arctic sites not included in the original 21 Giorgi regions.

Regionally aggregated series analyses highlight a shift in 20CRv2c around the early-mid 1940s in N. America (in agreement with Ferguson and Villarini (2012)) and also in many other regions (Figure 4.11). This abrupt shift is much reduced in both 20CRv3 (Figure 4.12) and ERA-20C (Figure 4.13). Overall, 20CRv3 shows the best agreement with aggregated station series across most regions of the globe (Figure 4.14) and this performance extends far further back in time compared to the prior generation of sparse-input reanalysis products (compare Figure 4.12 to Figures 4.11 and 4.13). This is consistent with what has been observed for more traditional full-input reanalysis products whereby newer versions, learning from prior iterations and benefitting from innovations in data assimilation techniques and improved models, have markedly improved in various metrics relative to previous generations (Simmons et al., 2017). For 20CRv2c there would be plausible questions about its application for homogeneity assessment prior to the mid-20th Century. In comparison, ERA-20C shows useful performance. However, it is time limited to

1900. In contrast, series from 20CRv3, at least in most regions of the globe, can likely be applied until much earlier and likely to at least 1850 or the instigation of measurements (whichever is the later date).

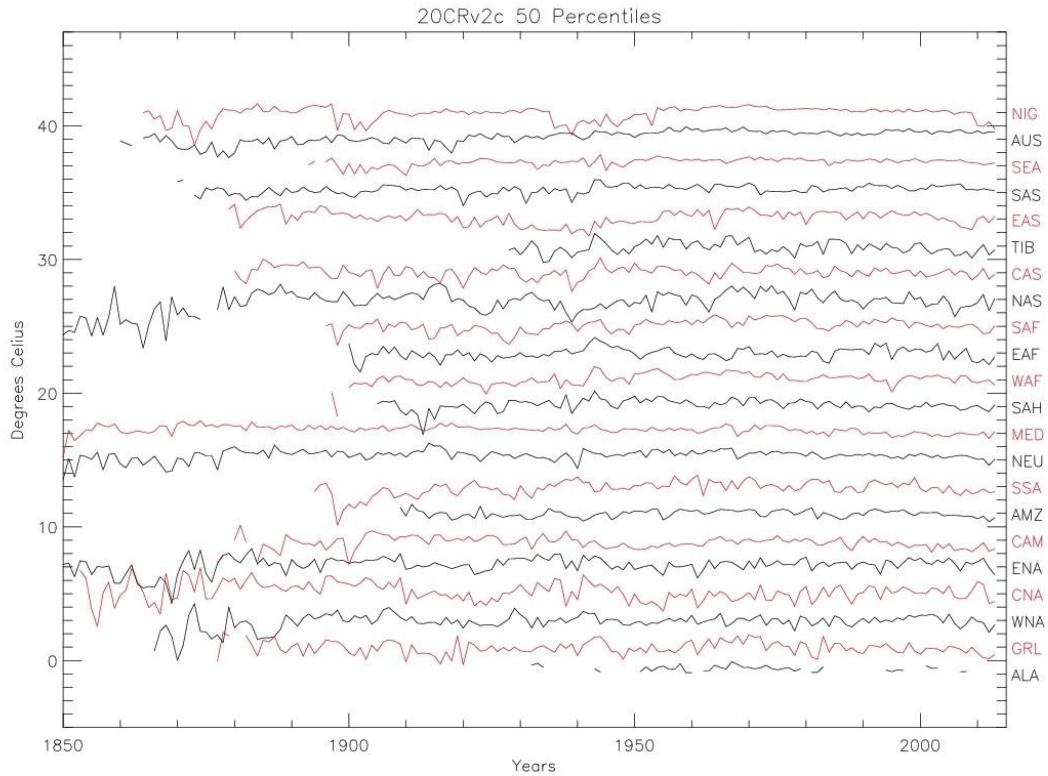


Figure 4.11. The median value (50th percentile) of the regionally-aggregated differences series between 20CRv2c ensemble mean and the station anomalies at each timestep aggregated over the Giorgi regions. Each series is vertically offset for clarity. There is a marked degradation in apparent performance over many regions in the mid 20th Century. For region definitions see main text.

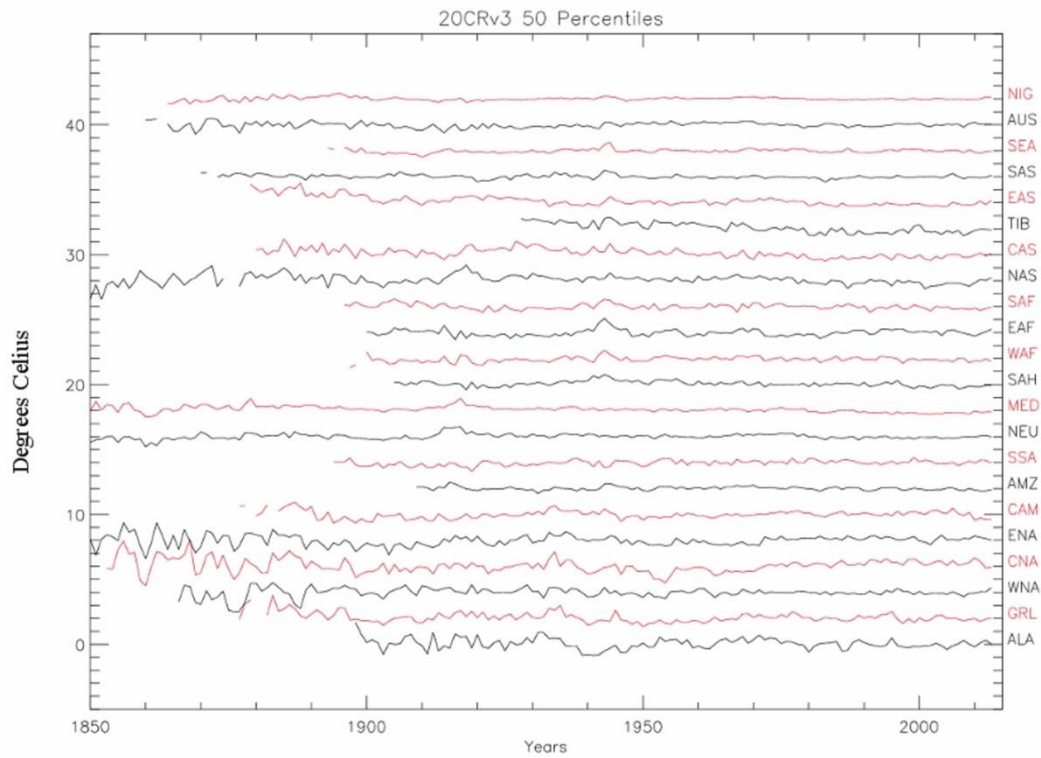


Figure 4.12.. As figure 4.11 but for 20CRv3. The 20CRv3 product shows better performance than either ERA-20C or 20CRv2c across all regions with stable behaviour back to at least 1900 across all regions. The mid 20th Century is much more stable than either of the other sparse-input reanalysis products.

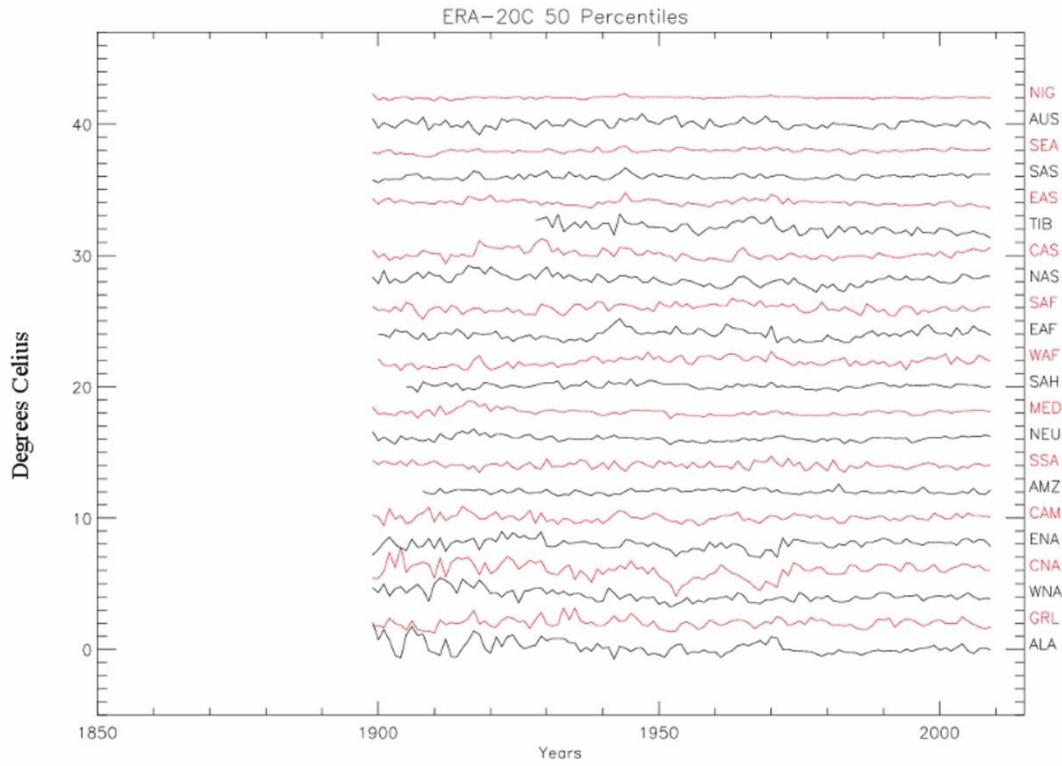
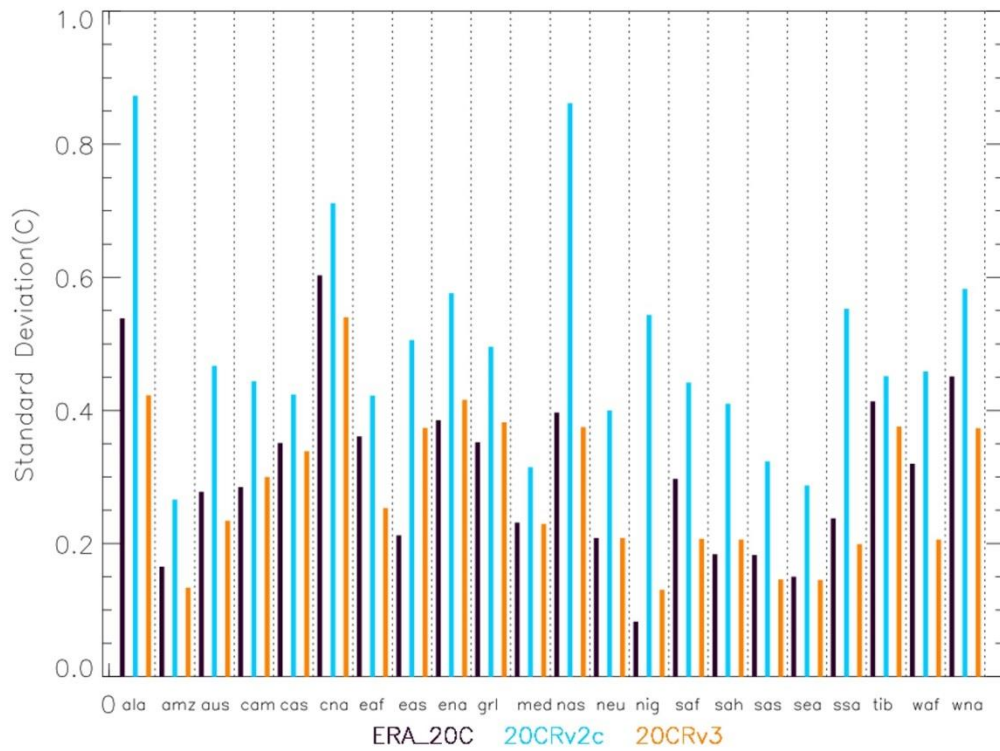


Figure 4.13.. As Figure 4.11 but for ERA-20C which starts in only 1900. This is a clear limitation on the use of ERA-20C compared to the two NOAA sparse-input reanalysis products. Although ERA-20C contains apparent decadal variations in the mid 20th century, the degradation, in this case, is much less marked than for 20CRv2c in most regions (c.f. Figure 4.10).

Giorgi Regions from 1851



Giorgi Regions from 1900 only

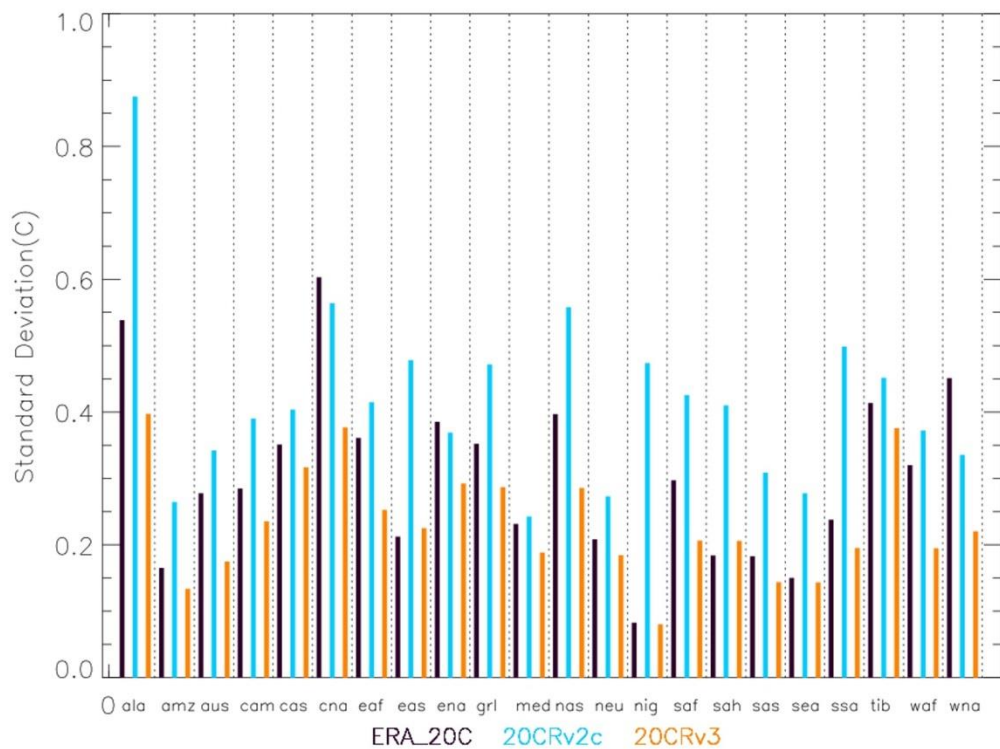


Figure 4.14..For each of the 22 Giorgi regions, the bars summarize the standard deviation of timeseries shown in Figures 4.11 through 4.13 for the three reanalysis products from 1851(Top Panel.) The 20CRv3 product exhibits the lowest standard deviation for almost all regions . The bottom panel is the same analysis for a reduced period from 1900 to concur with the start of ERA-20C

4.5.3 Comparison between densely and sparsely sampled regions

The Giorgi region analysis in Section 4.5.2 highlighted the fact that the vast majority of available monthly-mean station records in the ISTI databank sample only a small percentage of the globe. This is of particular concern because global mean surface temperatures are calculated by area-weighting regional temperature records from a combination of land and marine sources. The influence of an individual station in a sparsely sampled area thus far exceeds the influence of individual stations in richly sampled areas in the calculation of the global mean (Cowtan et al., 2018). It is therefore of great importance that high-quality homogenisation of the stations in the sparsely sampled regions is undertaken. Analyses in the two preceding sub-sections imply that reanalysis-based approaches may have advantages here.

To investigate this further, we have randomly selected 100 stations from those Giorgi regions that can be considered sparsely sampled and 100 stations from those regions that can be considered densely sampled for comparison. The mean distance between the selected stations and their neighbours in richly sampled regions is 79.6 km with a standard deviation of 37.6 km. For poorly sampled regions, the mean distance between a station for homogenisation and its neighbours is 567 km with a standard deviation of 356 km. Prior work has shown that inter-station correlation decreases roughly exponentially with distance with correlation halved on monthly timescales typically within 500 km distance (Hansen and Lebedeff, 1987, New et al., 1999).

We quantify how 20CRv3, which the Giorgi region analysis highlighted constituted the best potential reanalysis product for this task, compares to neighbour-based approaches based upon network sparsity (Figure 4.15). Given that the station series is the basic ISTI databank monthly-mean data without having had homogenisation or quality control applied, some proportion of this spread will inevitably arise from data issues in the candidate and/or neighbour series. Using the median neighbour distance to indicate network sparsity, there is far less of a marked drop in correlation/increase in standard deviation when using 20CRv3 as the estimator than when using neighbours. In dense regions, it is clear that neighbour-based approaches will tend to have more power (higher correlations, lower standard deviations). Conversely, in sparse regions, the 20CRv3 estimates likely have more power. The cross-over

between the two occurs somewhere around the 600-800km distance to the median neighbour. Furthermore, there is a much reduced gradient in these diagnostics with network density when using 20CRv3, implying that any analysis is likely to be more globally homogeneous in its application using 20CRv3, even if this came at the expense of reduced breakpoint detection power in data dense regions of the globe.

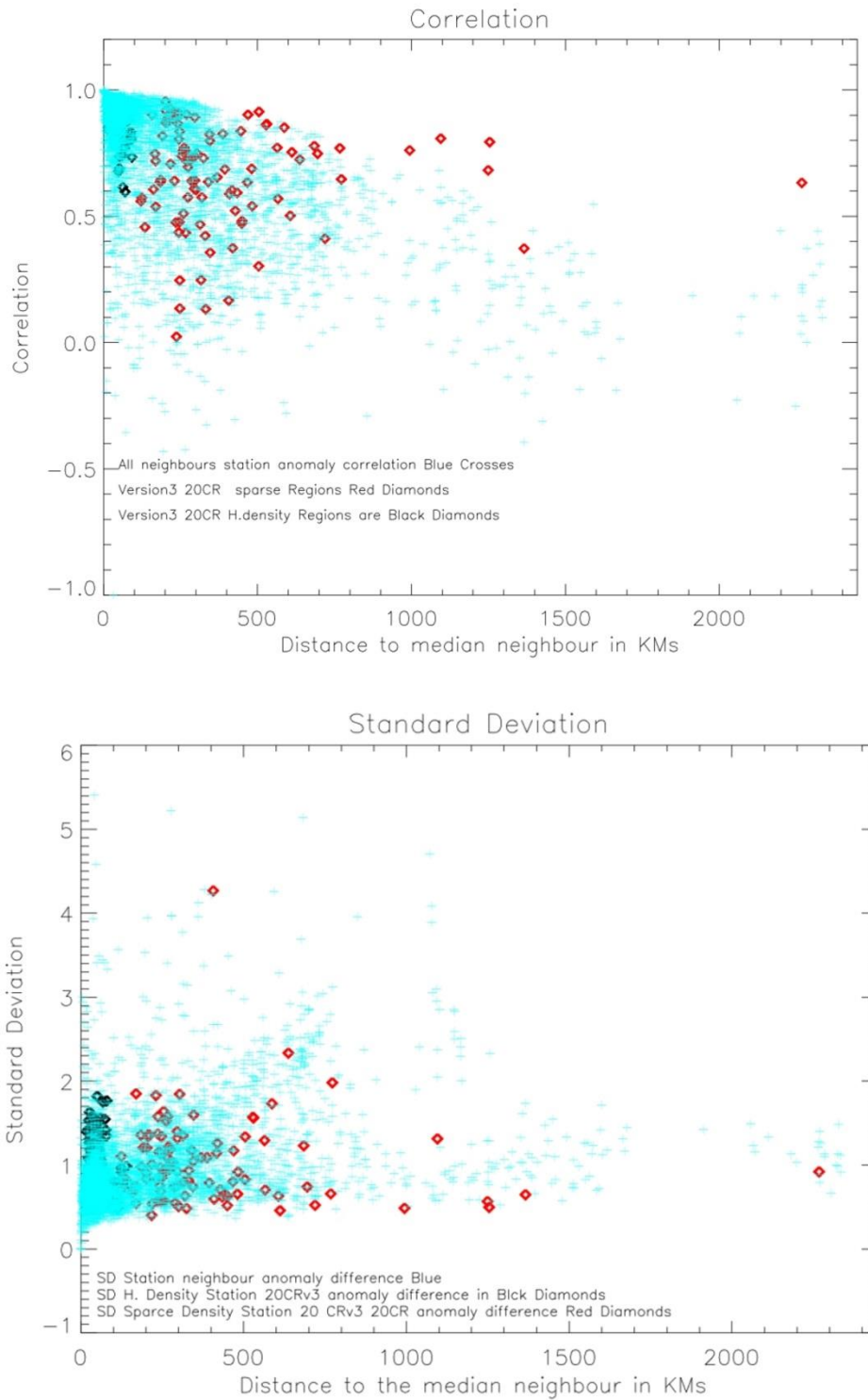


Figure 4.15 Top Panel is a pooled comparison of the correlations (r) (Blue x's) between each station and its 25 nearest neighbours across both for the 100 densely-sampled and sparsely-sampled stations (200 times 25 independent values). Overplotted are correlations between the 20CRv3 product and the candidate series (Red Diamonds). These are each displaced in the x-axis by the distance to the median neighbour such that for stations in densely sampled regions the reanalysis is closer to $X=0$ and for progressively sparser station locations the reanalysis estimate is further displaced from $X=0$. The bottom panel is the same comparison as in the top panel, but for the standard deviation of the difference series. Neighbour-based pairwise comparisons are likely better when the distance from a candidate station to its neighbours is less than 350km and, conversely 20CRv3 reanalysis performs better when the distances are c. 700km or greater.

Another way to look at this issue is by splitting into densely and sparsely sampled stations and comparing the correlation between the series (Figure 4.16) and standard deviation of the difference series (Figure 4.17). Assuming that the median neighbour is indicative of the likely overall typical performance of neighbour comparisons, it is evident for many sparsely located stations that the correlation between the candidate station anomalies and 20CRv3 product anomalies are comparable to or higher than the correlation with the median neighbours' anomalies. The standard deviation of the difference series of the anomalies is also often lower. Conversely, in well-sampled regions, the correlation with the median neighbour series is almost always higher and the standard deviation almost always lower. The variation in absolute performance between well-sampled and sparsely-sampled regions is smaller for 20CRv3 implying the potential for lower regional variations in performance in subsequent homogenisation applications.

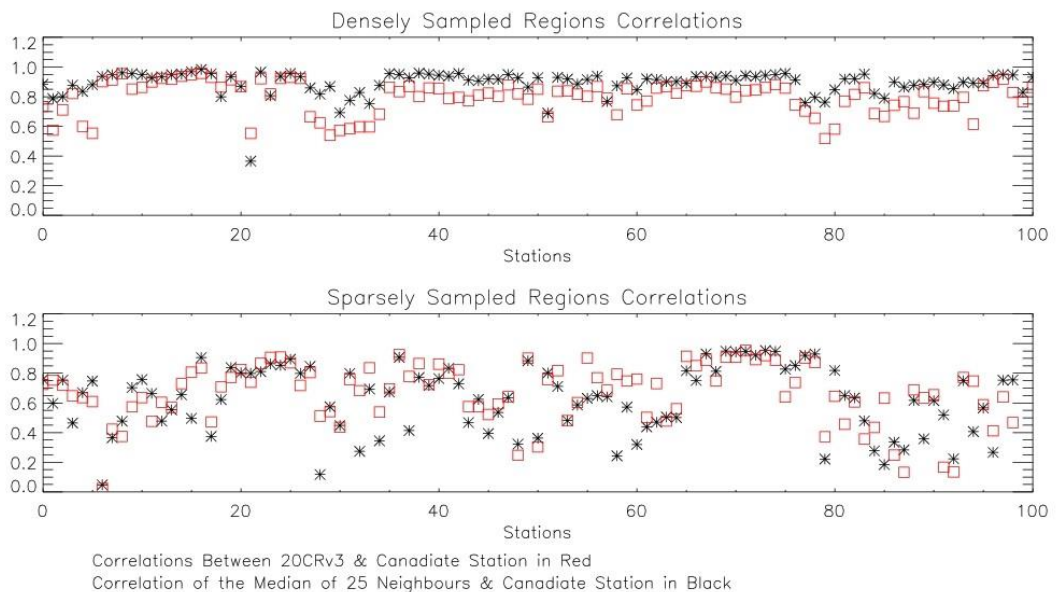


Figure 4.16 Correlations between the candidate station anomalies and the median neighbour compared to the correlation between the candidate station anomalies and that of 20CRv3 in well sampled (top panel) and sparsely sampled (bottom panel) regions. The median neighbour value is denoted as a black star and the 20CRv3 value as a red box. Higher values denote better agreement and thus greater suitability as a reference series to remove common climatic events to perform relative homogenisation.

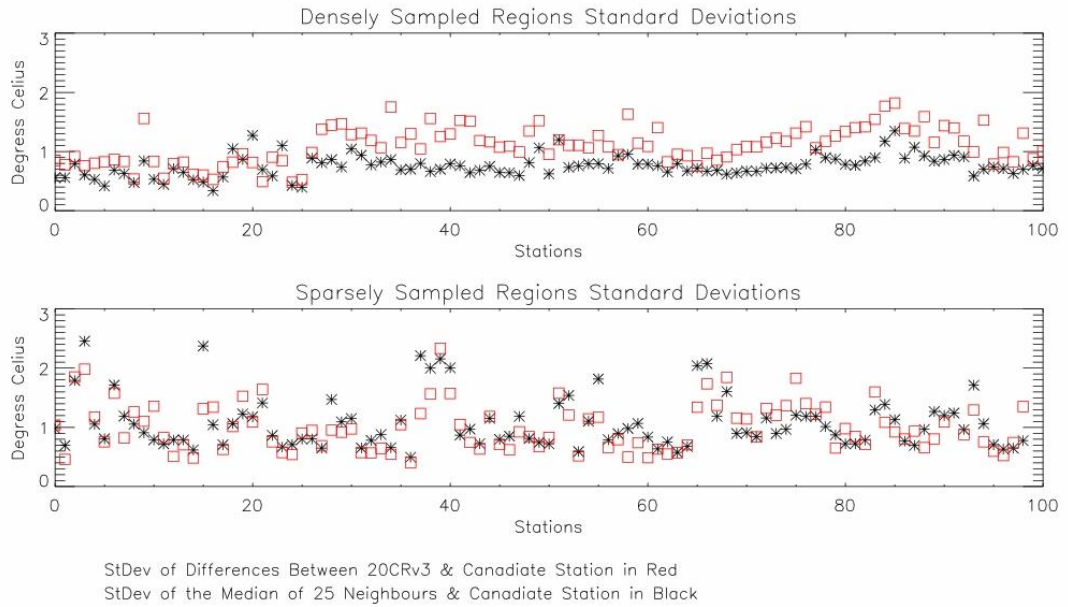


Figure 4.17. As Figure 4.16 but now considering the standard deviation of the difference series. Lower values would, all else being equal, lead to smaller breaks being able to be detected and reliably adjusted in the candidate series.

4.6 Discussion

The global land surface records of monthly mean temperature are not evenly distributed in either space or time. Much of the ISTI databank of raw data is made up of short term records, with the majority of stations not extending back before the 1950s. This uneven spatio-temporal distribution creates challenges, not least in the homogenization of the longest records. This is particularly acute in sparsely sampled regions. All current state-of-the-art homogenization techniques use some form of a neighbour-based approach. However, these approaches work best in densely sampled regions and for periods where a sufficient number physically correlated series are available as comparators. Hence, neighbour-based approaches will work best in the recent past and in areas such as Europe, North America, and Japan where a high-density network of meteorological measurement stations is available.

Herein we have shown that perhaps for the first time, the most recent generation of sparse input reanalysis products, represented by the NOAA-CIRES-DOE 20CRv3 data set, likely has broadly comparable power to neighbour-based approaches based upon individual station comparisons and regionally aggregated characteristics. The 20CRv3 achieves this while being independent of observational temperature records

over land. This independence will be an aid in cases where changes that are broadly consistent in nature apply quasi-contemporaneously across broad regions. An example of such a change is the transition from cotton region shelters to maximum-minimum temperature systems (MMTS) sensors across the US cooperative observer network that occurred over a decade or so in the late 1980s to early 1990s (Quayle et al., 1991). However, the lack of direct use of surface temperatures observations means that care is needed to firstly ascertain the quality of the sparse-input reanalysis data. This analysis, building upon precursor analyses (Compo et al., 2013, Simmons et al., 2017, Parker, 2016, Zhou et al., 2018), provides the evidence basis that the quality of 20CRv3 is likely sufficient.

However, 20CRv3 is a modelled based product and Slivinski et al. (2019) detail the parameters used in 20CRv3 and how they differ from previous versions. These updates include changes to the boundary condition needed to run the model that are taken from evolving sea surface temperatures (SSTs) and sea ice concentrations which themselves contain uncertainties. SSTs are taken from an analysis by Hirahara et al. (2014) with input from the ocean reanalysis 1851-2013 by Giese et al. (2016), both of which come with a level of uncertainty, that is not least as a result of lack of observations in time and space requiring interpolation into regions lacking observations. The daily estimates are interpolated often from monthly values giving rise to additional uncertainty.

Reanalysis product creation is continuous work to produce new products and improve on previous versions (Slivinski et al., 2019, Poli et al., 2016, Laloyaux et al., 2018). While reanalyses are not observations and must not be confused with observations (Parker, 2011) they constitute an increasingly realistic estimation of the state of the atmosphere at any point in time and the ensemble spread of the reanalysis product should represent the error of the ensemble mean (Laloyaux et al., 2018). Reanalysis are therefore perhaps the most accurate and homogenised datasets available (Dee et al., 2011a). and any systematics issues with reanalysis products are reasonable will quantify compared to observations.

A substantial further advantage in the use of these reanalyses spanning more than a century is the availability of a comparator series at each and every month for which a temperature value is available in the target station series. Conversely, Figure 4.18

shows that this is a major challenge for fixed neighbour constellations. For a neighbour set consisting solely of the 25 nearest neighbours, the most frequent number of neighbour observations at any given timestep is zero. The least frequent occurrence is to have all 25 neighbours available.

However, an advantage of neighbour approaches is that multiple independent assessments are possible, meaning that if any single comparison is compromised by a poor comparator series other independent comparisons can rectify the issue. If, on the other hand, the reanalysis contains a systematic artefact, it is harder to identify and remedy. To try to ascertain the risk of this, robust regionally-aggregated analyses were performed. These assume that station issues will become pseudo-random when averaged over a sufficient sample of stations leaving behind an indicator of regional issues in the reanalysis fields. Such comparisons point to issues in the previous generation of reanalysis products, in agreement with prior analyses (e.g. Ferguson and Villarini (2012)), which are much less evident in the newest 20CRv3 product.

It is also possible to use the ensemble products from the reanalyses which give a population of estimates, although questions as to their dispersiveness may remain. Our analysis of the 20CRv2c ensemble indicates that, as expected from Ensemble Kalman Filter theory (Compo et al., 2011) the ensemble mean tends to be a somewhat better estimate of the station series than the individual ensemble members. This is likely to hold for 20CRv3 which has a larger ensemble size that is designed to be more dispersive to more reliably capture the true climate state. This latter ensemble was not available at the time of the 20CRv2c analysis being performed herein.

This has been the first analysis to directly compare the quality of 20CRv3 to earlier generation products for the ability to estimate observed land surface air temperature series. The interpolated-to-station estimates show improved correlations and reduced standard deviations of station minus reanalysis difference series. When aggregated over broad regions, 20CRv3 shows marked improvements in its ability to reproduce regional series behaviour prior to the mid-twentieth century, addressing previously stated concerns (Ferguson and Villarini, 2012).

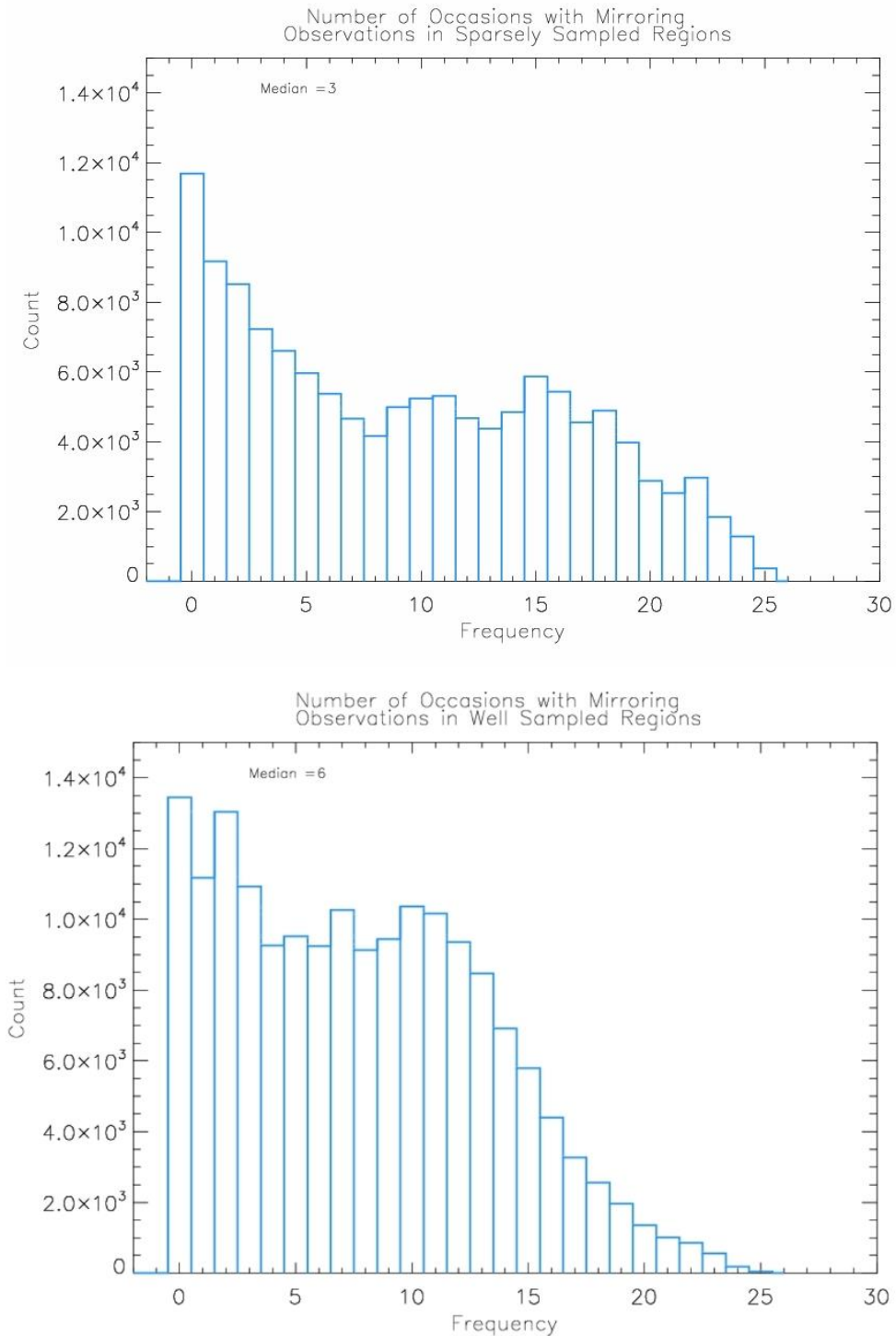


Figure 4.18. Histogram of the occurrence of the frequency of overlap between each station and its 25 nearest neighbours aggregated over the poorly sampled candidate station-neighbour pairs (top panel) and well sampled regions (bottom panel) from 1850 to 2012. The most frequent occurrence is for no overlap for both regions and the median value is 6 out of 25 comparisons being possible at any given timestep in well sampled regions, falling to 3 in poorly sampled regions.

New sparse-input reanalysis products are planned which, as has been the case with full-input reanalysis products (Simmons et al., 2017), will likely yield further substantial improvements. As new generations of sparse-input reanalysis data sets become available, it is thus increasingly probable that they will become an attractive proposition for homogenisation activities of surface air temperatures and potentially other surface meteorological series.

Our analysis points to 20CRv3 becoming potentially advantageous compared to neighbour-based approaches when stations are separated by 700km or more, while neighbour-based approaches are clearly better at separation distances less than 350km. This means that 20CRv3 is preferable for about 700 stations and competitive with neighbour-based approaches for a further c. 3000 out of a total of 28000+ stations retained following the analysis performed in Chapter 3. However, it is not the number of stations that matters, it is their spatial distribution. Figure 4.19 illustrates how those stations where 20CRv3 is likely preferable to or competitive with neighbour-based approaches account for the majority of the global land domain. There is clear potential value in using the 20CRv3 product for homogenisation if the target is a global estimate of changes. Chapter 5 goes on to assess the potential impact upon existing global surface temperature products by applying homogenisation approaches building upon Haimberger et al. (2012) using these series.

Stations differentiated by network density

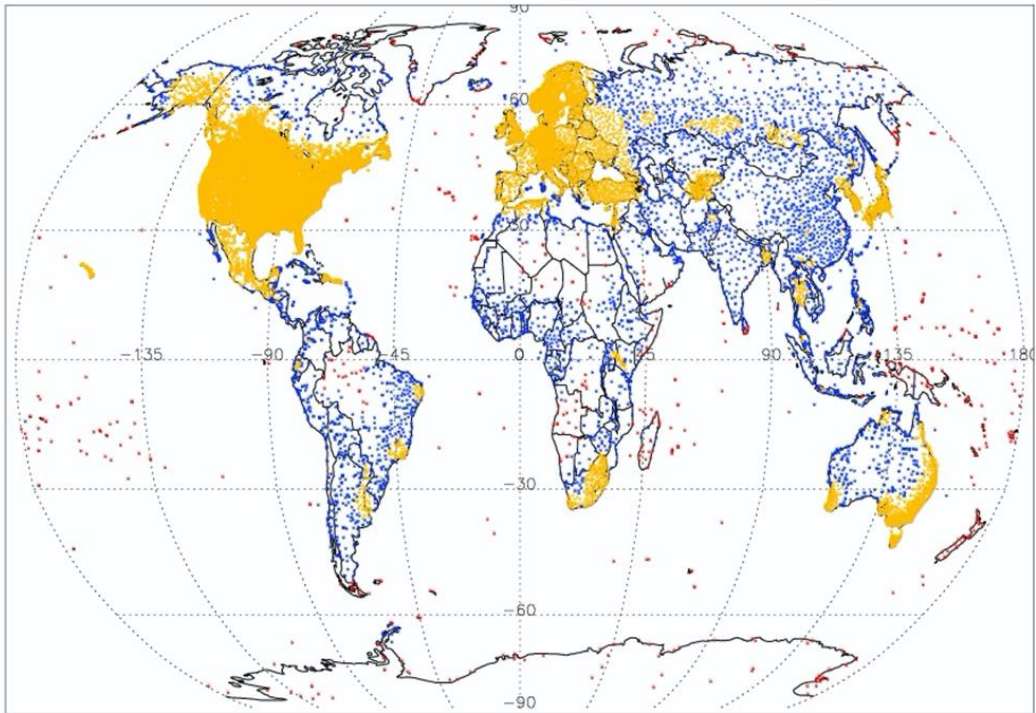


Figure 4.19 Stations with 25 or more stations within 350 km radius (yellow) for which pairwise approaches may be preferable. Stations with the 25 nearest neighbours within 700km (blue) in which pairwise and 20CRv3 based approaches may be of comparable power according to the present analysis. Stations in more data sparse regions (red) which likely will be more amenable to homogenisation using 20CRv3. As successive sparse-input reanalysis products improve over time progressively more points may become blue or red in similar future maps.

4.7 Conclusion

Homogenisation of long-term records of land surface temperature is a challenging proposition. A well-characterised estimate of the true underlying climate signal is required to separate real geophysical effects from non-climatic artefacts. The current state-of-the-art techniques utilise nearby station neighbours. While the majority of stations are in densely sampled regions where such techniques have proven effective, the vast majority of the global land surface is poorly sampled. Sparse-input centennial reanalysis products, which are independent of the surface temperature observations, offer an opportunity to address both this issue and the paucity of early records available as comparator series to assess early instrumental series homogeneity. Once interpolated to the point observation, we find that the best current such reanalysis data set - NOAA-CIRES-DOE 20CRv3 has clear potential value whereas earlier products had substantial potential limitations. Use of sparse-

input reanalysis products would also offer a valuable methodologically distinct approach which would allow improved exploration of structural uncertainty in reconstructions of global land surface air temperatures.

Chapter 5. Application of homogeneity assessments using sparse-input reanalyses fields and comparison to existing approaches

5.1 Introduction

For reasons outlined in detail in Chapter 2, homogenisation of land surface air temperature time series is essential prior to their application to long-term monitoring. In brief, over multiple decades there are inevitable changes in multiple facets of a station series such as the local micro-environment, instrumentation, observers, methods of observation etc. etc. (Vincent, 1998, Karl and Williams, 1987, Quayle et al., 1991, Guttman, 1998, Peterson and Easterling, 1994, Bronnimann, 2015). Even if every effort is made to minimise the impacts of such changes, it is all but inevitable that non-climatic data artefacts shall be present in the record for many stations, and that in very many cases such artefacts will matter for long-term climatic time series analysis (Causinus and Mestre, 2004, Trewin, 2010, Domonkos and Coll, 2017, Hunziker et al., 2017, Lawrimore et al., 2015).

State-of-the-art homogenisation methods generally use a neighbour-based approach, typically based upon pairwise comparisons as described in the benchmarking paper of Venema et al. (2012) to identify and then adjust for breakpoints. All such approaches are predicated upon the availability of sufficiently similar neighbour estimates and an assumption of non-coincidence of these data artefacts (Causinus and Mestre, 2004). Such assumptions cannot be guaranteed, and it is of value to explore alternative approaches which may better maintain independence. One such potential approach is to use sparse-input reanalyses which do not ingest or use land surface air temperature measurements and yet provide dynamically and physically constrained estimates of land surface air temperatures (Trewin, 2010, Compo et al., 2013, Compo et al., 2011).

Chapter 4 has illustrated that the most recent sparse-input reanalysis products are increasingly viable and credible candidates for use as reference series for the homogenisation of land surface air temperature station series. Specifically, it showed that the correlations and standard deviations of the difference series between the stations and sparse-input reanalyses products were similar to those between stations

and their neighbours. It highlighted particular potential benefits in data sparse regions and epochs. The present chapter therefore now goes on to apply the 20CRv3 sparse-input reanalysis product (which Chapter 4 highlighted to be the most appropriate product) to the task of homogenisation, and to compare the results to those from the application of NOAA NCEI's PHA method (Menne et al., 2018) to create GHCNMv4, to the same set of fundamental data holdings (Chapter 3). It goes on to compare globally aggregated results to the full suite of existing Global Land Surface Air Temperature (LSAT) products.

The sparse-input reanalysis homogenisation approach applied starts from the established methods applied to full-input reanalyses to homogenise radiosonde data records by Haimberger et al. (2012). These are modified for the particular circumstances of sparse-input reanalyses and land surface stations. The quality control and breakpoint detection steps are outlined in Section 5.2. Section 5.3 summarises and assesses the overall application of the adjustments. The approach yields 4 estimates of the required adjustments. These estimates all use the same method to detect breakpoints but diverge in how adjustments are then calculated. Section 5.4 assess the efficacy of the resulting homogenisation techniques by assessing adjustment behaviour, station series and gridbox anomalies and trends. GHCNMv4 station series are used as a comparator therein. In Section 5.5 a regional, hemispherical and global analysis is conducted. Section 5.6 compares the results at the global mean aggregation with estimates from other published LSAT products variously used in monitoring and assessment activities. Section 5.7 includes a discussion of limitations, outstanding questions and potential next steps and, finally, Section 5.8 concludes.

5.2 Quality control and breakpoint detection

5.2.1 Removal of gross outliers

Before attempting the identification of breakpoints, it is necessary to remove outliers that may arise from a number of sources of random error such as e.g. inadvertent keying of 15.7 as 25.7, erroneous instrument readings or missing recordings entered as zero rather than a missing data indicator. Quality Control is applied to the difference series between the station series anomalies and the matched 20CRv3

anomalies, both normalised to their common period-of-record. This is the same series as was analysed in Chapter 4. The inter-quartile range of this series was calculated and a QC threshold set at three times this value. All points where the absolute differences are greater than this threshold were set to missing. 15,537 stations had one or more observations quality controlled as detailed in Table 5.1 and Figure 5.1

Percent of Observations Removed	Number of Stations
0.0 % Obs Removed	12102
0.0% > Obs Removed < 0.2%	2775
0.2% ≥ Obs Removed < 0.5%	5296
0.5% ≥ Obs Removed < 1.0%	4153
1.0% ≥ Obs Removed < 5.0%	3133
5.0% ≥ Obs Removed < 10.0%	152
10.0% > Obs Removed	28

Table 5.1 Summary of the frequency of different percentage intervals of observations removed by Quality Control from individual stations.

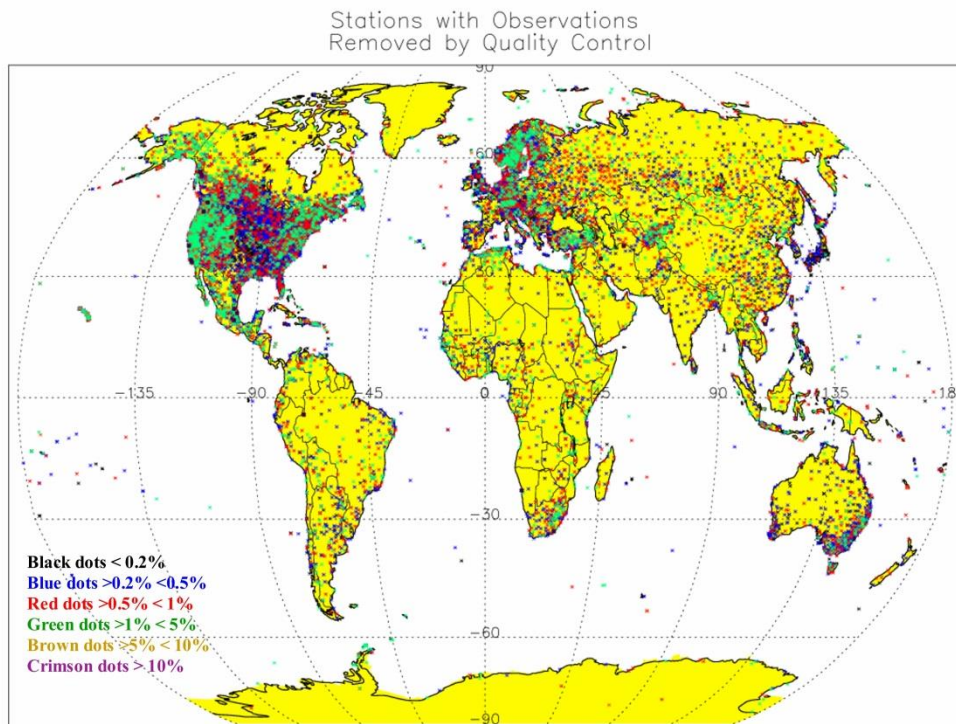


Figure 5.1 Map detailing locations of stations Quality Control using the same intervals as Table 5.1. Stations with more data removed overplot those with fewer data removed.

5.2.2 Breakpoint detection

The present analysis makes use of a variant of the Standard Normal Homogeneity Test (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 method performs well, particularly in well sampled regions, and has consistently scored highly in benchmarking and comparison tests (Venema et al., 2012, Ducre-Robitaille et al., 2003).

The SNHT test was applied to 20CRv3-station difference series where 20CRv3 is interpolated to the station coordinates (Chapter 4). The SNHT can be described mathematically as follows:

$$T_i = \frac{N}{S_i^2} * (\bar{X} - \bar{Z})^2 + (\bar{Y} - \bar{Z})^2 \quad \text{Eqn 5.1}$$

Where:

T_i = Test Statistic (SNHT Score)

N = Total number of observations

S_i = Sigma of valid points before & after the break

\bar{X} = Mean of valid points before the break

\bar{Y} = Mean of valid points after the break

\bar{Z} = Mean of valid points before & after the break

The test is applied iteratively to consecutive segments of the series of equal intervals and progresses one timestep at a time through the series. The test thus cannot be applied to, or detect breaks at, the start or end of the series (Toreti et al., 2011). In the present analysis the segment length either side of the tested point is set to 5 years 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 of X and Y contains the same number of points and seasonal sampling is identical. The test is only applied if more than 20 points remain in both X and Y after this sample matching has been applied. Otherwise, the test statistic is set as missing.

Long gaps in the station series would *a priori* increase the chance of a break in the statistical properties of the series. Such long gaps, unless resulting solely from poor records management, will typically be associated with a change of one or more of: station observer, station instrumentation, and even station location. 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 in all such cases. This is necessary because, as applied, the SNHT test is unable to elucidate upon the presence of a break across such a substantive gap in the data record.

The SNHT test is a t-test family test looking primarily for mean shifts. Unlike a standard t-test, it has no recognised significance assessment criteria. Calculating the test statistic is not difficult, but deciding what threshold to assign as a critical value is more challenging. This can be informed by the nature of the resulting test statistic time series, but ultimately it is a matter of judgement. Past analysts have tended to choose values between 8 and 18 for their particular applications (Tuomenvirta, 2002, Klingbjer and Moberg, 2003).

In the present analysis, the SNHT 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 to be employed. Firstly, the histogram of all SNHT scores exceeding a critical threshold of at least 6 was examined (Figure 5.2). If there were an obvious breakpoint in the behaviour of these scores then that would provide a rationale for selection of a threshold. Examination of the histogram fails to indicate any such behaviour to aid in the selection of the critical value. Rather it shows a highly skewed distribution of values that varies smoothly.

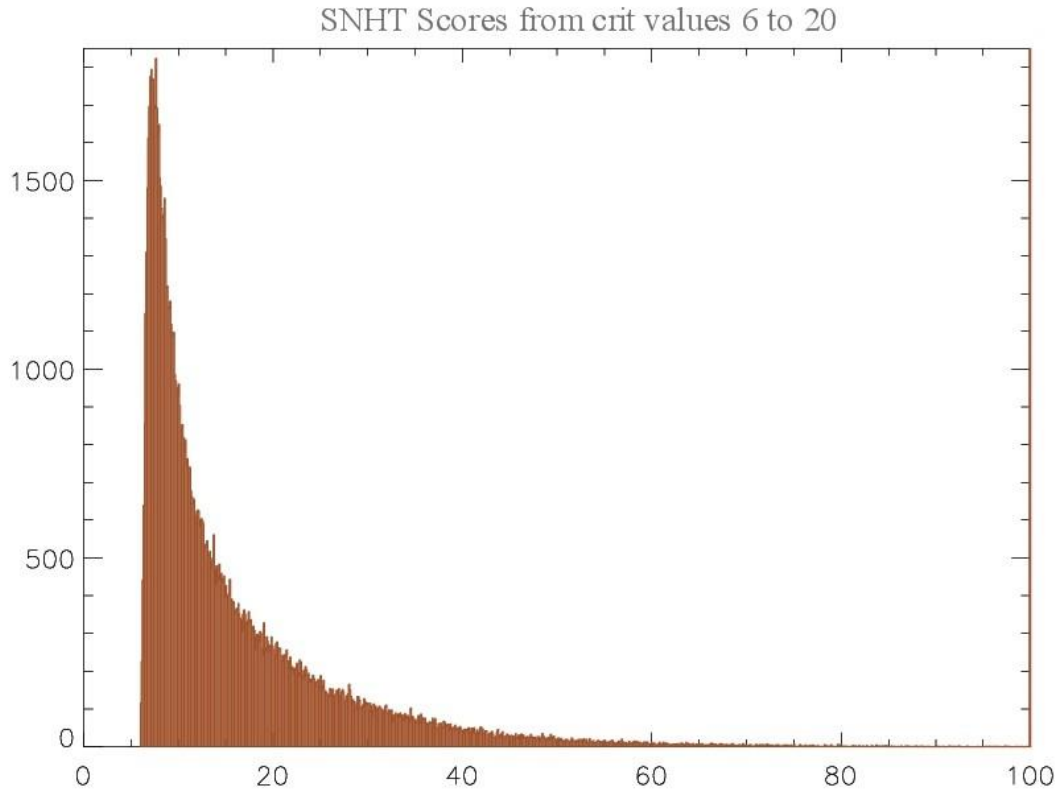


Figure 5.2 Examination of the SNHT scores for critical values from 6 to 100 producing a highly skewed smooth distribution that provides no clear rationale for the selection of a critical SNHT score value. The spike at almost 100 relates to breakpoints assigned to account for timeseries cessation and resumption over a period of >36 months duration which are given a value of 99.9 to force a breakpoint to be assigned.

Secondly, the distribution of implied segment adjustments (Figure 5.3) and cumulative adjustments (Figure 5.4) using the average of the station minus 20CRv3 difference series for each segment were examined. For the individual break segments (Figure 5.3) the mean becomes increasingly non-zero and the standard deviation greater as the critical value increases. The accumulated breaks series that is the total number of breaks detected for each critical value from a critical value from 6 to 20 at intervals of 2 (Figure 5.4) shows no obvious behaviour with the choice of critical value threshold for the SNHT test. As expected from Figure 5.2, the total number of breakpoints adjusted drops with each incremental increase of the critical value. A frequency of the order found for critical thresholds of 12-16, equivalent to a break every 15-20 years, would be consistent with the typical frequency of breakpoints reported in prior studies such as GHCNMv4 (Menne et al., 2018). However, benchmarking exercises (Williams et al., 2012, Venema et al., 2012) highlight a propensity for all present algorithms to underestimate the number of breakpoints in a

broad array of synthetically produced test series such that it is reasonable to assume that this behaviour also extends to the real-world.

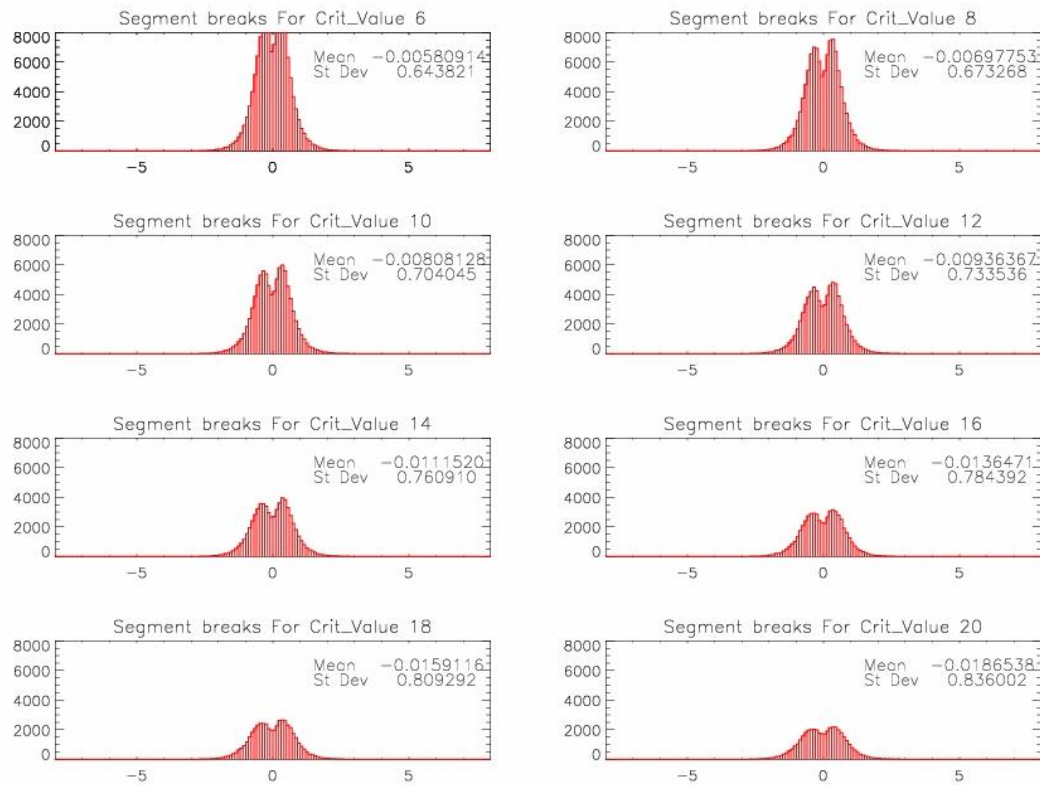


Figure 5.3 Histograms of breakpoint sizes inferred from the station minus 20CRv3 difference series at each breakpoint identified for different SNHT critical values (panels). Within each panel, the mean calculated adjustment and the standard deviation of the distribution are shown.

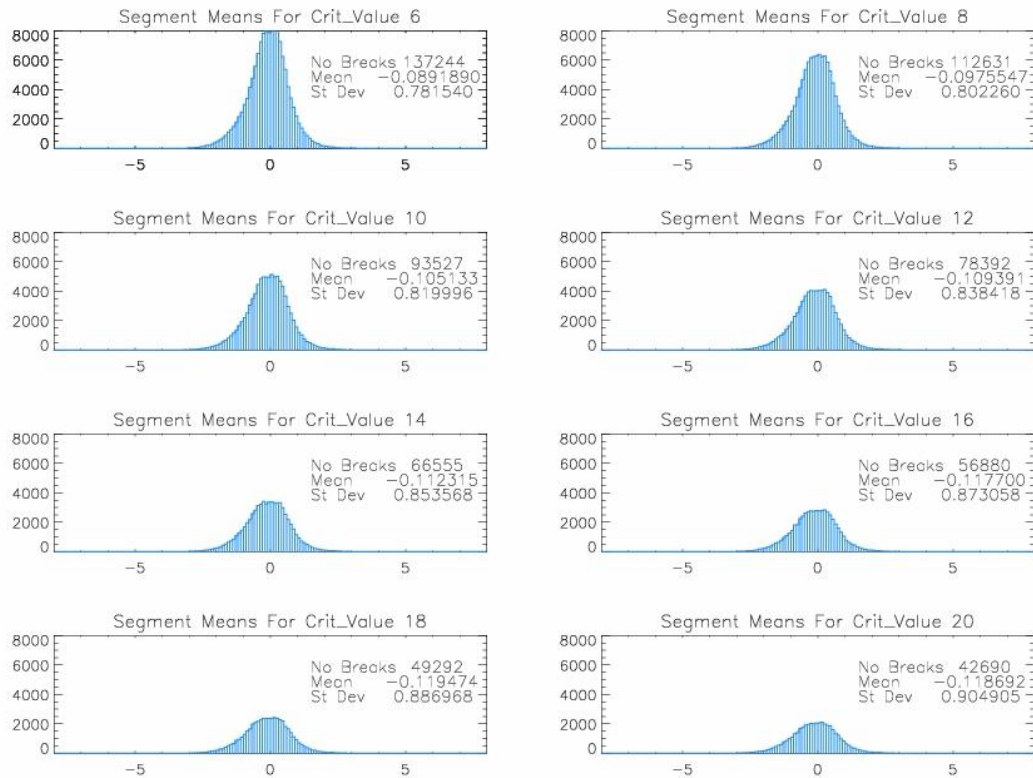


Figure 5.4 Analysis of the cumulative segment adjustments for each critical value from 6 to 20 in intervals of 2. Shown in-line within each panel is the total number of breaks identified, the mean cumulative adjustment and the standard deviation.

Given the lack of a robust basis from the data to select a critical value, recourse was made to randomly generated series containing no breaks. A hundred thousand runs of random numbers were created, equivalent to contiguous series availability over 1850 to 2014 at monthly resolution. A series mean, sigma and AR(1) similar to the values found in the difference series between the station anomalies and 20CRv3 interpolated anomalies in the initial 29 case study stations discussed in Chapter 4 was used to seed these series. All series contain no artificial breakpoints and so, for a perfect breakpoint detection test statistic, would not return any breakpoints. The maximum SNHT score attained was assessed against the autocorrelation and standard deviation of the seeded series (Figure 5.5). The analysis clearly rules out low values of the SNHT score below 12 as every single synthetic series would return one or more breakpoints for such low values. But it does not greatly help further beyond confirming that time series with higher standard deviations and, in particular, higher autocorrelation will yield higher false positive rates which has long been recognised (Karl and Williams, 1987, DeGaetano, 2005, Caussinus and Mestre, 2004). This analysis highlights that no critical threshold is likely to be optimal for all

stations as these properties may reasonably be expected to vary spatially and possibly through time.

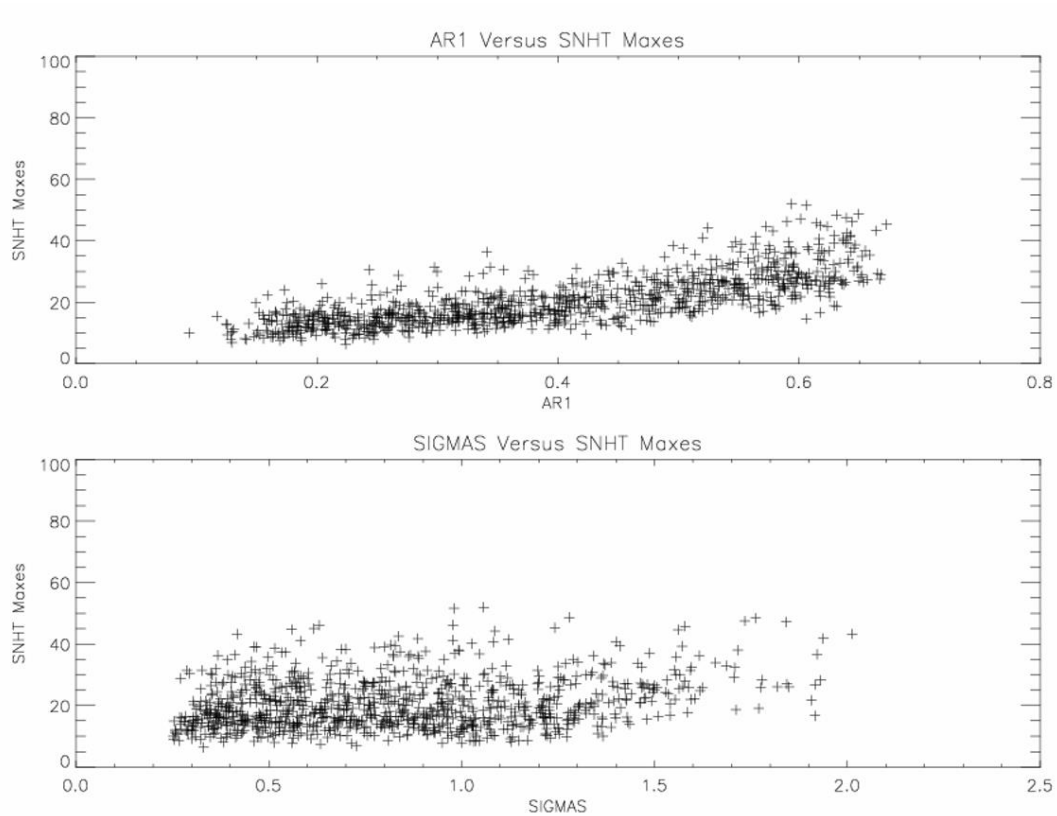


Figure 5.5 Analysis of the application of the SNHT test to 100,000 simulations of homogeneous time series with similar statistical properties to the difference series for the case study stations used in Chapter 4. The maximum SNHT score has been retained from each series and is plotted against the autocorrelation of the synthetic series (top panel) and the sigma of the synthetic series (lower panel).

In summary, there is substantive evidence that rules out values lower than 12 for the SNHT score for the present application. The value of 16 is relatively close in terms of intervals of occurrence of identified breaks in the station series with prior assessments using neighbour-based approaches, so in the absence of other criteria, a critical value of 16 was selected for use in the present analysis.

5.2.3 Break Assignment

Breaks are only assigned if three or more consecutive values exceed the critical value threshold. Breakpoints are associated with the timing of the maximum test statistic value attained within such a contiguous string. To further minimise 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.

Examples of the results of the application of the algorithm are given for three selected stations in Figures 5.6, 5.7, and 5.8 selected from data sparse regions where the algorithm may have most value over traditional approaches (Chapter 4). Also included in Figure 5.9 as a typical example from a well sampled region is the De Bilt station used as a case study in chapter 4.

Starting with a simple example, Figure 5.6 illustrates a case with a single detected break in the series which is visually obvious in the difference series and associated with a large exceedance of the critical threshold for the SNHT test. The SNHT test assigned breakpoint is in good accordance with the apparent break location from simple visual inspection of the series.

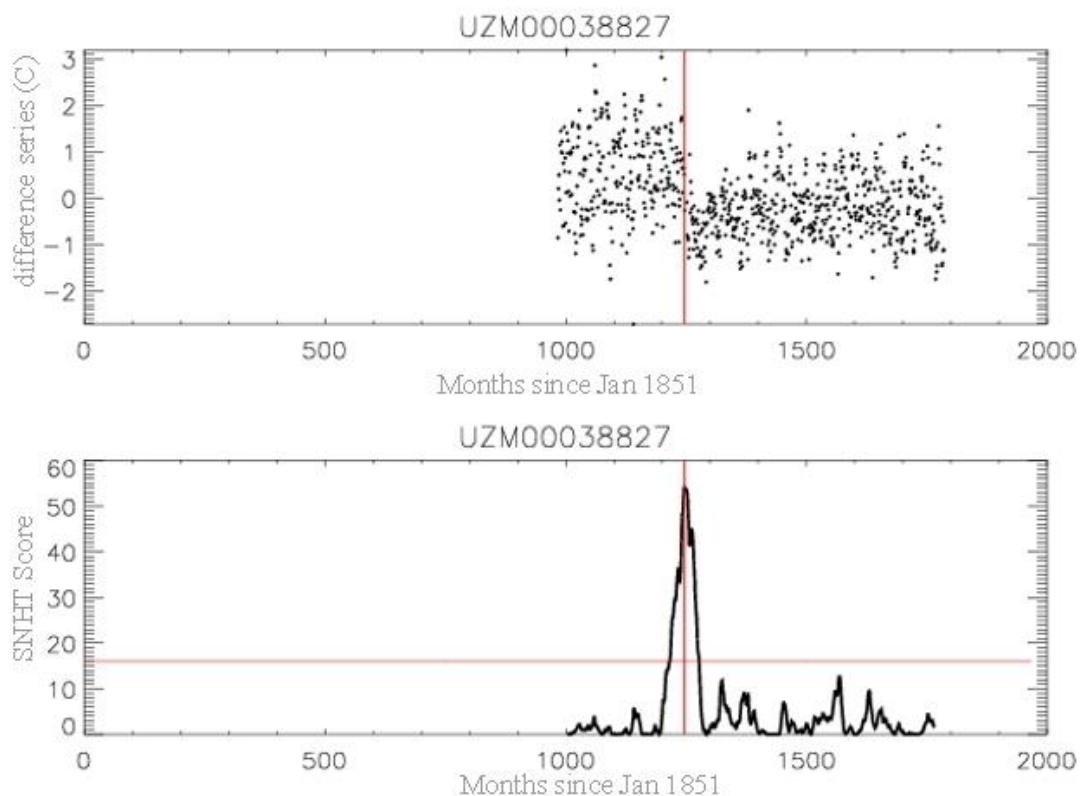


Figure 5.6 Station at Baisun from Uzbekistan (38.2° N, 67.2° E, 1241 m.a.s.l) with 803 observations over December 1932 until October 1999 with a single break assigned in November 1954. The top panel shows the station minus 20CRv3 difference series where each monthly value is plotted as a dot. The lower 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.

Figure 5.7 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 often somewhat less clear, but some of the breaks are visually obvious and, again, the assigned locations seem reasonable in these cases.

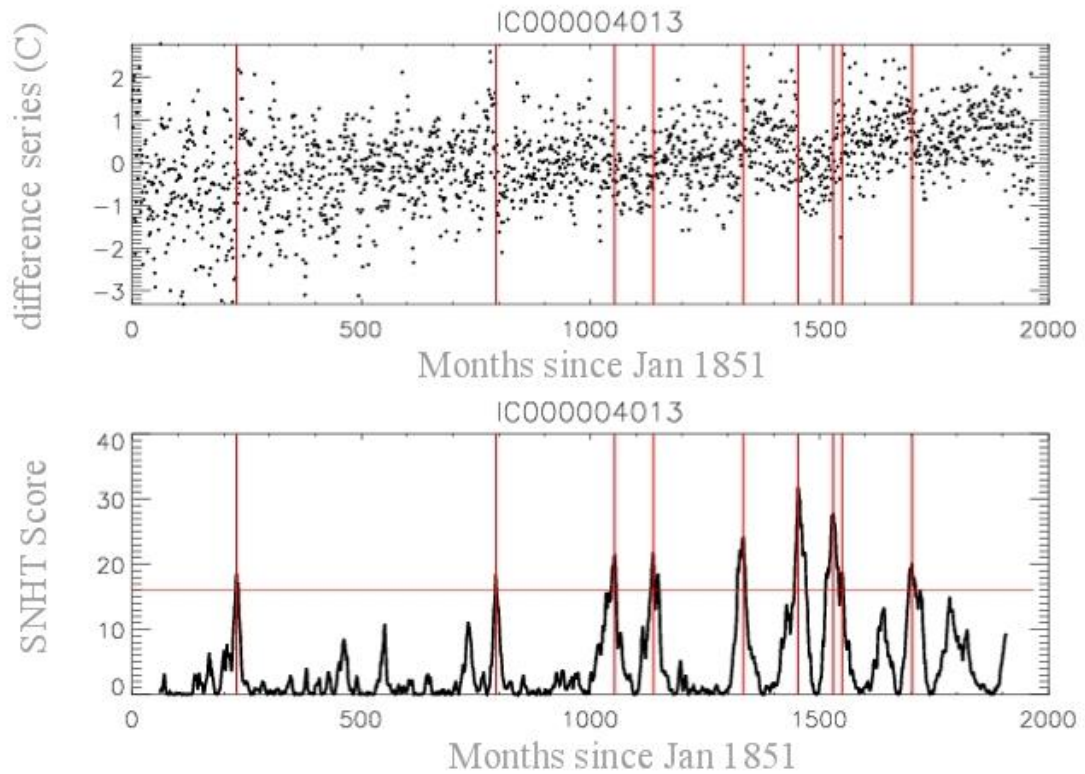


Figure 5.7. As Figure 5.6 but for Stykkisholmur, western Iceland (65.073°N , 22.725°W , 15 m.a.s l). The site has 2051 observations, commencing before January 1851. The site has minor data gaps from August to December 1921 and between December 1940 and April 1941 that are not clearly visible. Breakpoints were detected in December 1869, January 1917, September 1938, September 1945, February 1962, February 1972, July 1978, February 1980 and November 1992

Figure 5.8 shows a case with a break in the series availability, in this case owing to the second world war, and highlights how the forced insertion of a breakpoint upon resumption ensures an adjustment will be estimated in the case of such a cessation and resumption of operations. In this case, visually there is a potential minor discontinuity between the series before/after the world war. Again, breakpoints within the series appear to be reasonably well detected for this site.

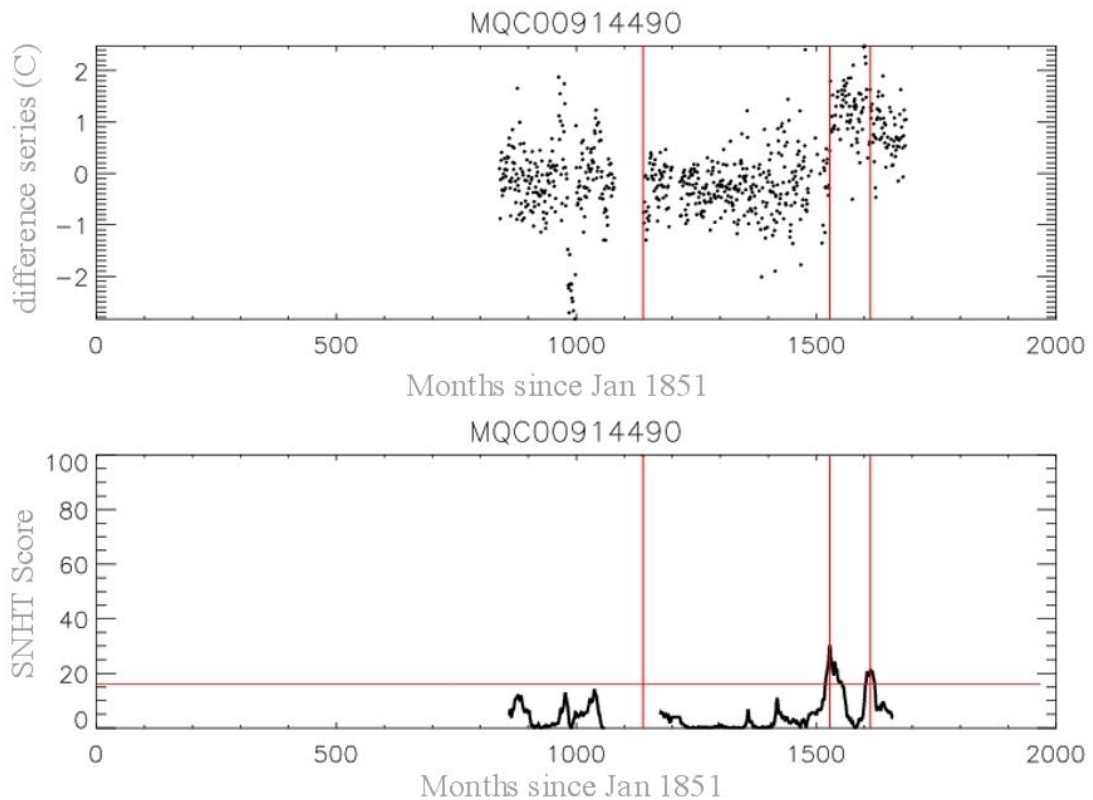


Figure 5.8 As Figure 5.6 but for a site on Midway Sand Island at Midway in the Pacific Ocean (28.217°N, 177.35°E, 3 m.a.s.l.). The Site has 701 monthly observations commencing in December 1920 until August 1991. Note the gap in the observations from December 1940 to December 1945 (month 1080 to 1140) 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.

Figure 5.9 shows the De Bilt series in the Netherlands, as previously discussed and analysed in Chapter 4 (see Chapter 4 for a more in-depth discussion of this series). There are two detected breaks over the 1895 - 1905 period, in concordance with the visual interpretation previously undertaken in Figure 4.8 and available metadata. 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), although the SNHT test is only able to detect the mean shift aspect in such cases by design. The breakpoint detection method also indicated an additional break around 1984.

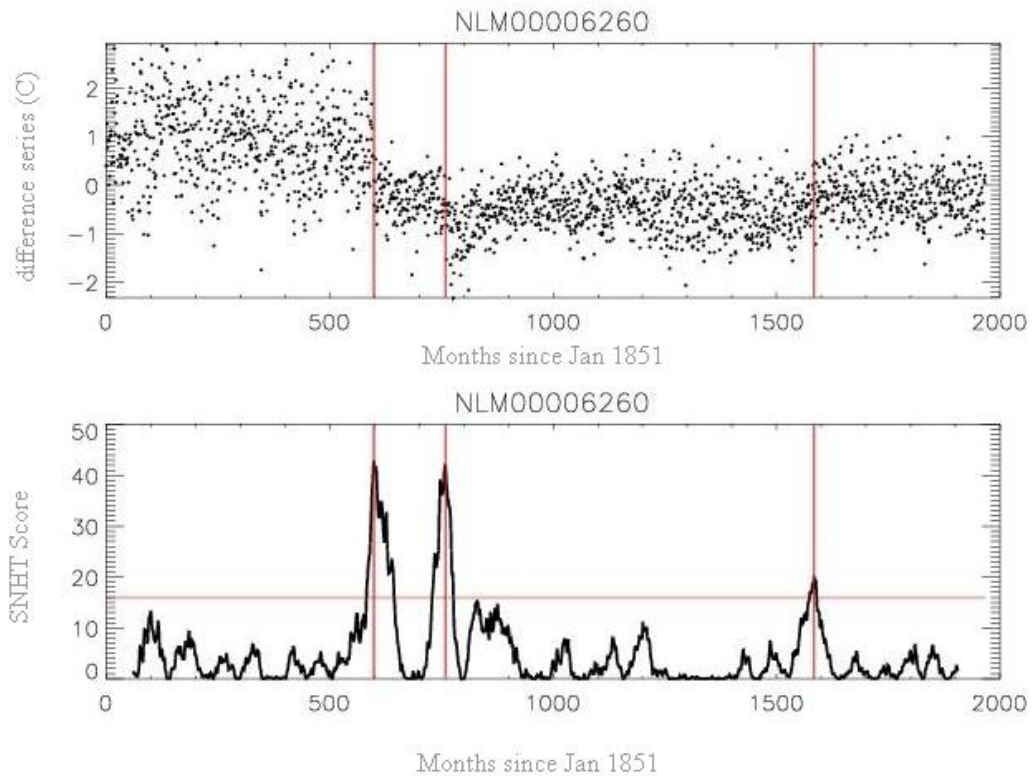


Figure 5.9. As Figure 5.6 but for De Bilt (52.1014°N, 5.1867°E 2 m.a.s.l). The De Bilt site is continuous before and after 1851 to 2014, the period under examination for this work

Visual inspection of very many additional series confirms the results shown in these station examples discussed here. At an SNHT score threshold of 16 and with the stipulations over data-breaks and near coincident breakpoints employed, the SNHT algorithm appears to detect obvious breaks in the series and does not have a propensity to overtly over-estimate the occurrence of breaks.

5.3 Application of adjustments

Adjustments are applied to all points preceding each identified breakpoint. They are applied progressively back through the series such that the resulting series mean state, if the adjustments are adequate, should end up as 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 exist in many series as is visually obvious from e.g. Figures 5.8 and 5.9 where in both cases

the earliest segment shows higher variance. 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 herein to create 4 distinct adjustment estimates: 2 based upon the RAOBCORE approach and 2 based upon the RICH approach.

5.3.1 20CRv3_{long}

20CRv3_{long} is based upon the RAOBCORE methodological approach, although uses the reanalysis estimate rather than the background forecast. In the RAOBCORE case, the target data is assimilated such that use of the reanalysed field would introduce an overt circularity. In 20CRv3 the surface temperatures are not assimilated and thus the use of the reanalysis estimate directly is still independent. In fact arguably because it never assimilates the temperature observations from the station it is more independent than the RAOBCORE approach where the background contains some residual information from prior observations from the site. The same difference series of the neighbour anomalies minus 20CRv3 anomalies as was used to determine the break locations is used to estimate segment mean adjustments that need to be applied. The full segment series irrespective of length is used to estimate the means prior to and after the break to infer the required adjustment. The difference in segment means (after-prior) is sequentially added backwards to all valid data points prior to the current breakpoint in the series. The whole process is applied backwards from the final identified breakpoint to all identified breakpoints such that, if the adjustment estimates are well-defined, the entire series should be homogeneous with the final segment permitting modern measurements to be directly compared to older series.

5.3.2 20CRv3_{short}

20CRv3_{short} is identical to 20CRv3_{long} except that segments are cut if they are longer than 5 years. Thus only the (up to) 5 years either side of the breakpoint is used to infer the required adjustment. If there is long-term drift in the 20CRv3 reanalysis product at the station location then this minimises the impact that this can have on

the returned adjustment estimates and the resulting homogenised series. However, use of short segments may introduce noise to 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 resulting in somewhat noisier adjustment estimates being applied.

5.3.3 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 defined by a great circle distance search. This distance search is performed irrespective of neighbour directions, so in some cases may preferentially sample from some quadrants. Each neighbour that contains sufficient data within +/-5 years of the breakpoint and itself does not have identified breakpoints within 20 months either side of the breakpoint is used to create an adjustment estimate. The difference series between the target station anomalies and the neighbour series anomalies (both calculated relative to their own station series availability to maximise station retention) is used as the basis for this estimate (again using the difference in means after-prior). There must be at least twenty points prior to and after the break to calculate an estimate of the required adjustment. Differences in station data availability lead to slight station-to-station climatology differences which are assumed sufficiently small as to be unimportant in this context and when averaged over a sufficient sample to become pseudo-random in nature.

Assuming one or more neighbour-based estimates are returned the median of the individual estimates is applied as the adjustment. This minimises 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 application of a 20CRv3 based estimate is preferable to no adjustment being applied. This occurs for 5154 breakpoints across a total of 4986 stations across the full suite of stations retained following the analysis described in Chapter 3. The overall proportion of deferral to 20CRv3_{short} is just under 10% of all identified breaks (c.f. Figure 5.4 SNHT critical value 16 panel). Figure 5.10 illustrates that the propensity of deferral to 20CRv3 generally increases back in time as the availability

of neighbours decreases. Figure 5.11 shows maps of stations that deferred at least once to 20CRv3 to homogenise a break and this produces some surprises in so far that particularly in North America post 1950 (lower right map) 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 (discussed further in Chapter 2) whereby in several cases no suitable neighbours that were themselves homogeneous presumably remained.

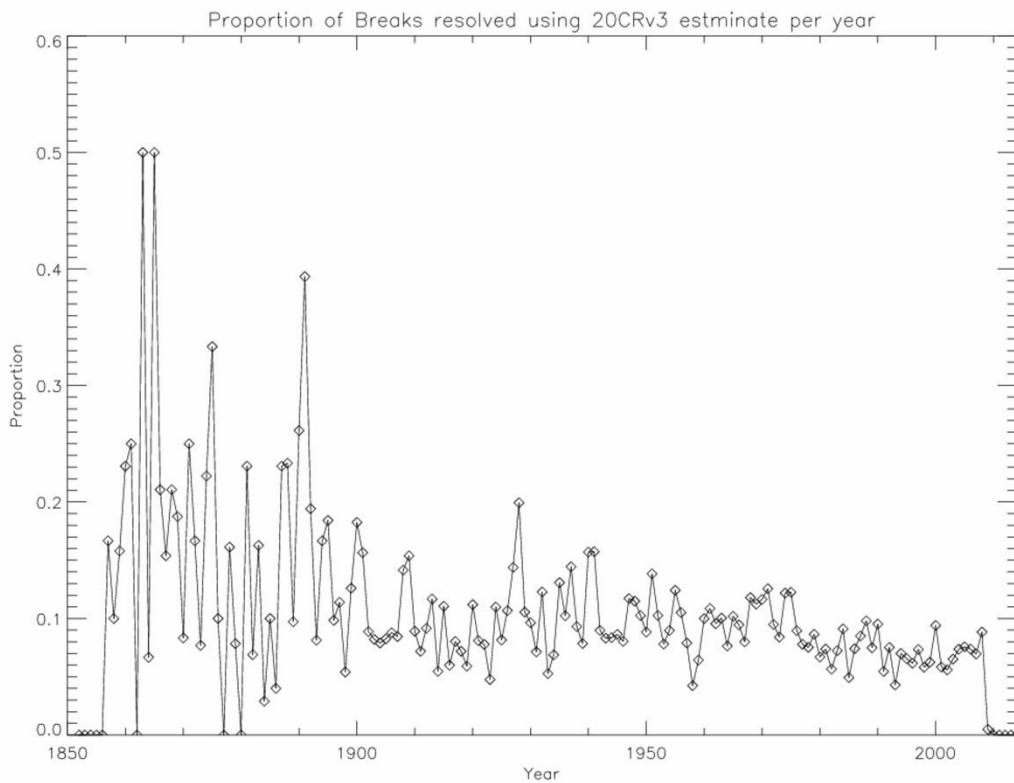


Figure 5.10 Proportion of deferral to 20CRv3 for homogenisation with time as a proportion of total break counts in each given year (which increases substantially as the station density increases after 1950). Note that no breakpoints are detected and hence no adjustments applied in the first and last 5 years of the series.

5.3.4 Neighbour_{double-diffs}

Neighbour_{double-diffs} is broadly based upon the RICH_{tau} method of Haimberger et al. (2012). It differs from 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 quite widely in Numerical

Weather Prediction (Saha et al., 2010, Kanamitsu et al., 1991). The assumption is that although the model being used (NWP or, in this case, 20CRv3) may be biased that this bias varies smoothly. Taking the difference of the differences between reasonably proximal locations removes the common bias component in the reference model (in this case 20CRv3) leaving a purer estimate of the true instrumental differences assuming that the model reasonably approximates true geophysical gradients.

Other than the series being applied all other details of neighbour_{double-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.

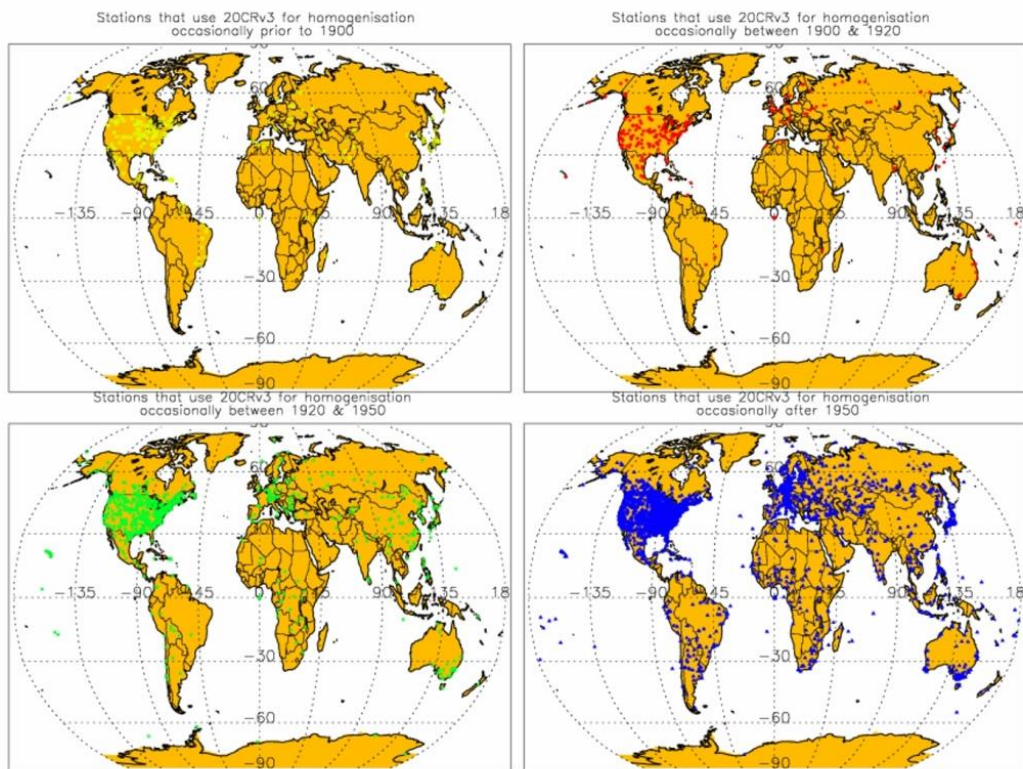


Fig 5.11 Map showing stations the deferred to 20CRv3_{short} for homogenisation using the two neighbour-based approaches before 1900, (top left), between 1900 and 1920, (top right), between 1920 and 1950, (bottom left) and after 1950, (bottom right) for the homogenisation of at least one identified breakpoint.

5.3.5 Adjusting Climatology to 1961-1990

Station time series up to and including the application of adjustments have been used either as actuals or as anomalies relative to their own data availability. This serves to maximise the station count that can be used. However, to further analyse the long-term series and perform aggregations to regional and global series requires the application of a consistent climatology. Standard climate normals are averages generally compiled over thirty years or longer. A thirty-year period is long enough to smooth out year to year variation that inevitably occurs. But conclusions are sensitive to the choice of reference period selected and use of different reference periods may produce substantially different interpretations (Hawkins and Sutton, 2016).

The method employed herein is to calculate a climatology based on 1961- 1990. However, only approximately 11,000 stations 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 1961 to 1990. To incorporate as many stations as possible, 20CRv3 temporally complete station-equivalent timeseries estimates for the full period of 1851 to 2014 were calculated for each station location. Where any station did not have enough data present in the 1961-1990 period to calculate the climatology directly, a proxy climatology adjustment was calculated from the full 20CRv3 series. The 1961-1990 climatology and the climatology matched to the station availability were calculated for 20CRv3. The difference between these two estimates was then subtracted from the station anomalies to normalise the homogenised 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. This is returned to in Section 5.7.

5.4 Assessing the efficacy of the approach

The results of the application of the breakpoint detection and adjustments steps are shown and evaluated herein. In the absence of a benchmarking exercise, this revolves around assessing with increasing fidelity the possible adequacy and caveats around the adopted homogenisation approaches. The section starts in Section 5.4.1 by considering the breakpoint detection and adjustments applied across the full suite of stations to assess whether any of the approaches appear anomalous. Next, resulting time series for selected individual stations are illustrated in Section 5.4.2 to consider whether the homogenised station series are reasonable. Following this, consideration is given as to the spatial patterns of individual monthly anomalies in Section 5.4.3, whereby any obvious issues should become apparent. Finally, in Section 5.4.4 the influence upon gridded trends – where the cumulative impact of adjustments will be most readily apparent - is assessed. This assessment of efficacy is aided by recourse to GHCNMv4 adjusted series as an independent comparator series in Sections 5.4.2 through 5.4.4. In all cases in this section GHCNMv4 homogenised station series have been aggregated identically to the series which have been homogenised herein to allow direct comparability. Section 5.4.5 briefly summarises.

5.4.1 Assessment of adjustments

Applying the SNHT with a critical value 16 leads to a total of 58,325 breakpoints being identified in 19,241 stations in total. Of the 27,639 files analysed, 8,398 stations had no breaks and 1,909 stations had breaks only associated with missing data of greater than 36 continuous months. Table 5.2 shows that the majority of stations contain either no breaks or at most a couple of breaks. 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 (Figure 5.12). Overall, breakpoints occur on average every 21 years taken across the ISTI databank as a whole using the SNHT score of 16. This return period includes breakpoints that have been forced to account for timeseries cessation and resumption as described in Section 5.2.

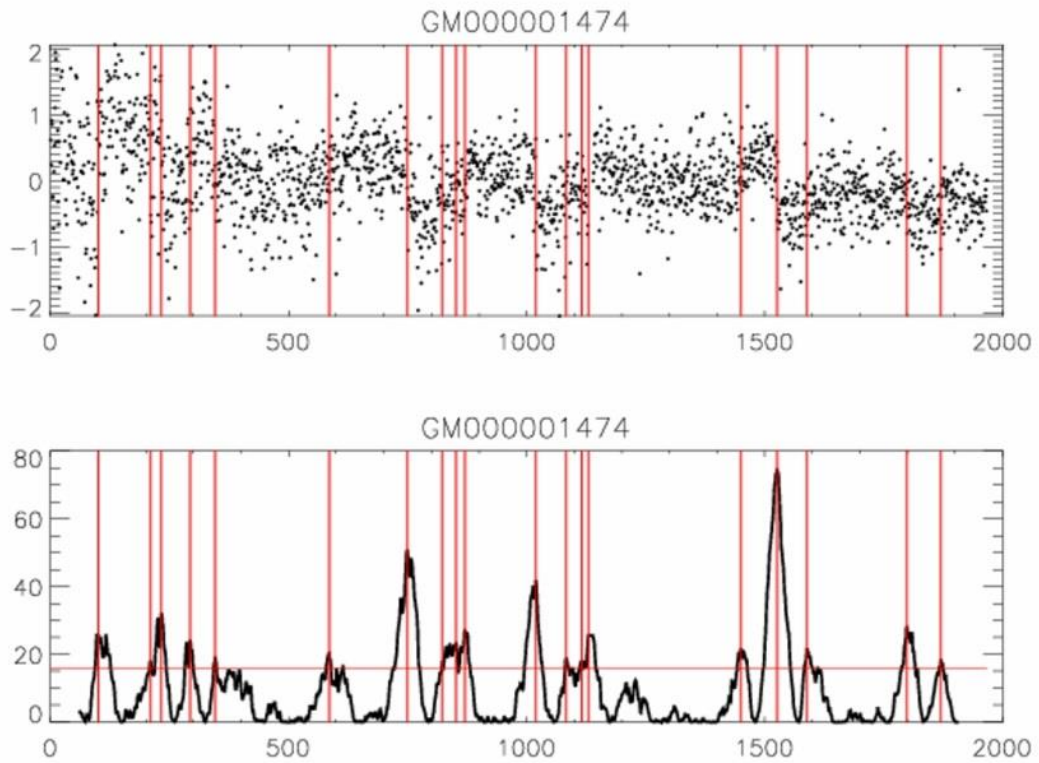


Figure 5.12. As Figure 5.6 but for GM000001474 Bremen, Germany (53.0464°N, 8.7992°E 4 m.a.s.l) with the highest number of returned breaks at 19.

Number of breaks identified	Number of stations
0	8398
Data gaps breaks of > 36 continuous months only	1909
1	5915
2	4292
3	2944
4	1464
5	919
6	598
7	416
8	274
9	160
10+	350
Total	27639

Table 5.2. Summary of the preponderance of breakpoint detection at an SNHT critical value of 16 across the raw ISTI databank stations retained following the analysis undertaken in chapter 3. 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 summarised in Figure 5.13 for the four adjustment approaches. All four histograms contain an identical total number of adjustment estimates by construction. The four distributions are clearly distinct from one another, with the $\text{neighbour}_{\text{segment}}$ based approach as a clear outlier in comparison to the other three. The $\text{neighbour}_{\text{segment}}$ approach histogram contains no dip centred close to zero adjustment size – a feature present in all remaining distributions and also noted in the PHA technique as applied to GHCNMv4 (Menne et al., 2018). This so-called ‘missing middle’ dip is largest in $20\text{CRv3}_{\text{short}}$ adjustments. Given that the detection and adjustment periods are identical for that approach, this would be expected by construction. 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 which must apply to all statistical breakpoint detection techniques. 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 $\text{neighbour}_{\text{segment}}$ approach also has the largest standard deviation and more large adjustments (fatter tails) than the remaining three techniques. The $\text{neighbour}_{\text{double-diffs}}$ technique shows the next highest standard deviation. The lowest standard deviation is from the $20\text{CRv3}_{\text{long}}$ adjustment technique. All four techniques show a very slightly negative mean adjustment of between -0.011°C and -0.022°C compared to the mean adjustment of -0.023°C reported for GHCNMv4 (Menne et al., 2018).

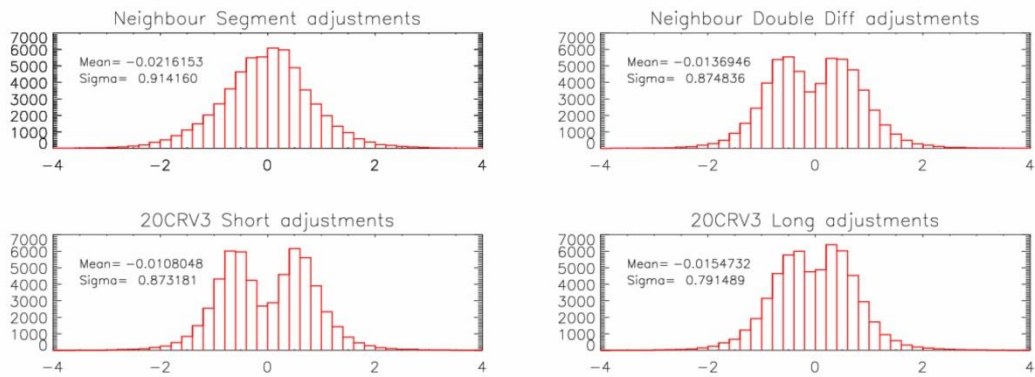


Figure 5.13 Distribution comparison of adjustments for the four adjustment approaches employed in this analysis using an SNHT critical value of 16.

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.

5.4.2 Evaluation of impacts on individual station series

Having ascertained that in aggregate the 4 adjustment approaches appear reasonable, next the impacts on individual station series must be ascertained. Firstly, consideration is given to those stations introduced in Section 5.2.3 to illustrate the impacts of the 4 homogenisation techniques utilised here on these time series.

Recourse is made also to the GHCNMv4 neighbour based adjusted series returned by NCEI’s PHA algorithm as described in (Menne et al., 2018). Then a number of further stations are introduced to illustrate additional important features. For illustrative purposes, in this section, the adjusted series are normalised to be equal over the final homogeneous segment and then compared. This better highlights the time-varying nature of adjustments than comparing series normalised 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 this analysis is simply to determine overall reasonableness of the adjustment approaches via comparison. New benchmarking studies such as those performed by Venema et al. (2012) and Williams et al. (2012) may permit a

more absolute characterisation were the set-up to enable participation using 20CRv3 as the background field. Such an approach is beyond the scope of the present analysis.

For the single break case study station introduced in Figure 5.6, all four homogenisation methods produce a series that reduces the mean shift, as does the entirely independently produced GHCNMv4 neighbour based series estimate which clearly identified the same breakpoint (Figure 5.14). None of the adjustment techniques (including GHCNMv4), by construction, is able to deal with the obvious shift in variance in the time series associated with the identified breakpoint. Figure 5.15 shows that for this station all techniques adjust the early period of record to be cooler than the raw record. This adjustment is visually obvious and the various adjustment estimates are in broad agreement. Figure 5.16 is similar to Figure 5.15 but only shows on adjustment for clarity, in this case 20 CRv3 Long.

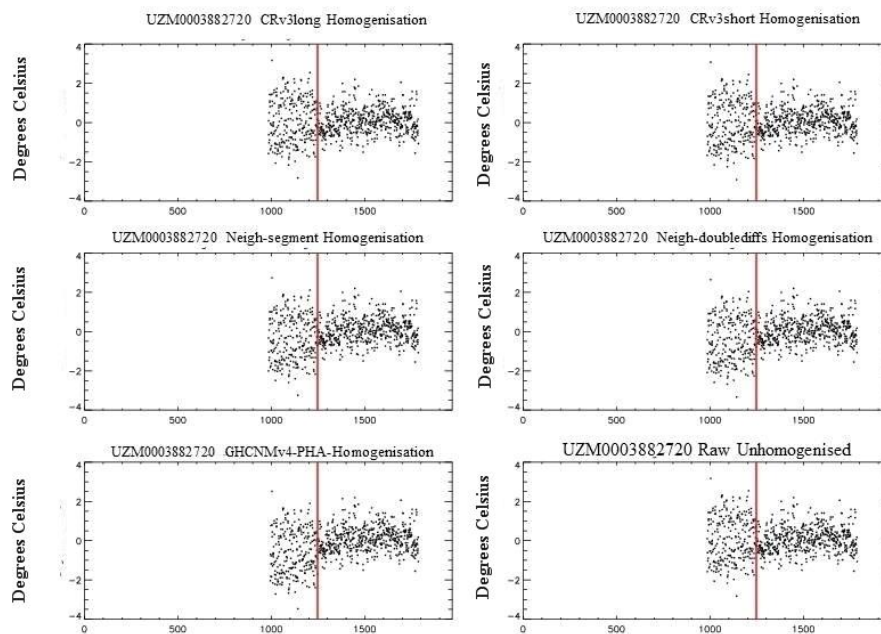


Figure 5.14 The resulting set of 20CRv3 minus station difference series for station Baisun Uzbekistan, UZM00038827 (top left panel 20CRv3 long; top right panel 20CRv3 short; middle left panel neighbour segment; middle right panel doublediff; bottom left panel GHCNMv4 adjusted; bottom right panel raw unadjusted series is reproduced from Figure 5.5). The single breakpoint location identified in the present analysis is denoted by the solid red vertical line. Individual monthly values are shown.

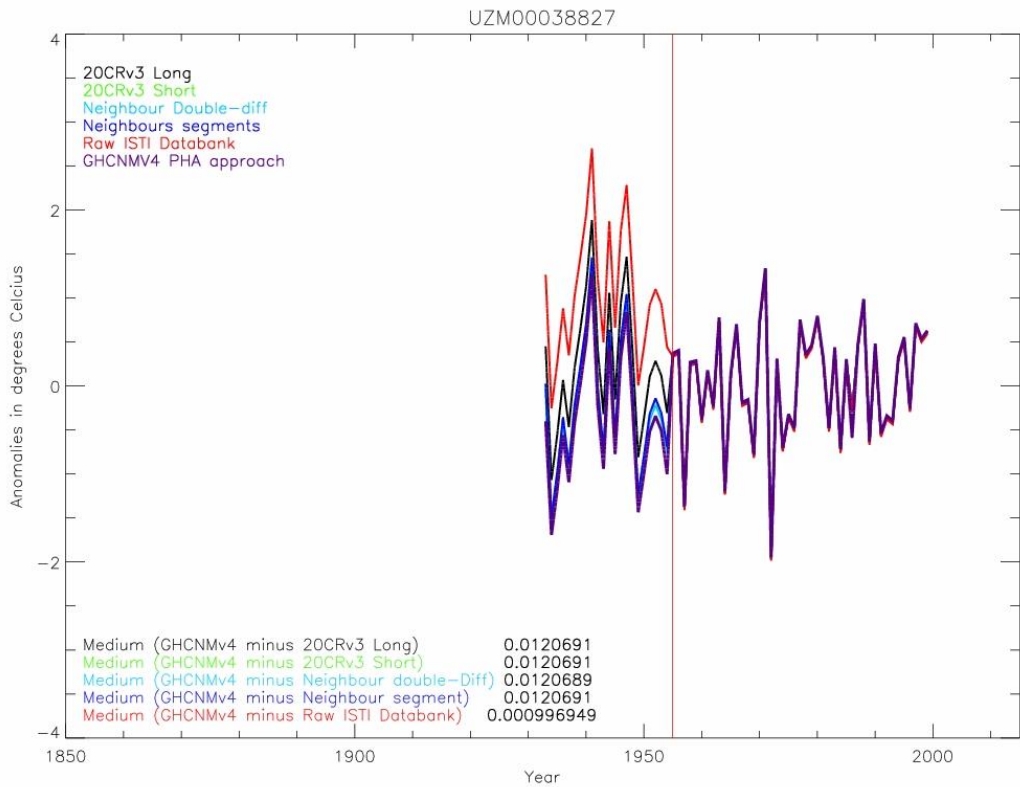


Figure 5.15 Annual time series of anomalies following application of adjustments (except for the raw series) and renormalisation to a 1961-1990 climatology followed by matching all series to be identical for the final homogeneous portion for illustrative purposes. Locations where breakpoints have been assigned and thus adjustments applied are denoted by solid red vertical lines .

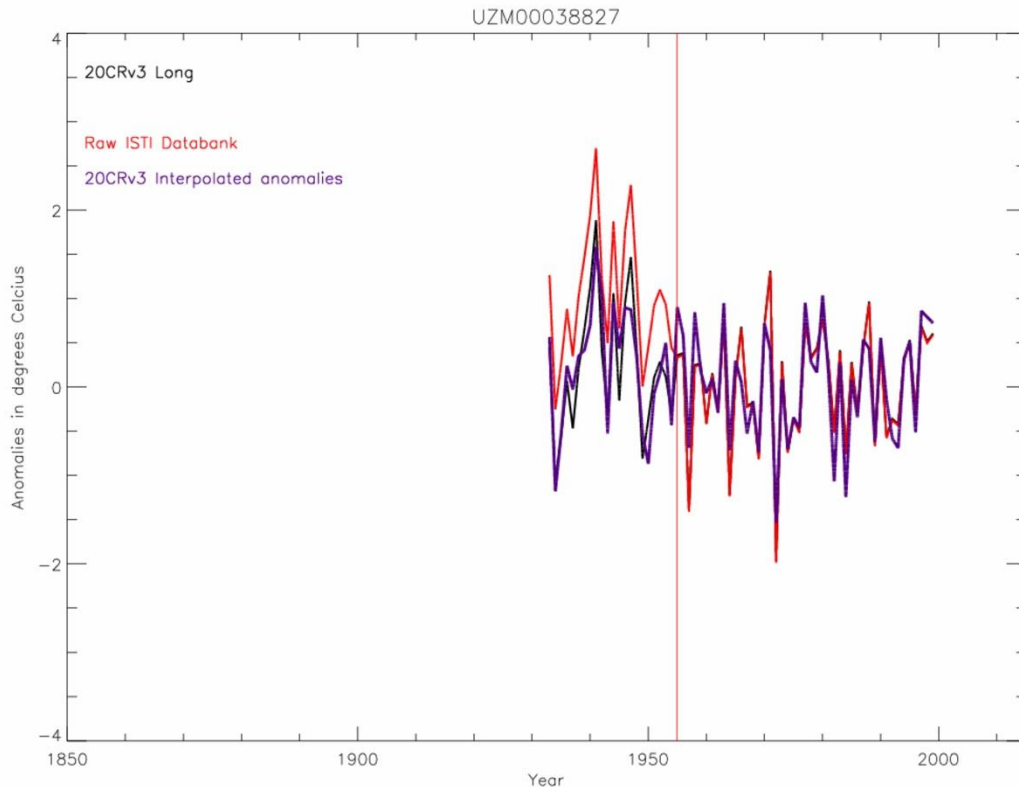


Figure 5.16 As of Figure 5.15, but only showing one adjustment (20CRv3_{long} , the ISTI raw series and 20CRv3 interpolated anomalies

Moving on in complexity to Stykkiosholmur in western Iceland which had multiple breaks identified by SNHT (Figure 5.7), the resulting adjustments show considerable spread (Figure 5.17) that grows back in time and particularly prior to a breakpoint identified in the early 1960s for which a considerable spread in estimated adjustments exists, and then again for a breakpoint identified in the late 1930s. Multi-decadal variability, driven by the Atlantic Multidecadal Oscillation (Yang et al., 2020, Knight et al., 2006, Allison et al., 2014) 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 raw. 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.

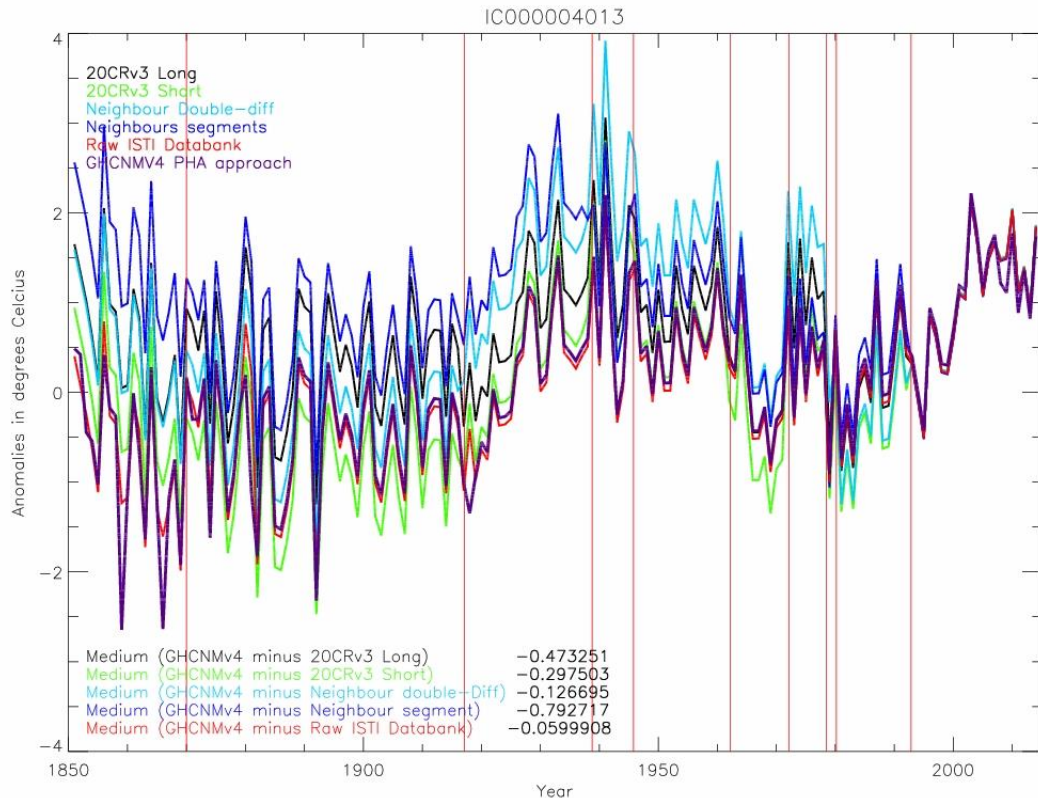


Figure 5.17. As Figure 5.15 but for Stykkiosholmur western Iceland. Differences in the final homogeneous segment relate to QC differences for raw and GHCNMv4 which alter some annual values.

For the Midway Island site with a long duration cessation of operation over WW2, GHCNMv4 has removed the entire pre-WW2 segment (Figure 5.18) 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 (systematic shift in mid-1950s). GHCNMv4 also either: i) has a much larger adjustment estimated for the brief homogeneous segment between the late 1970s and mid-1980s than any of the 4 solutions developed herein; or ii) did not detect and adjust for a break here.

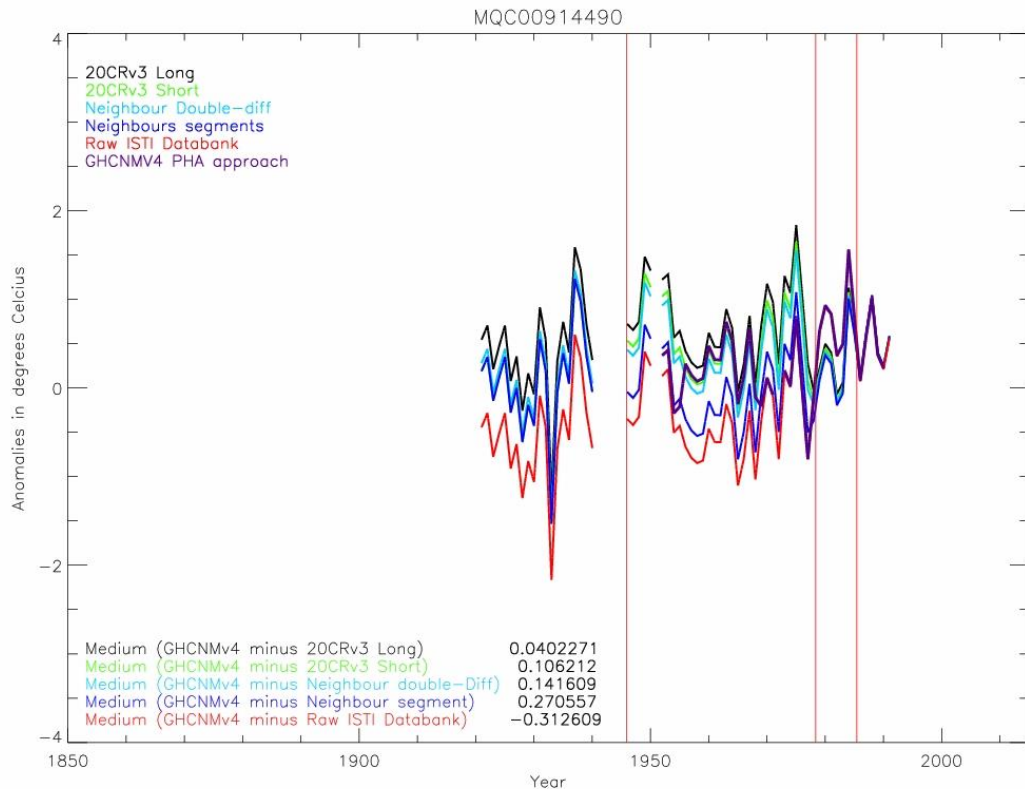


Figure 5.18 As Figure 5.15 but for Midway Island.

Turning attention to De Bilt (Figure 5.19) 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 (Chapter 4), all four adjustment methods are in broad agreement with each other and with GHCNMv4. Again, by construction, none of the methods accounts for variance effects although GHCNMv4 appears to do so perhaps through more aggressive quality control application. The individual time series plots are all similar to one another (Figure 5.20) with differences being considerably smaller than the inter-annual variability in this maritime mid-latitude location. 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 to agree closely but adjust to be substantively cooler than the raw data, by of the order 1°C.

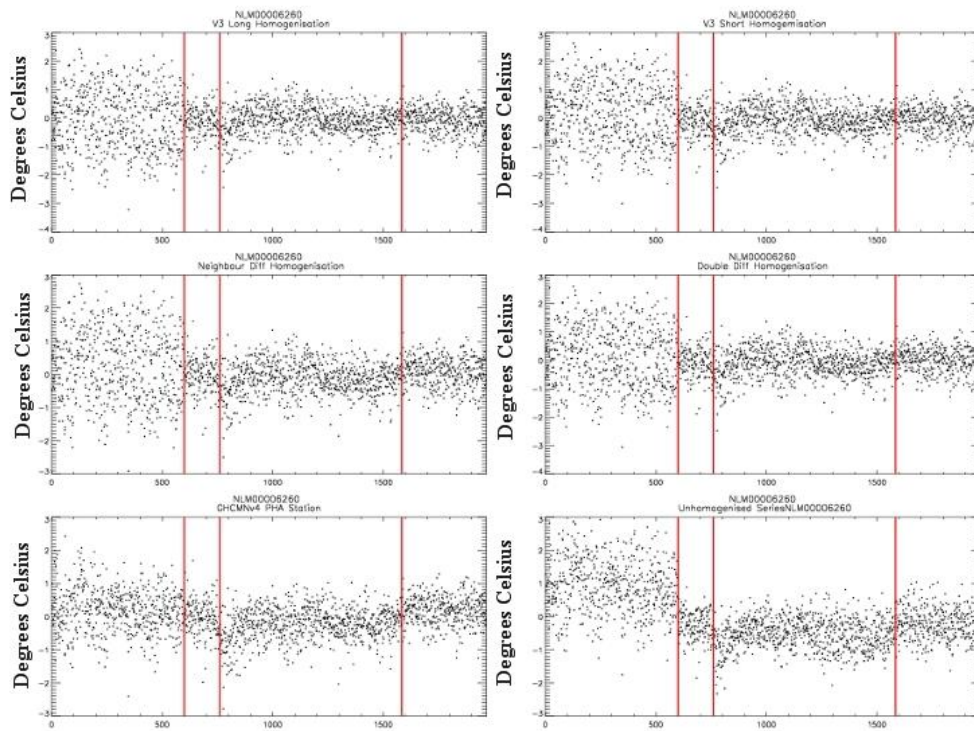


Figure 5.19 As Figure 5.14 but for station NLM00006260 De Bilt, Netherlands.

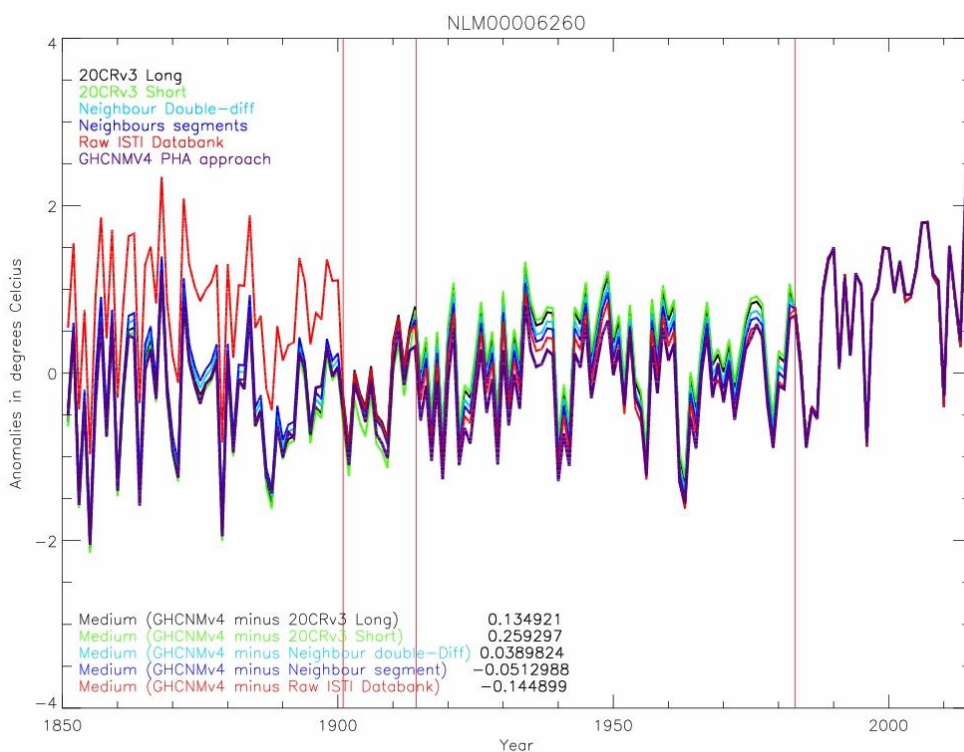


Figure 5.20 As figure 5.15 but for De Bilt, Netherlands. Note that this series extends back further than 1850 but the assessment of homogeneity herein has been truncated to 1850 so the series is accordingly truncated here.

A challenging proposition for any homogenisation procedure is a station that repeatedly starts and ceases (apparent) operations such that the data consists of repeated short bursts of records. However, such stations also tend disproportionately to be in data sparse regions where retention of as much data as possible is very valuable to the estimation of global and regional averages. Station KEM00063708 located at, Kisumu, Nyanza, Kenya (0.1°N, 34.75°E) consists of six individual segments of data in short bursts of twenty to thirty years. Within two of these segments, further breaks have been assigned by the present method. All the adjustment approaches yield distinct solutions here with no obvious pattern (Figure 5.21). There are large differences in adjustment estimates for the segment of record over the 1930s to 1950s with also divergence in estimates of the required adjustment for the breakpoint identified within this segment. Differences are also present in the first segment, including behaviour in GHCNMv4, presumably arising from differences in QC choices. Long-term trend estimates arising from this site are highly uncertain and it is not clear what value if any, any of the various homogenisation approaches has in this case.

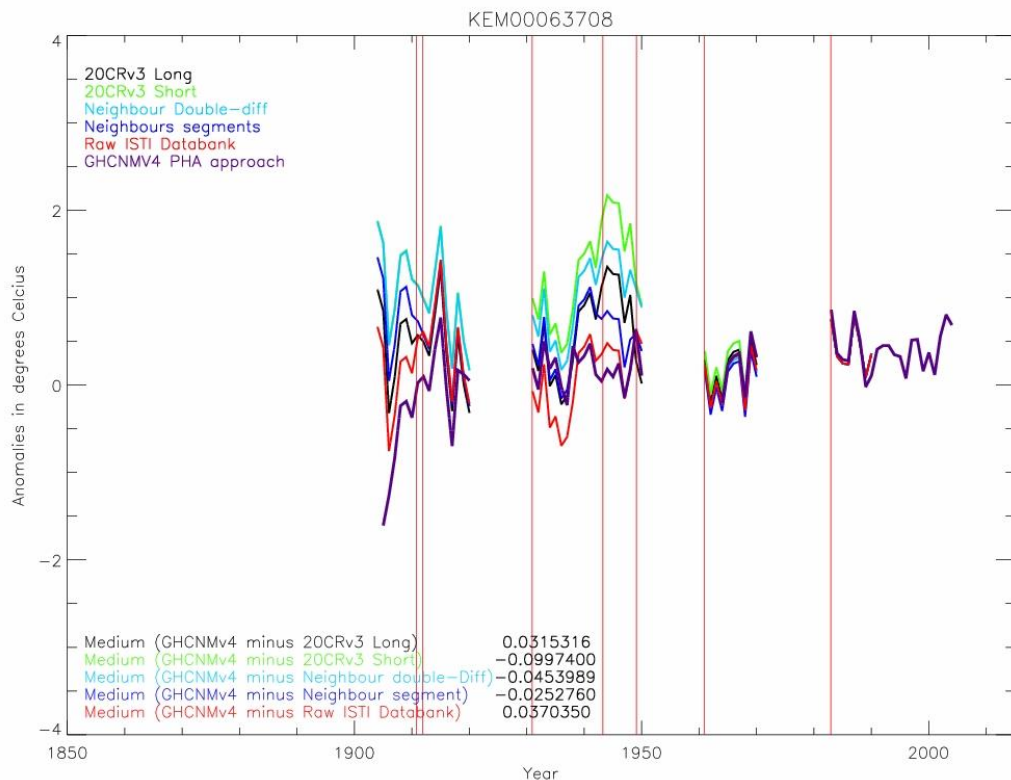


Figure 5.21 As Figure 5.15 but for Kisumu, Nyanza, Kenya (0.1 N°, 34.75 E° 1146 m.a.s.l)

Up to now, the focus was predominantly on stations identified as located in sparsely sampled regions, as defined in Chapter 4 analysis, where the nearest neighbours are located at a distance greater than 700km distance. But to investigate the performance of these methods more completely it is important to consider in addition stations located in well sample regions that traditionally would be well homogenised by pairwise methods which, as Chapter 4 indicates, may remain preferable in such cases. Station LG000026422 (Figure 5.22) located in Riga, Latvia (56.9625°N, 24.04° E, 17 m.a.s.l) shows all methods agree reasonably well throughout the series despite the presence of 5 detected breakpoints. Total early series divergence amounts to less than 1 degree between solutions. There is somewhat greater divergence between estimates in the early-to-mid 20th Century. Differences between solutions are minor in comparison to the large inter-annual variability at this mid-latitude location. All approaches appear reasonable.

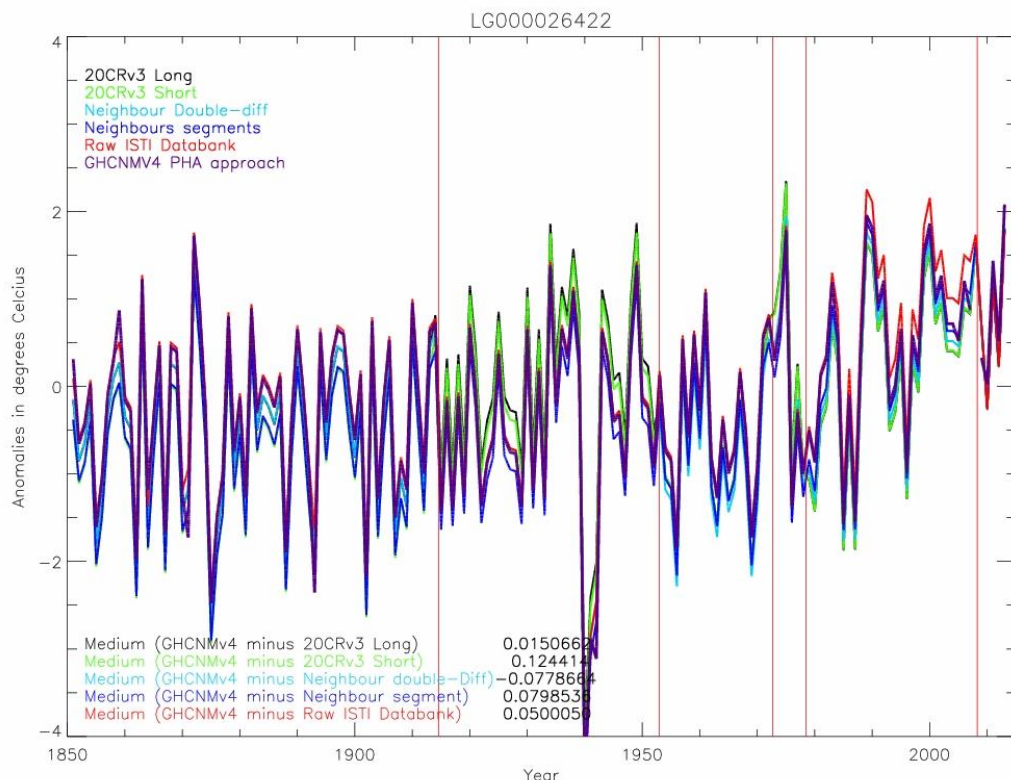


Figure 5.22 As Figure 5.15 but for LG000026422 at Riga ,Latvia at (56.9625°N and 24.04°E, 17 m.a.s.l). Note the temperature drop circa 1940. This is not an identified break.

Station ITE00001729 (Parma, Italy 44.8°N, 10.32°E 54 m.a.s.l) (Figure 5.23) exhibits a broad spread in solutions with GHCNMv4 being a marked outlier from around 1910 to the early 1990s. This presumably relates to GHCNMv4 either not

detecting a breakpoint identified by the present technique in the early 1990s or, more likely, adjusting for it in a very distinct manner. The four estimates of adjustments used here are in closer agreement with the raw data than with GHCNMv4 and often disagree on the sign of the required adjustment, being spread both above and below the raw data. The two neighbour based approaches share some characteristics with GHCNMv4 for most of the series but are much less extreme. The two 20CRv3-based sets of adjustments diverge significantly between 1890 and 1910. Before 1890 they are indistinguishable. Again, for this station, all approaches would appear to be at least as reasonable as GHCNMv4, which perhaps is questionable, at least for much of the 20th Century.

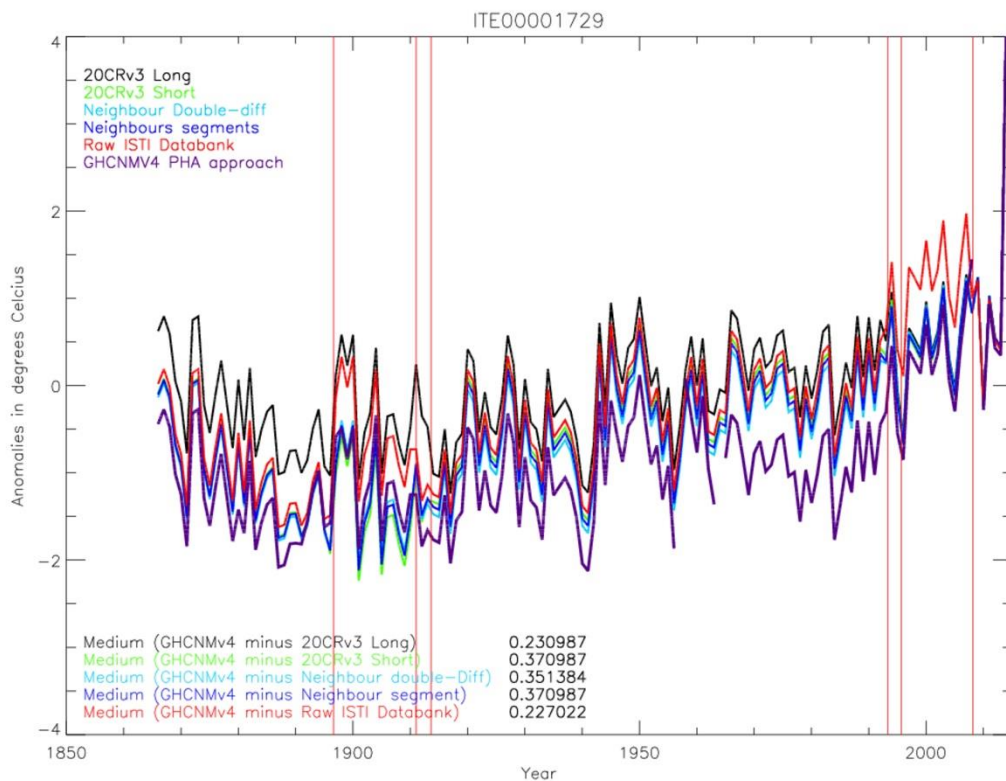


Figure 5.23. As Figure 5.15 but for ITE00001729 at Parma, Italy at (44.8°N, 10.54°E, 54 m.a.s.l.)

Station USW00014837 (Madison Dane County Regional Airport, Wisconsin, USA, 43.14°N 85.35°W, 264 m.a.s.l.) (Figure 5.24) was likely relocated to the airport from town at some time in its history given that the airport only opened in 1939 yet the station commenced in the late 1860s. The site has six breakpoints identified during its history. The break in 1939 at a SNHT score of 19.65 is the likely station move to the airport. GHCNMv4 removed data between January 1942 and November 1945. Homogenisation by 20CRv3_{long} is very comparable to GHCNMv4 method except for

a short period in the 1880s. $20CRv3_{short}$ and $Neighbour_{double-diff}$ diverge from other estimates to be significantly warmer before approximately 1922. Before the first break at 1884 significant divergence exists between estimates with $20CRv3_{short}$ and $neighbour_{double-diff}$ being marked outliers that ostensibly appear warm biased in this period.

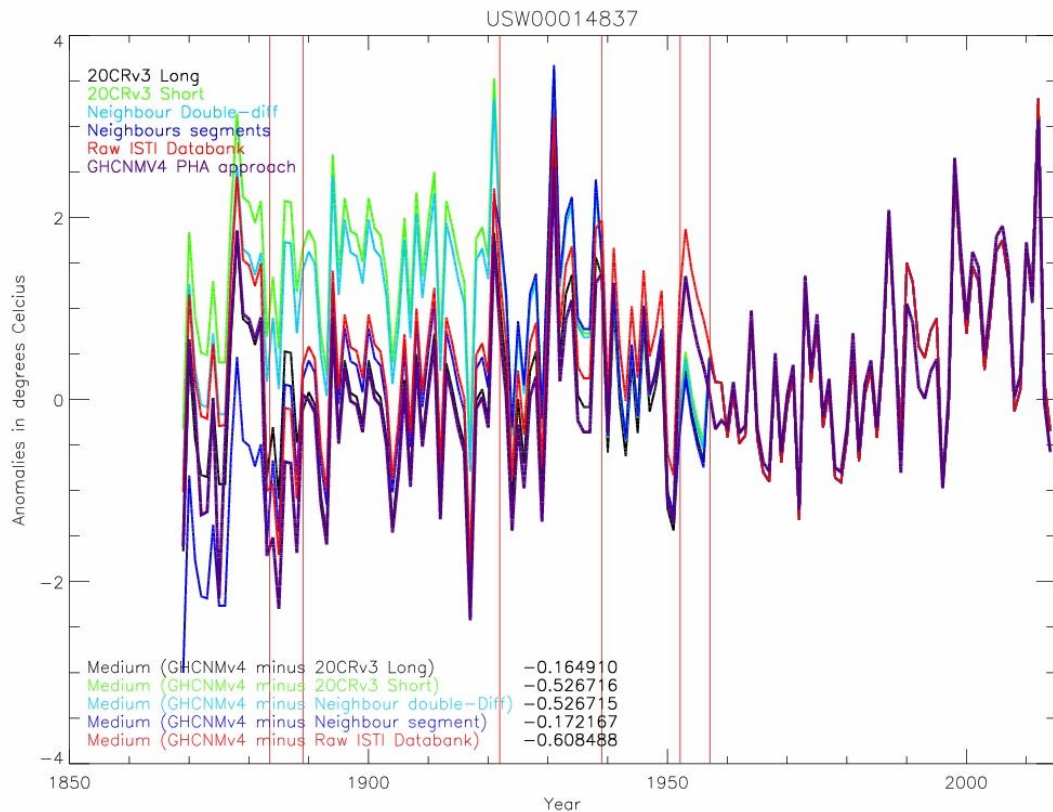


Figure 5.24 As Figure 5.15 but for USW00014837 at Madison Dane County Airport, Wisconsin, USA (43.14°N 85.32°W at 264 m.a.s.l)

The above examples are broadly indicative of a larger sample of stations that were considered manually. Overall, from a consideration of this broad sample of stations, the approaches deployed herein appear reasonable for the majority of cases. Where there are potential issues they tend to occur more prominently in the two neighbour-based adjustment approaches, but not always so as shown by the final example. However, more often than not across the subsample of manually inspected stations, it would be hard to consistently question the value of any the approaches in a manner which may lead to their rejection as a scientifically reasonable approach. Furthermore, the resulting adjustments more often than not better agree with those estimated from the neighbour-based GHCNMv4 analysis than they do with the

original raw data, particularly so in well-sampled regions where PHA is expected to perform best and 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 4 adjustment methods do, however, show considerable spread in some cases justifying analysis as distinct possible approaches to the derivation of adjusted series.

5.4.3 Spatial anomalies

Having ascertained that all techniques appear to undertake at least reasonable adjustments at the station level the resulting series were gridded for the ‘raw’ data, GHCNMv4 and all 4 adjustment techniques. This gridding used simple grid box binning of 1961-90 normalised series (Section 5.3) to a 5 degree by 5 degree resolution with no attempt made at interpolation. There is a degree of mismatch between the GHCNMv4 station availability and that arising in the 4 adjusted versions arising from the analysis undertaken herein (Figure 5.25). Stations with values in GHCNMv4 but absent from this present analysis result from the additional blacklisting described in Chapter 3. There are 1280 stations present in the current analysis that are not included in the GHCNMv4 dataset owing to not meeting their 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 explaining some missing grid boxes in the GHCNMv4 maps that follow.

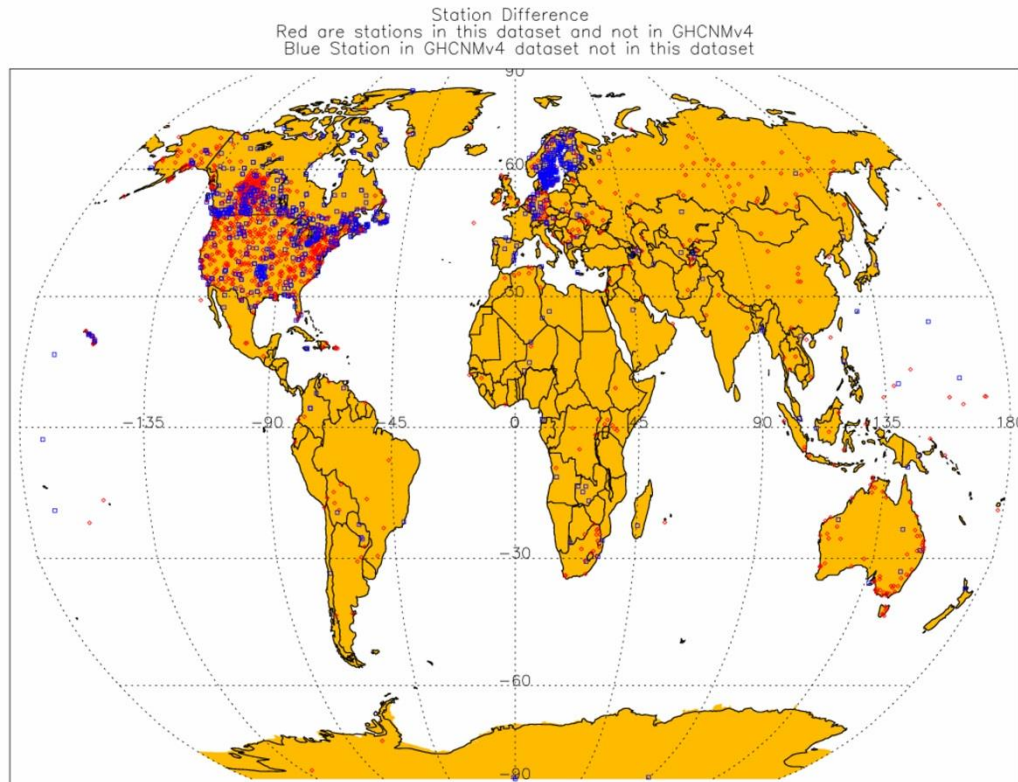


Figure 5.25 Summary of differences in station inclusion between GHCNMv4 and the present analysis. Red stations are present only in the current analysis. Blue stations are present only in GHCNMv4.

Anomaly fields were compared for all six gridded products for June in 1900, 1970 and 2000 respectively (Figures 5.26 through 5.28). June was chosen to minimise the absolute range to be plotted given the marked seasonality in Northern Hemisphere mid-latitude LSAT variability. Thus any impacts of poor or questionable adjustments would be expected to be more apparent. 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.

Differences would be expected to grow with distance from the 1961-90 climatology period and be largest early given that the cumulative effect of differences in applied adjustments typically grows back in time (Section 5.4.2). Some differences are apparent for example over Greenland in June 1900 but, in general, distinctions between the products are smaller than the colour-scale resolution necessary to span the full range of anomalies experienced. This behaviour holds for all estimates and over a broader range of months and years than are possible to show here for illustrative purposes. The impact of the choice of data product upon large-scale

monthly mean anomalies is thus small relative to monthly variance and the consideration of monthly anomaly field behaviour does not give rise to any concerns about the efficacy of the methods.

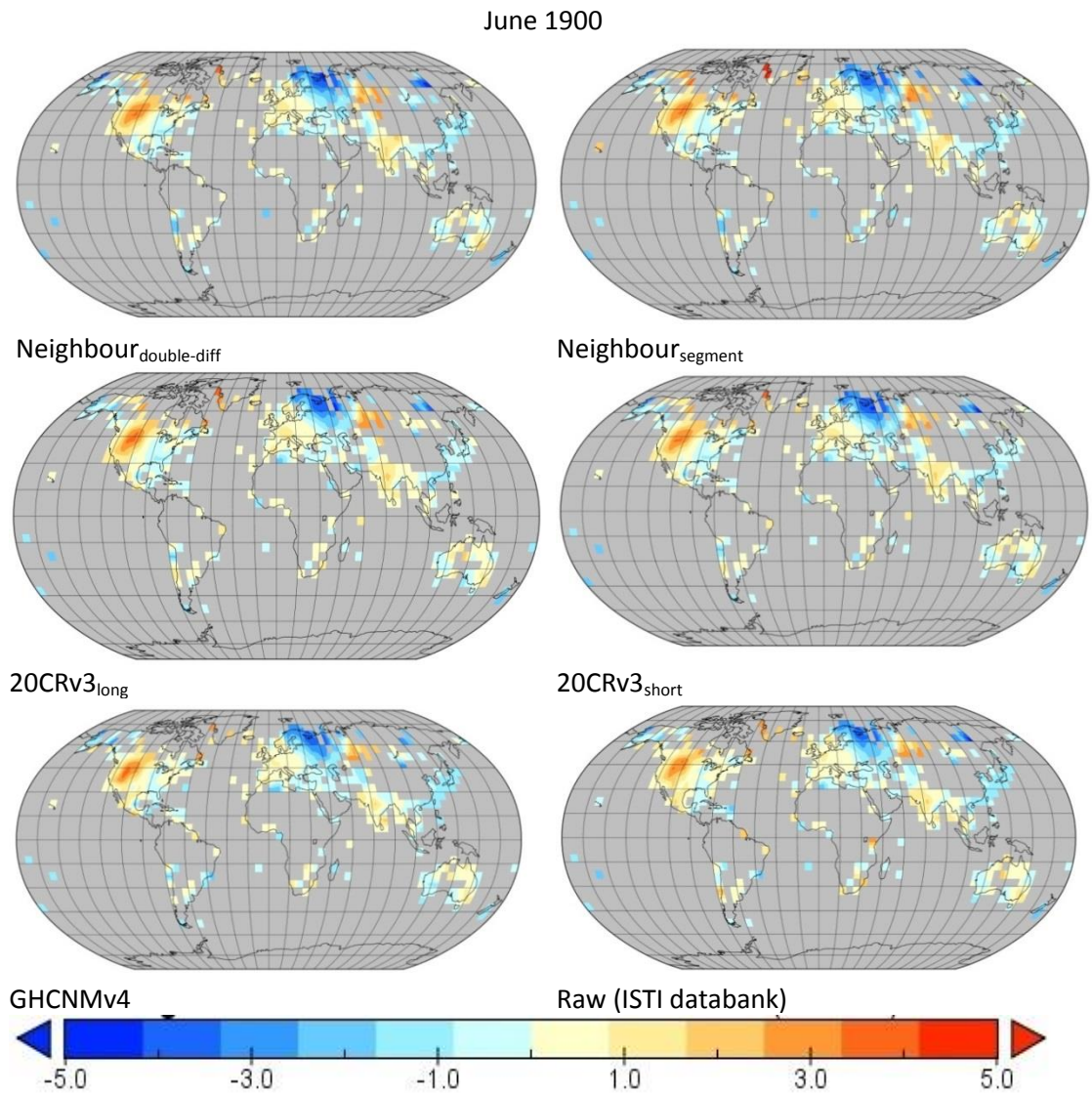
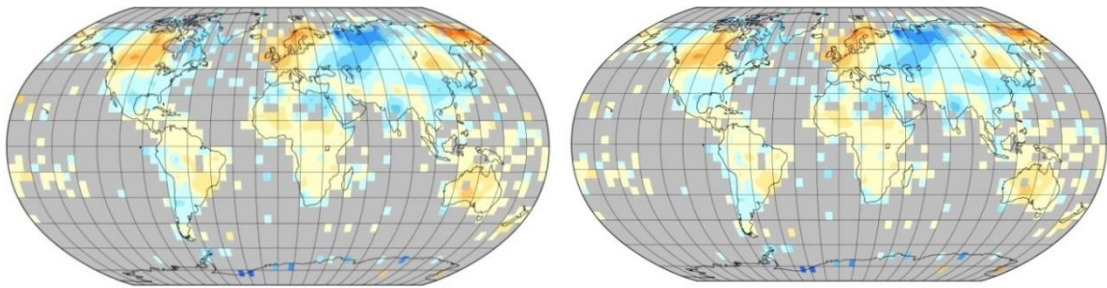
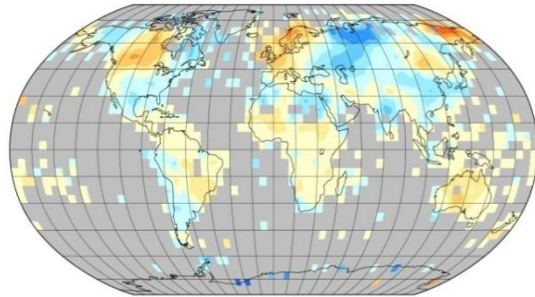


Figure 5.26 Maps of June 1900 gridbox anomalies from a 1961-1990 climatology for (from top left to bottom right): double differencing; neighbour segment; 20CRv3 long; 20CRv3 short; NOAA NCEI's GHCNMv4 product and the original raw ISTI databank holdings. Plots produced using Panoply version 4.10.12 for windows.

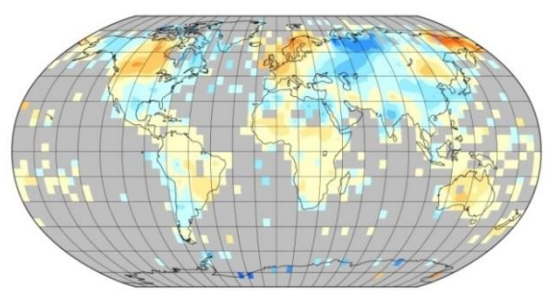
June 1970



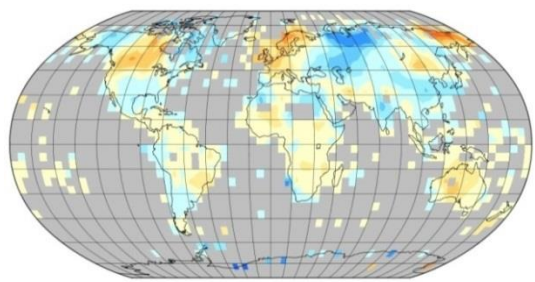
Neighbour_{double-diff}



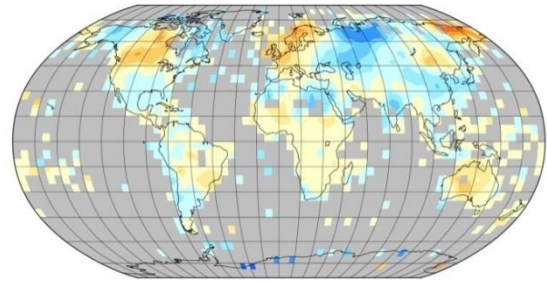
Neighbour_{segment}



20CRv3_{long}



20CRv3_{short}



GHCNMv4

Raw (ISTI databank)

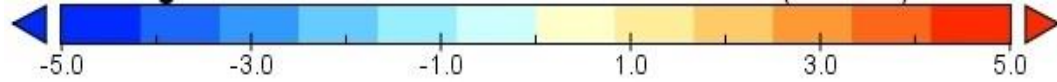


Figure 5.27 As Figure 5.26 but for June 1970

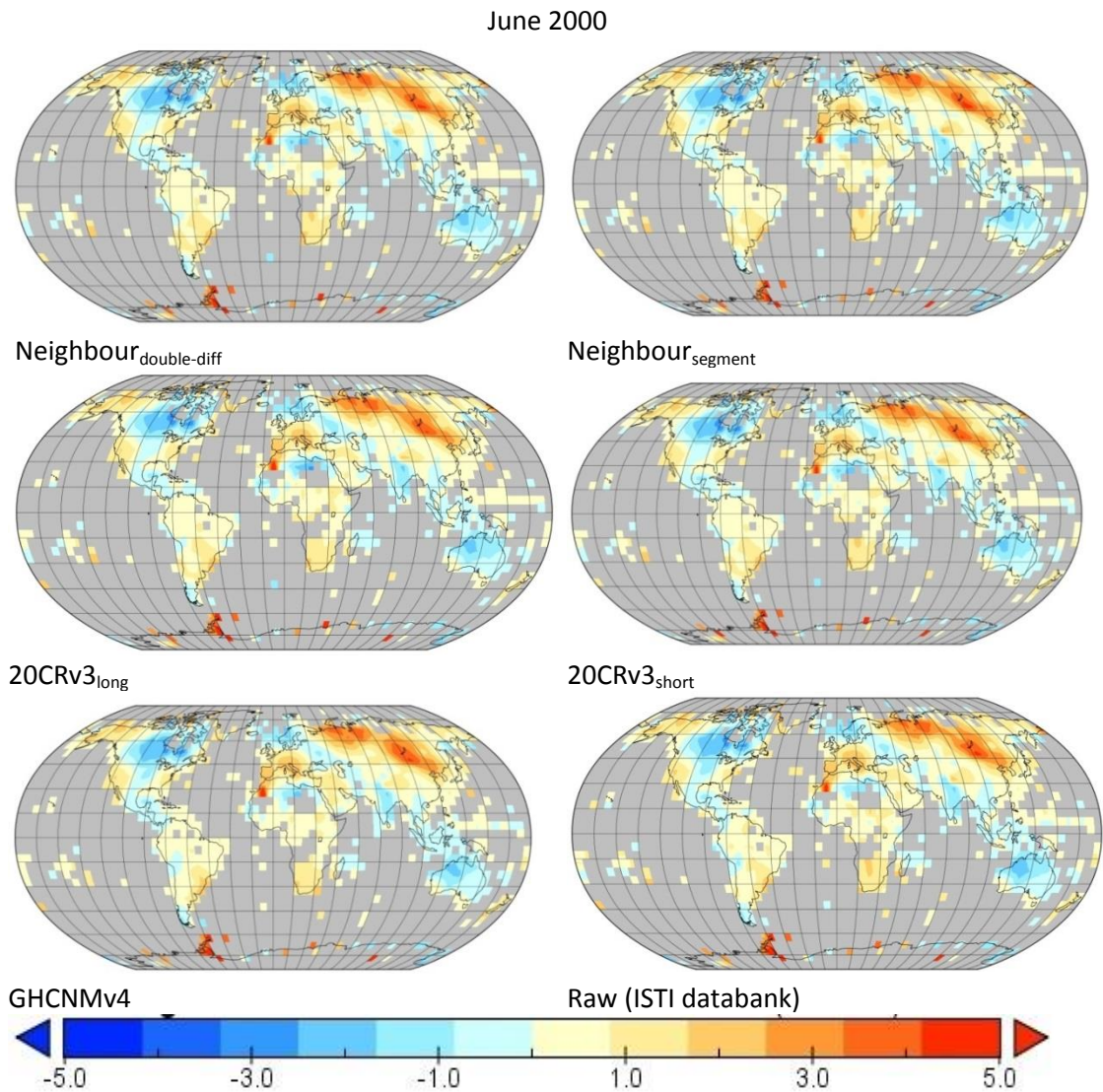


Figure 5.28 As Figure 5.26 but for June 2000

5.4.4 Spatial trends

Having ascertained that there are minimal differences in individual monthly anomaly fields, next spatial trend maps are considered over several periods. Differences arising from homogenisation choices, which act as red noise, should project much more strongly onto spatial trends than onto individual monthly anomaly fields, where geophysical anomaly patterns will tend to dominate. Examining trends thus helps to identify locations where variation exists between the four homogenisation methods and also in comparison to the independently produced GHCNMv4 product.

Starting with the very longest trends from 1851 to 2014, for which relatively few grid boxes contain sufficiently complete records (Figure 5.29), over Europe and

Russia, the 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 4 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 individual homogenisation decisions would be expected to be larger in this region. To varying extents, Neighbour_{double-diff}, Neighbour_{segment} and 20CRv3_{short} all suggest considerably less warming or even long-term cooling in the continental interior and suggest far fewer areas exhibit 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 little by way of systematic differences.

Trend 1851 to 2014

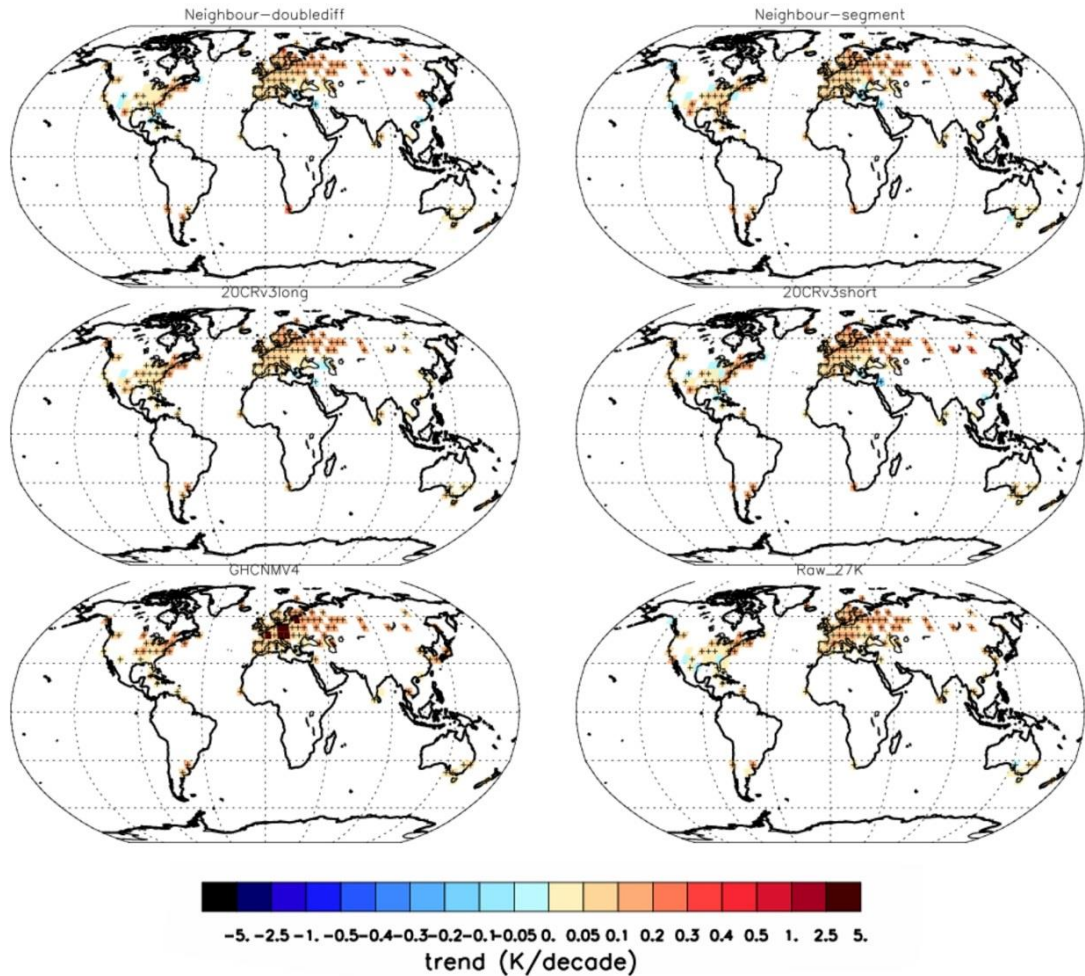


Figure 5.29 Gridbox trend analysis from 1851 to 2014. Trends have been calculated using OLS regression and based upon a requirement for 70% reporting with some reports in the first and final decile. Trend significance is denoted by + signs and ascertained from AR(1) corrected uncertainty estimation following Santer et al. (2008). Maps from top left to bottom right are for: Neighbour_{doublediff}, Neighbour_{segment}, 20CRv3_{long}, 20CRv3_{short}, GHCNMv4 and the raw ISTI databank holdings.

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, although there remain substantive coverage gaps (Figure 5.30). Trends over Europe are, again, largely consistent across all 5 estimates both in terms of trend magnitude and trend significance. However, 20CRv3_{long} does suggest less warming than the other estimates in the Balkans, 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 over this period. The cooling is also present, but to a much lesser degree, in 20CRv3_{long}. Conversely, GHCNMv4 estimates a robust warming in this region for this period. Over North America, the patterns, significance and even sign of the trend differ between the 5 estimates in a zone from the south-east of the USA through the central to upper plains. GHCNMv4 warms everywhere, but that warming is not significant across the south-east – the well documented warming hole (Pan et al., 2004, Kunkel et al., 2005, Mascioli et al., 2017). 20CRv3_{short} and Neighbour_{double-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-diff} show a slight cooling, in contrast to all remaining estimates.

Trend 1900 to 2014

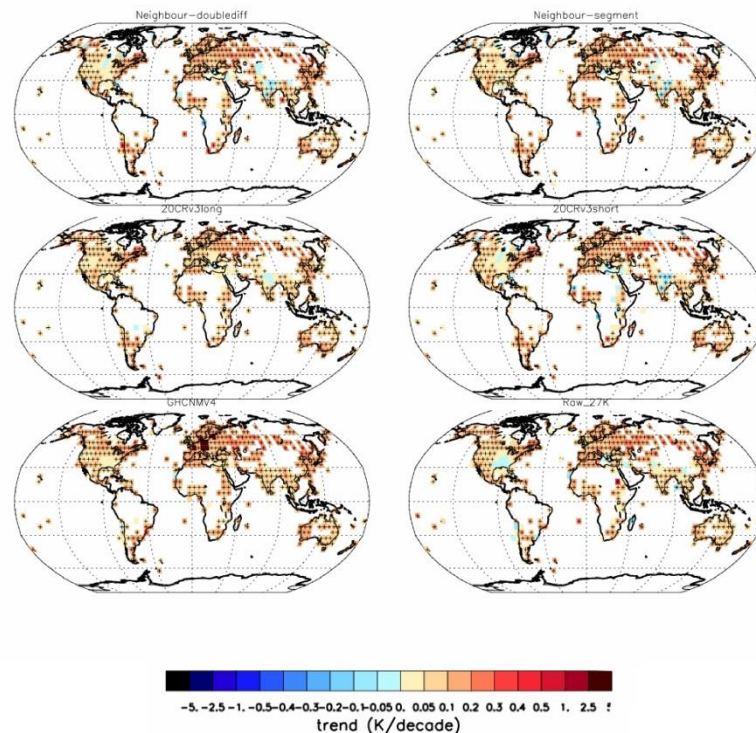


Figure 5.30. As Figure 5.29 but for the period 1900 to 2014

Moving forward to 1951-2014, when almost global land domain coverage is achieved, at least over the inhabited continents, trends are much more consistent between the 4 estimates produced herein and with GHCNMv4 than for the earlier periods (Figure 5.31). But there do remain some differences. Three solutions: $20CRV3_{short}$; $Neighbour_{segment}$ and $Neighbour_{double-diff}$ all exhibit no warming to varying degrees in parts of northern India and the Himalaya / Tibetan plateau whereas to an extent $20CRV3_{long}$ and, more markedly, $GHCNMv4$ imply significant warming. Some Pacific Islands in the four solutions developed herein show no warming whereas $GHCNMv4$ indicates somewhat greater warming. Otherwise, there is a remarkable degree of coherence between the 5 sets of estimates.

Trend 1951 to 2014

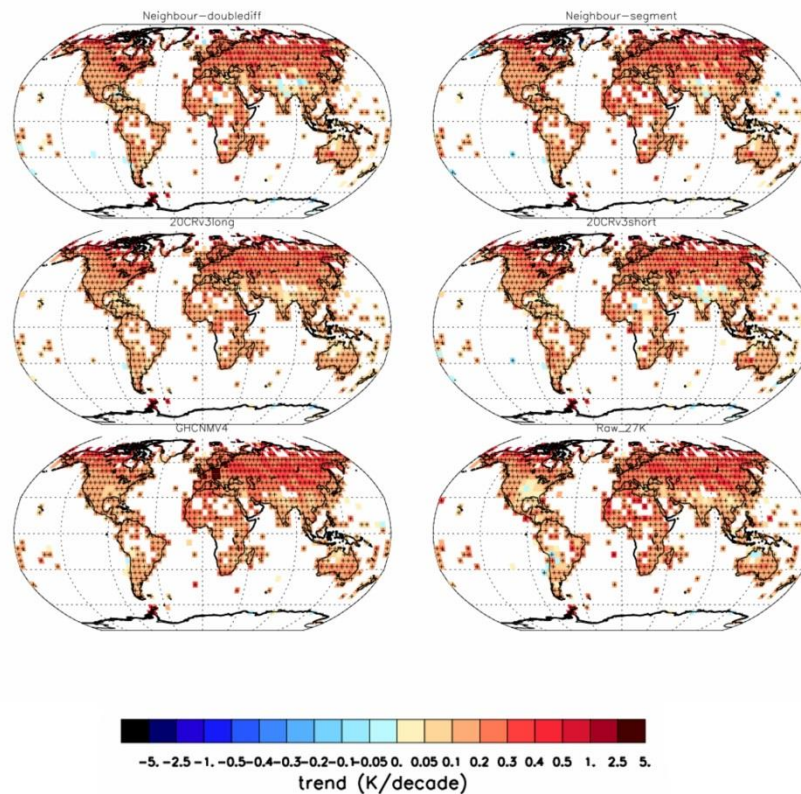


Figure 5.31 As Figure 5.29 but for the period 1951 to 2014

Over the much more recent period of 1980 to 2014 (Figure 5.32), all four new solutions agree with $GHCNMv4$ over most of the global landmass, particularly in Europe, Russia and the Far East, and generally in the most well sampled regions. The differences that are present are confined mostly to the more sparsely sampled

regions of the Southern Hemisphere. On occasion, in the Pacific Islands, the sign and significance differ between the four solutions developed herein and GHCNMv4.

Trend 1980-2014

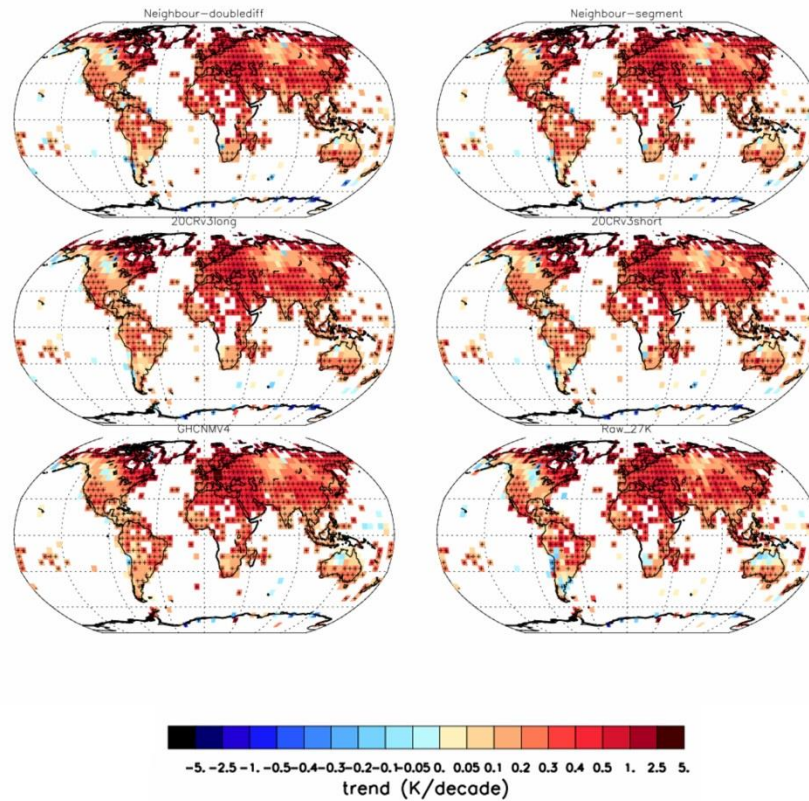


Figure 5.32 Intercomparison of trend analysis between the different methods of homogenisation and GHCNMv4 on a global basis for the period 1980 to 2014

However, again, it is overall very hard to distinguish between the 5 estimates over this period.

In summary, globally, all four adjustment techniques show broadscale trend agreement between each other and with GHCNMv4 across a range of timescales. Where 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, back in time. While there exist interesting differences, that are most apparent in trends over 1900-2014, and predominantly arise over the Indian sub-continent and N. America, these differences are not sufficient to

disqualify any of the approaches as constituting reasonable approaches to homogenisation. 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 spatio-temporal structure (or ‘spottiness’), then it would have provided grounds for rejection of that approach.

5.4.5 Summary of assessment of the efficacy of approaches

Section 5.4 has, with increasing fidelity, assessed the efficacy of the 4 approaches to homogenisation developed herein. In Section 5.4.1 it was shown that the populations of returned adjustment estimates across all stations was reasonable in all cases, although with some caveats around the nature of the statistical distribution of estimates for the Neighbour_{segment} approach. Section 5.4.2 showed a number of case study station series. Within these case studies, and also a broader sample that were inspected, in individual stations there were obvious cases where individual solutions appeared potentially questionable. In some cases this included GHCNMv4. However, there was no consistent pattern sufficient to call into question any given approach. Section 5.4.3 showed that the choice of approach had little impact upon spatial anomalies on a monthly basis. Finally, section 5.4.4, while highlighting some distinctions between estimates, showed that all approaches provide reasonable spatial trend estimates. Based upon these analyses it is concluded that all approaches developed herein may constitute reasonable approaches to estimation of bias corrected LSAT series, although the limited verification undertaken cannot absolutely confirm this to be the case.

5.5 Regional, hemispheric and global analysis

Having ascertained in Section 5.4 the overall reasonableness of the approaches developed herein, this present section goes on to undertake a selection of regionally aggregated analyses. The regionally aggregated analysis considers the same series as Section 5.4 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.

5.5.1 Regional analysis

Regional analysis is restricted to Europe, North America and Australia, where in all cases some limited (in the case of Australia, extremely limited) data extend back to 1850. In all cases, the regions are defined by simple bounding boxes and for each month the available grid boxes have been averaged using $\cos(\text{lat})$ weighting. Resulting monthly mean regional series have been aggregated to annual mean time series. Following this regional analysis, a hemispherically aggregated analysis is performed which can bring in additional information from more data sparse regions. Finally, globally aggregated series behaviour is considered. In these comparisons, GHCNMv4 has been aggregated from the station series as in Section 5.3 to ensure that any implied differences arise from a combination of any station selection mismatches (Section 5.4.3) or the impacts of differences in station series adjustments, and not from additional post-processing choices to create hemispheric and global averages undertaken by NOAA NCEI.

5.5.1.1 European domain

Over Europe, all time 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 5.33). 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 to the production by Haimberger et al (2012) of the RICH approaches which assure a degree of independence in adjustment. This is more critical for RAOBCORE given that the reanalysis system ingests prior observations from the target station – an issue that does not apply here. 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 4 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 for most of the periods whereby 20CRv3 will *a priori* be well constrained throughout the record.

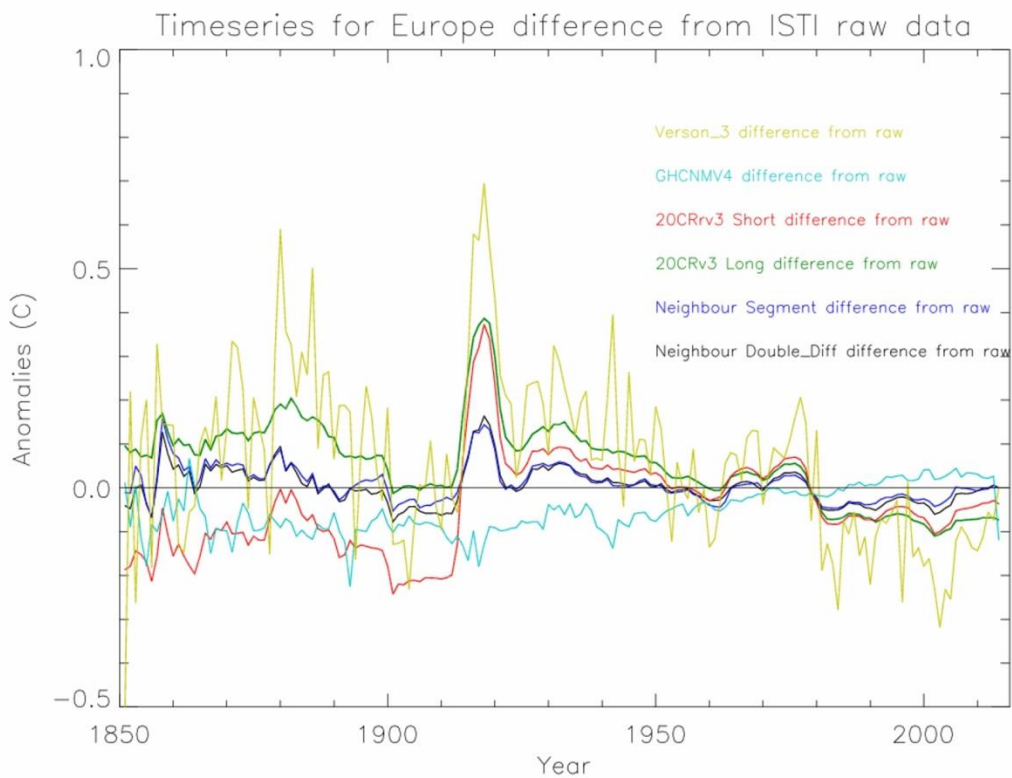
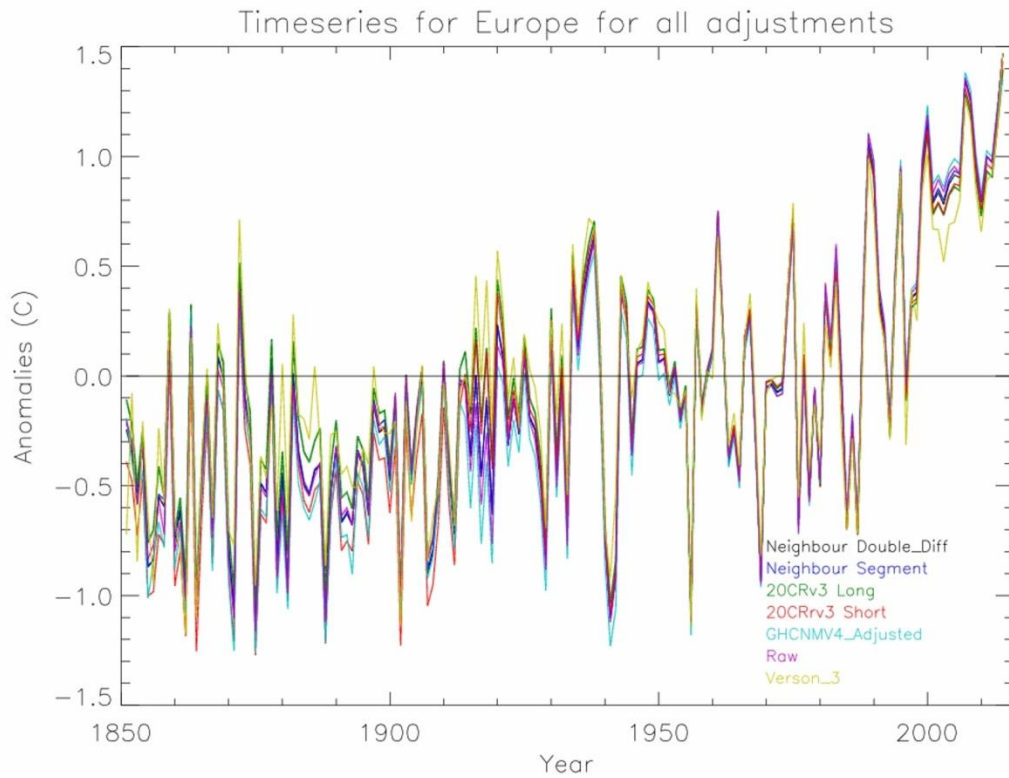


Figure 5.33. Top Panel Annual anomalies relative to 1961-1990 for the European domain defined as 35° N to 70° N and 10° W to 70° E relative to a 1961-1990 climatology for the four products developed herein, GHCNMv4, the raw ISTI databank and 20CRv3 interpolated to station locations and spatially matched to observational availability. Bottom panel is difference between GHCNMv4 and other data series

Considering long-term trends, all four solutions broadly agree with GHCNMv4 (Table 5.3). Over 1851 to 2014 20CRv3_{short} is very similar to GHCNMv4, but the other three solutions indicate slightly less warming over that period. 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 4 new solutions collectively bracket the raw data trends, and to a lesser extent, GHCNMv4.

Data Set	OLS trends in °C per decade				Change 1851-1900 to 2005-14
	1851-2014	1900-2014	1950-2014	1980-2014	
Neighbour _{double-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

Table 5.3 trend analysis for four time periods for the European region defined as 35°N to 70°N and 10°W to 70°E. 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 1850-1900 and 2005-2014 (final column).

5.5.1.2 North American domain

Over North America post-1970 there is strong agreement (partially forced by the choice of 1961-90 climatologies) between all timeseries, including GHCNMv4 (Figure 5.34). 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. The four distinct approaches to adjustment developed in the present analysis start to show sufficient systematic differences prior to c.1920 to be 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°C and 0.3°C than the four new adjustment techniques (Table 5.4, final column). Trends from 1900 onwards are reasonably in concordance between the various products.

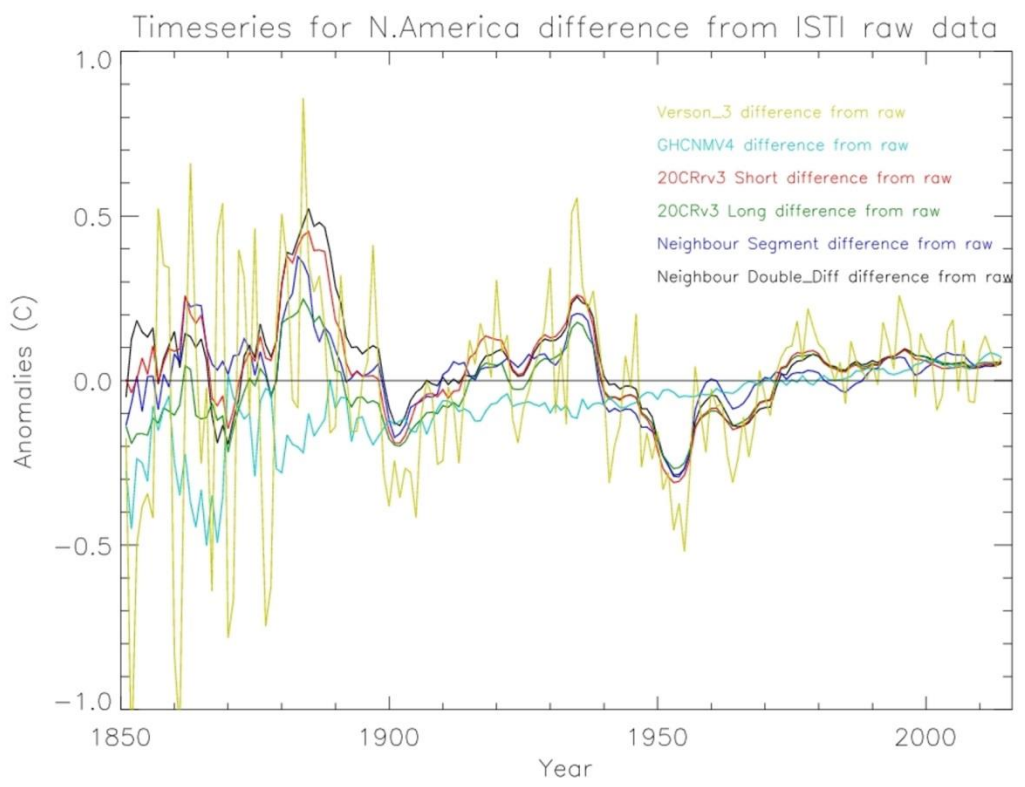
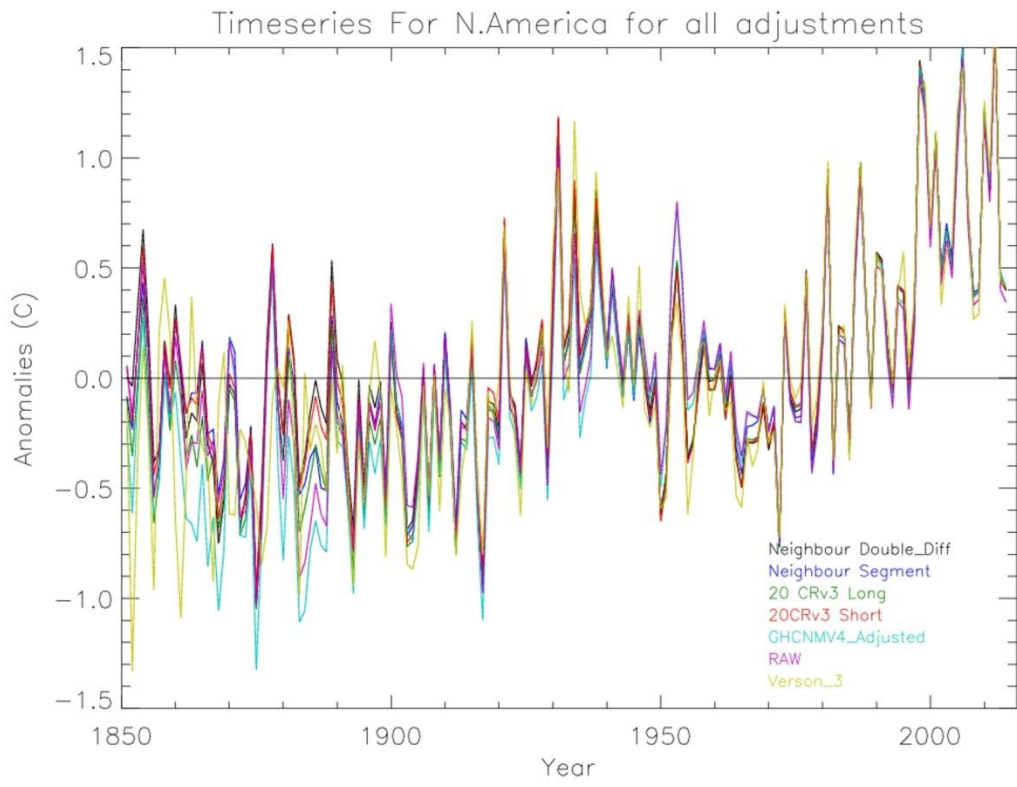


Figure 5.34 as Figure 5.33 but for the North American region defined as 25°N to 60°N and 45°W to 135°W

Data Set	OLS trends in °C per decade				Change 1851-1900 to 2005-14
	1851-2014	1900-2014	1950-2014	1980-2014	
Neighbour _{Doubled} iff	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 5.4 As Table 5.3 but for the North American region defined as 25°N to 60°N and 45°W to 135°W

5.5.1.3 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 5.35). The divergence between estimates prior to 1900 may be more due to scarcity of observations for that period and the difference in station availability between the different gridded series. The very earliest period data relies upon a single 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 databank series (Table 5.5). GHCNMv4 indicates somewhat greater warming between 1851-1900 and 2005-2014 by 0.16-0.23°C than the four new adjustment 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.

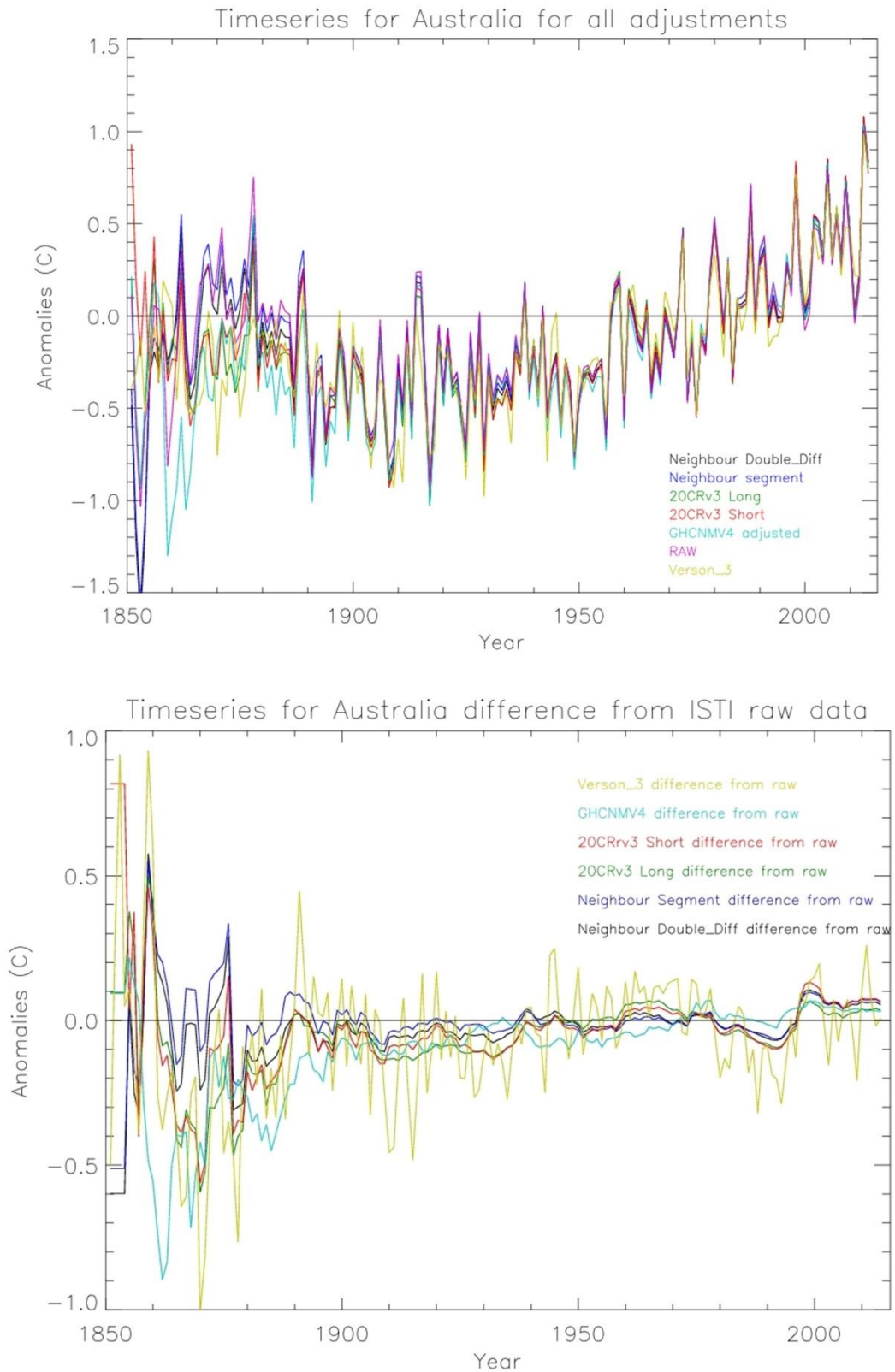


Figure 5.35. As Figure 5.33 but for the Australian region defined as 10°S to 45°S and 110°E to 155°E. Note that there are very few observing sites early in this series and great caution should be exercised in interpretation.

Data Set	OLS trends in °C per decade				Change 1851-1900 to 2005-14
	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

Table 5.5 As Table 5.3 but for the Australian region defined as 10°S to 45°S and 110°E to 155°E.

5.5.1.4 Hemispheric analyses

In the well-sampled Northern Hemisphere (Figure 5.36), all series are indistinguishable from one another after the mid-20th Century. Prior to 1950 GHCNMv4 becomes systematically cooler than remaining estimates 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 5.37) where the earliest records in the ISTI databank currently arise from the single station in Tasmania, although data rescue efforts can, hopefully, improve this situation in the future (Brönnimann et al., 2019). In the Southern Hemisphere, different products can periodically be distinguished from one another throughout the series. Differences become particularly marked prior to the 20th Century. The 20CRv3 reanalysis estimated at the locations and times of station observations is distinguishable periodically from remaining products. 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 remaining estimates.

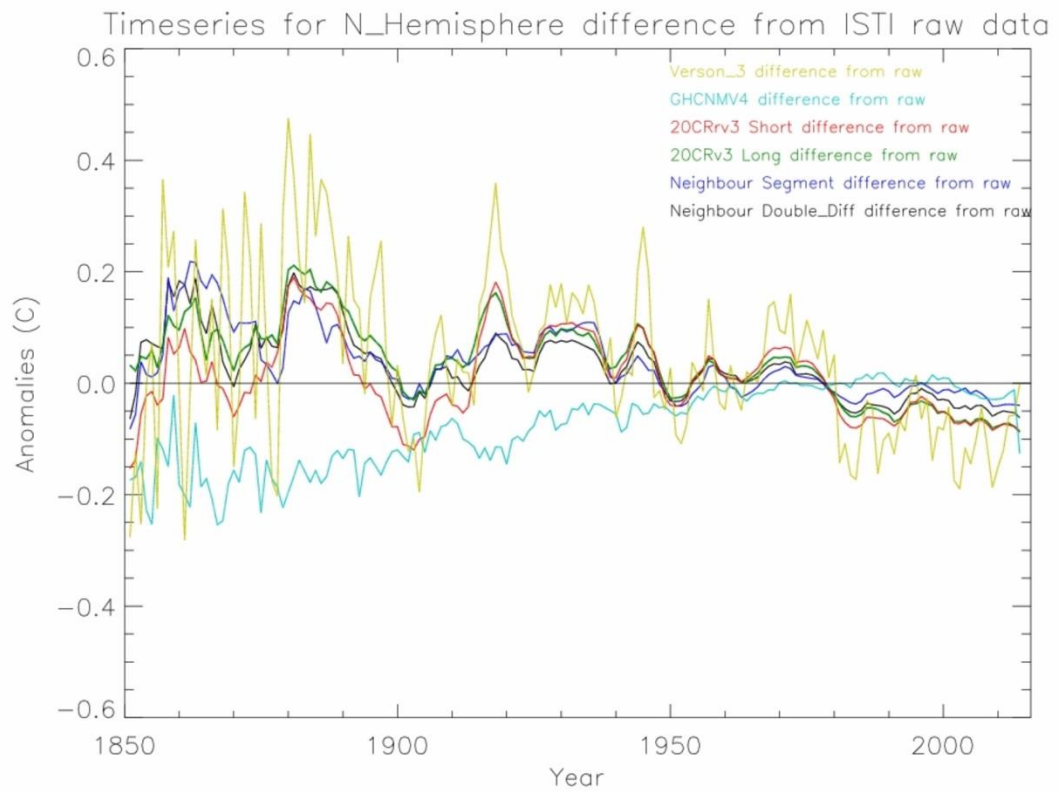
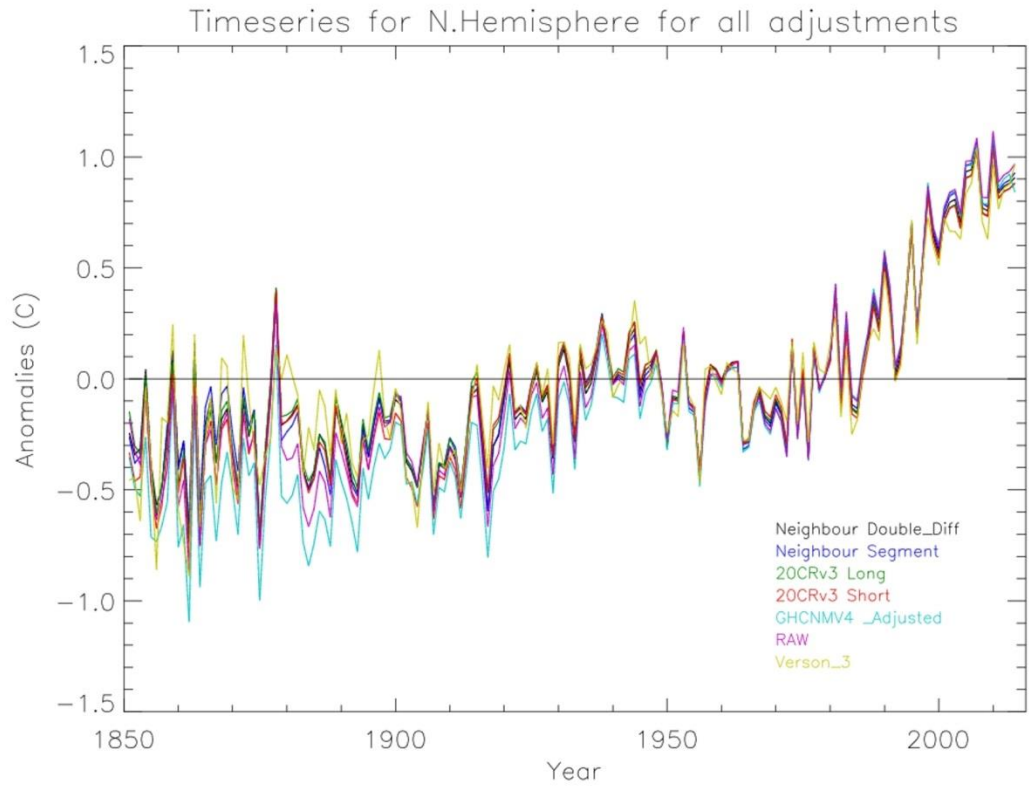


Figure 5.36 Northern Hemisphere Annualised time series shown for all 4 homogenised adjustments, 20CRv3 sparse input reanalysis, GHCMNV4 and the raw unadjusted time series.

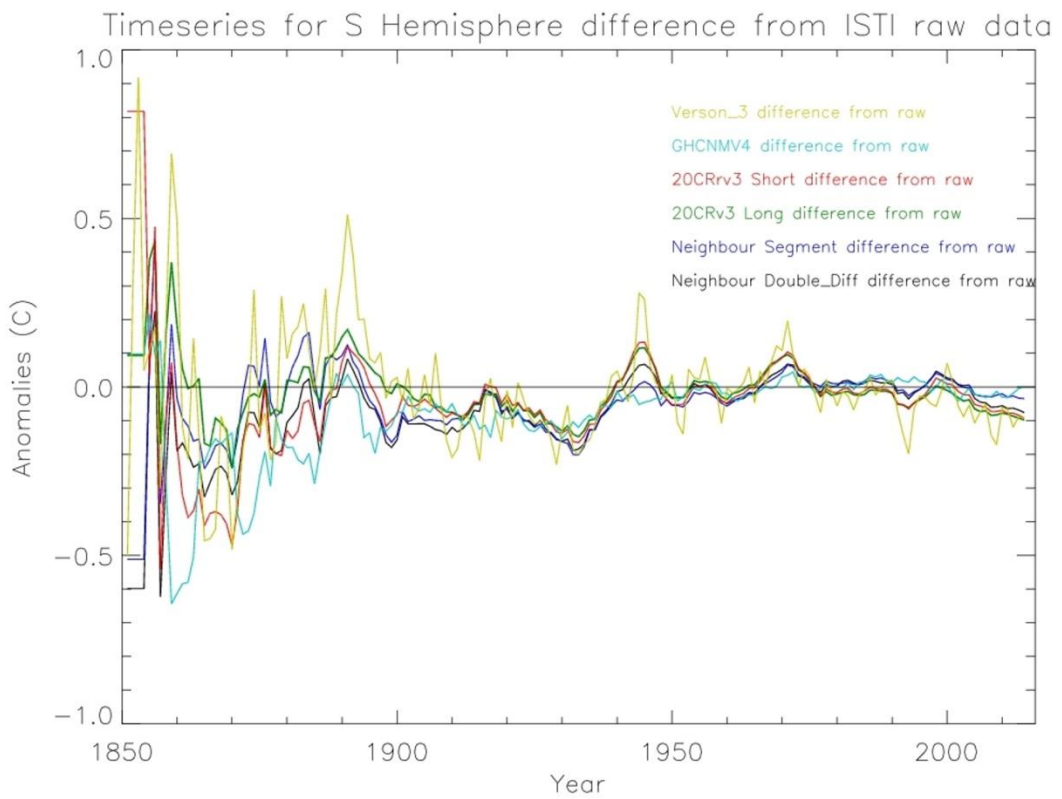
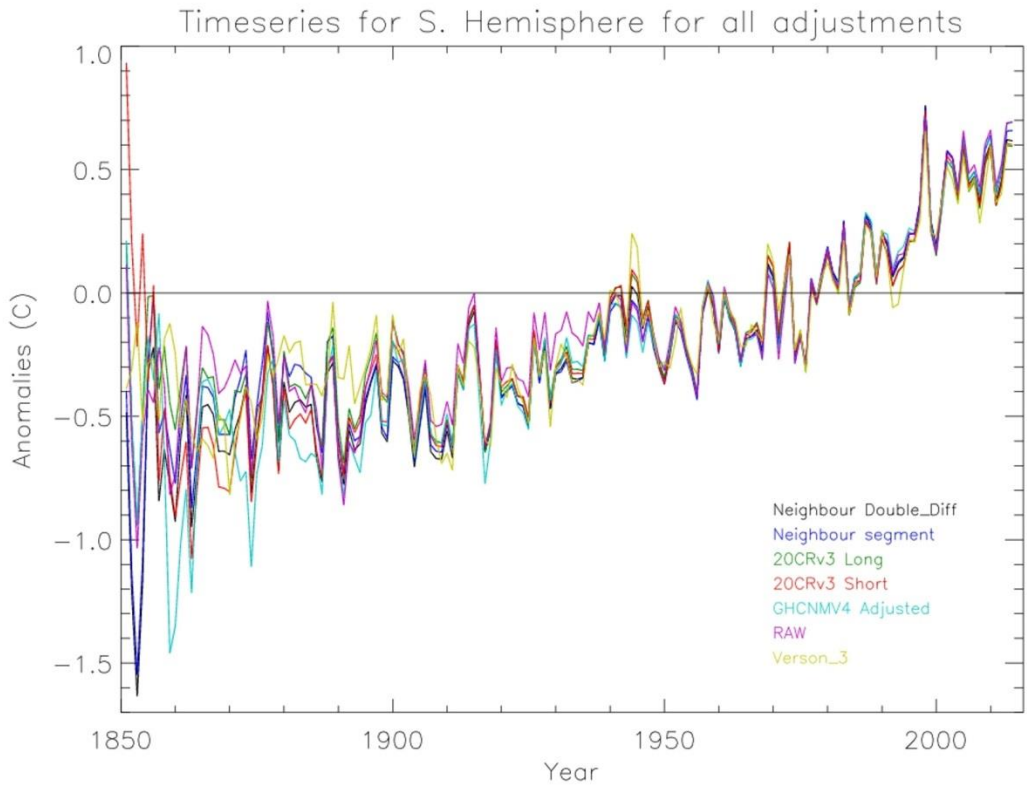


Figure 5.37. As Figure 5.36 but for the Southern Hemisphere

Considering hemispheric scale trends (Table 5.6), GHCNMv4 estimates more long-term warming than the remaining four adjustment techniques in the Northern Hemisphere. The disparity is larger for periods prior to 1950 but persists to a degree even for the 1950-2014 trend estimates. Then, in the most recent period, GHCNMv4 shows less warming than the 4 new estimates. The 4 new estimates are in all periods 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 5.4.3). In the Southern Hemisphere (Table 5.7), the estimates are considerably closer to one another than is the case for the Northern Hemisphere. There is more spread in estimates of the 1851-1900 to 2005-2014 changes in the SH across the 4 new estimates than is the case for the NH (compare the final columns in the two tables). This is also reflected in the OLS-regression based trend estimates for the same period.

Data Set	OLS trends in °C per decade				Change 1851-1900 to 2005-14
	1851-2014	1900-2014	1950-2014	1980-2014	
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

Table 5.6 As Table 5.3 but for the Northern Hemisphere.

Data Set	OLS trends in °C per decade				Change 1851-1900 to 2005-14
	1851-2014	1900-2014	1950-2014	1980-2014	
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	0.846

Table 5.7 As Table 5.3 but for the Southern Hemisphere

5.5.1.5 Global analyses

Global mean series (Figure 5.38) show characteristics in common with both Northern and Southern Hemisphere series, as would be expected by construction. But the behaviour is more influenced by the Northern Hemisphere mean series given that the method employed is a simple $\cos(\text{lat})$ weighted average of available gridbox series which are more prevalent in the Northern Hemisphere, particularly early in the record. Other approaches such as e.g. creating a weighted mean of the hemispheric means as is done in CRUTEMv5 would yield different characteristics (Section 5.6). All series are in good agreement after 1950 and diverge prior to this, with GHCNMv4 being cooler during this period than all remaining estimates as was the case particularly with the Northern Hemisphere analysis. All estimates support an assessment of long-term warming. Global trends are assessed in Section 5.6 in the context of remaining published estimates.

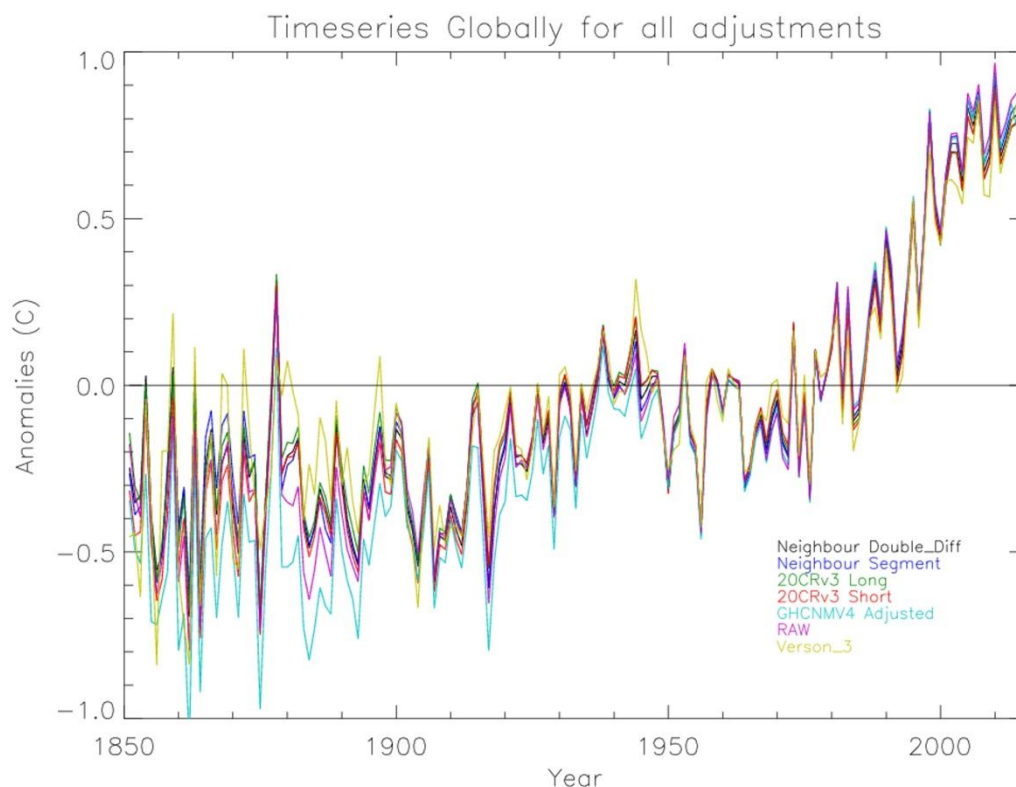


Figure 5.38 Annualised global analysis of all homogenised and raw time series

5.6 Intercomparison to other products

In section 2.5 of Chapter 2, five peer-reviewed land surface air temperature data sets were discussed and compared. All except the new Chinese land surface temperature dataset (CLSAT) produced by Xu et al. (2017) were used, albeit in now superseded versions, in IPCC AR5. In this section, these five datasets are compared to the four adjusted datasets created here using 20CRv3 sparse-input reanalysis. For these purposes, global-mean series have been sourced from public-facing repositories. Thus differences between the series may arise from some combination of: station selections, homogeneity assessments, and post-processing including choices over interpolation and averaging methods applied. For this section, the GHCNMv4 series for consistency is thus swapped out for the series made available directly from NOAA NCEI and used in their monitoring activities. Differences compared to the series in Figure 5.38 are minor apart from the series being cut in 1880 (although they shall shortly start reporting operationally from 1850 onwards).

There are substantial similarities in behaviour across a range of timescales between the estimates (Figure 5.39, although note that CLSAT, GHCNMv4, and GISS, at least in the public versions sourced do not extend all the way back to 1851). There are, however, some differences. The newly produced estimates are systematically lower than the remaining estimates post the mid-1990s. They are also systematically a little warmer prior to 1900 and markedly so over the 1880s and early 1890s. To some extent, these divergences are forced by the use of a 1900-2000 climatology which has been chosen to emphasise any dataset differences to the extent possible and practicable.

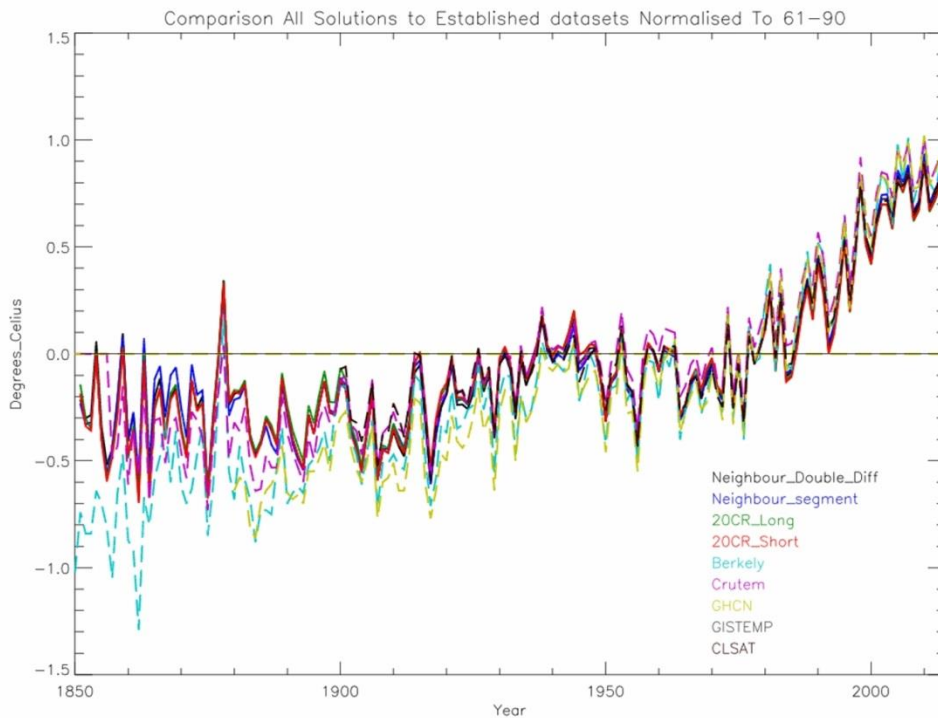
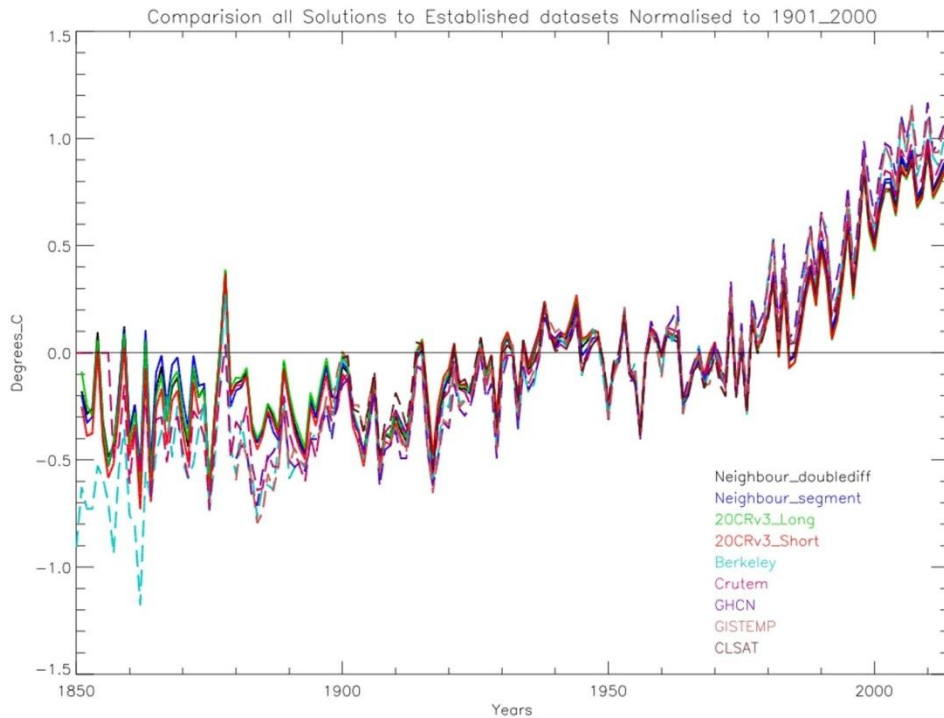


Figure 5.39 Top panel in a comparison of the established datasets of Berkeley, CRUTEMv5, GHCNMv4, GISTEMP, C-LSAT, with the 4 variants constructed herein: Neighbour_{double-diff}, Neighbour_{segment}, 20CRv3_{long}, and 20CRv3_{short}. All series have been normalised to 1901-2000 to try to highlight 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. Bottom panel is the same as the top panel but normalised to 1961-1990

Looking at trends over various periods (Table 5.8) highlights similarities and differences between the new estimates developed here and existing estimates. For the longest 1850 on period considered, trends and deltas comparisons can only be made to CRUTEMv5 and Berkeley Earth, both of which suggest considerably more warming than both the ISTI databank raw data and the new 20CRv3-based estimates. For trends starting in 1850 the effect is a reduction in long-term warming estimates of somewhere between 15 and 40% depending upon the choice of comparator product and which of the 4 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 4 remaining pre-existing estimates warm more than the raw ISTI databank holdings over 1901-2014 by about 10 to 25%, whereas the 4 new estimates all support CLSAT. For both 1851 on and 1901 on trends, the implied sign of required global aggregate adjustments is of opposite sign between the new estimates and all existing estimates bar CLSAT. 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 must result from some combination of the integrative effects of homogenisation uncertainty into the past and additional differences arising from station selection and post-processing choices.

Data Set	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	
GISTEMP		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-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	0.993

Table 5.8 Global trend analysis comparison with other Land Surface Air Temperature datasets

5.7 Discussion

The present analysis has undertaken an exploratory analysis of the application of sparse-input reanalyses to homogenise available land surface air temperature station records. The methods, based upon published approaches to radiosonde homogenisation using full input reanalysis (Haimberger et al., 2012), resulted in 4 estimates which share in common the breakpoint detection step but differ in the approaches used in the application of adjustments. 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% 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 standardised. The new exploratory estimates do not fundamentally alter existing conclusions that the global land surface air temperature has warmed or that this warming has accelerated over recent decades. However, if verified, they may have important implications for how close we are to Paris Agreement thresholds amongst other impacts.

Several steps, including further investigation of the early period records, would be required prior to the operationalisation of any products arising from the present analysis. Firstly, this analysis was performed as a PhD thesis and a thorough code review and refactorization would be advisable to ensure that no errors exist and to improve the processing efficiency. Secondly, further efforts to quantify and

understand the differences in the early period of record between the new set of estimates and prior published estimates would be required to be assured of the value of the present approach. Thirdly, to be complementary to existing products, most of which now come as ensembles, sampling parametric uncertainties, efforts would be required to build such an ensemble. Finally, to be operational a means of updating the analysis through present and then in a timely fashion thereafter would be required. In addition, subsequently, to be comparable to state-of-the-art approaches, efforts to create interpolated estimates would be required.

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 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 5.6. Station series from a greater range of these prior products should be compared to the new products developed here. Additional regional analysis should also be undertaken with an emphasis on locations, such as the Indian sub-continent, where the gridded trends appear to diverge either from one another or from GHCNMv4.

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 via co-variation of uncertain parameter choices would be consistent with state of the art approaches such as GHCNMv4 (Menne et al., 2018). Table 5.9 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 or otherwise. Of course, until such an ensemble were run these assessments cannot be verified.

The first set of choices pertain to the choice of sparse-input reanalysis and its interpolation. Chapter 4 justifies the selection of 20CRv3 sparse input reanalysis product as the best product to carry forward as the reference series. However, other reanalyses could have been used and, as the Chapter 4 analysis makes clear, this would have a large potential impact. How to interpolate the reanalysis to the station locations also involves a set of uncertain choices, but Chapter 3 analysis shows such

impacts to be relatively minor. In addition, it was opted subjectively to use the 9 nearest sparse input reanalysis gridded estimates, other choices could have been to use the 4, 16 or 25 nearest gridded values. Chapter 4 also considered the ensemble mean versus ensemble members, and that analysis justified the choice to use the ensemble mean in the present analysis. However, ensemble members could be used and may have some impact upon the resulting analysis.

Next comes Quality Control which is applied using a factor of three times the interquartile range of the difference series. For parametric uncertainty quantification, the critical value could be perturbed. However, equally the current quality control is relatively simple compared to state-of-the-art approaches and could be further developed and improved upon in future work.

For the breakpoint detection step there are a large number of choices that could be varied. Given that this step sets the number of adjustments to be applied the collective impact of these choices on the final output must be substantial irrespective of the adjustment step method being applied and choices therein. The most critical is probably the SNHT threshold score which, as noted in Section 5.2, is hard to pin down a robust basis for, and for which values between 12 and 20 appear reasonable that would cover roughly a 2-fold range in returned breakpoint location counts. Choices around: sample matching; when to force breaks for record cessation and resumption; sample size; window width and the required string of values above threshold requirements will also have potential impacts. But these are, at least individually, likely to be somewhat less important than the choice of SNHT threshold.

For the adjustments, there are relatively few choices that could have an impact for $20CRv3_{long}$ and $20CRv3_{short}$. For both approaches instead of using the difference between segment means the median or some other robust statistical estimate could be investigated. For $20CRv3_{short}$ the segment length could also be varied. Conversely, there are a fairly substantial number of choices that could be varied for the neighbour-based adjustment approaches. These pertain to the neighbour sample size, neighbour selection, number of estimates required, and what to default to when neighbours are unavailable. Another unique possible approach that could be considered is to produce a hybrid whereby selected neighbours could be combined

with the sparse input reanalysis. However, this possibility means reverting to a composite of neighbours reference series made up of neighbours that could be weighted by distances or correlation. In such a scenario a lot of consideration as to the weighting given to the sparse input reanalysis estimates would be required. This would be a substantive effort that is beyond the scope of this work. Collectively these choices probably have a high potential to impact the estimates resulting from the neighbour-based adjustments.

Finally, there are post-processing decisions that were made around the conversion to 1961-90 normalised series, their gridding and their averaging. Given the potential that spurious 20CRv3 behaviour could get aliased into the series, the choice to use 20CRv3 to adjust climatologies for stations with insufficient data over 1961-90 is probably the largest source of uncertainty in this set of post-processing steps. Future work could investigate alternative approaches such as using the differences to nearby stations as done in GHCNMv4.

Method step	Default choice	Other plausible choices	Expert judgement as to whether a primary control on long-term trends
Sparse-input reanalysis background field			
Choice of reanalysis	20CRv3	20CRv2 20CRv2C ERA-20C CERA-20C	High
20CRv3 Interpolation method	Inverse squared linear	Inverse linear, Kriging Spatial Averaging	Low
20CRv3 Interpolation distance	Use of 9 nearest grid boxes	Use of 4, 16 or 25 nearest grid boxes	Low
Use of ensemble mean or underlying ensemble members	Ensemble mean	Ensemble members	Medium
Quality control			

Quality control threshold	3*interquartile range	4 or 5 * interquartile range	Low
Breakpoint identification			
SNHT critical value	16	12-20	High
Matching samples before and after	Matching as in Haimberger et al.	Allow the number of observations before the break to differ from the number of observations after the break	Medium
Forced breakpoints for cessation and resumption (months) Forced	36 or greater	12- 36	Medium
Number of consecutive values to exceed the critical value	3	1,5,7	Medium
Test window width	60 months	36-120 months	Medium
Test sample minimum	20	15-30	Medium
Adjustments for 20CRv3 _{long} and 20CRv3 _{short}			
Difference in segment means	Means	Medians	Medium
Segment length in 20CRv3 _{short}	60 months	36-120 months	Medium
Adjustments for neighbour _{segments} and neighbour _{double-diff}			
Number of neighbours to consider	250	100-500	Medium
Neighbour geographical representation	Select from any quadrant	Each quadrant to be equally represented	Medium

Sufficient data in neighbours	20 observations within ± 5 years	12-36 observations	Medium
Number of valid neighbours required not to revert to 20CRv3 _{short}	1	2-25	High
Selection of best estimate from neighbours	Median	Mean Weighted Average	Medium
In the absence of a neighbour-based estimate revert to	20CRv3 _{short}	20CRv3 _{long} or choose not to apply any adjustment	High
Post-processing			
Calculation of climatology	Use stations where possible but revert to 20CRv3 to estimate elsewhere	Use solely stations and reduce number of stations accordingly	High
Gridding	Simple gridbox averaging	Weighted by distance from gridbox centre	Low
Creating area aggregate averages	Cos(lat) weighted mean	Zonally average then cos(lat) weight	Medium

Table 5.9 Possible sources of uncertainty which should be considered in the construction of a parametric uncertainty ensemble and an expert based estimation of their possible impact.

In lieu of a full parametric ensemble sensitivity of results to two plausible sources of high uncertainty identified in Table 5.9 have been investigated. The first is if the use of 20CRv3 to adjust the climatology for those stations which cannot derive a 1961-90 climatology estimate directly. This is assessed by simply gridding solely those stations for which the 1961-90 climatology can be calculated directly. The second is to use 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 5.10). The use of an SNHT threshold of 12 has a more considerable impact and would further increase

the difference to existing estimates. 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 symmetrical, could easily reconcile long-term warming rate estimates with many of the existing records. This highlights the critical importance of quantifying uncertainty in the new products prior to their application in an operational context. This comparison can be extended to other aspects such as gridded trend maps (Figure 5.40).

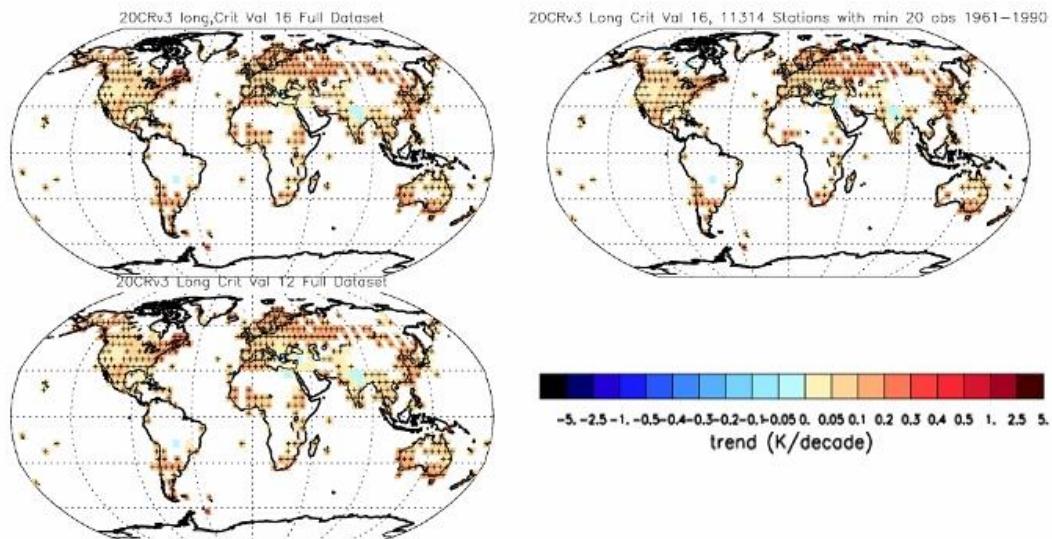


Figure 5.40 Trend analysis comparison for the period 1900 to 2014 using OLS trend estimation with AR(1) correction following Santer et al. (2008) for 20CRv3_{long} for the default version (top left), the same using only stations with 1961-90 stations (top right) and with an SNHT crit value of 12 (bottom left).

Data Set	OLS trends in °C per decade				Change 1851-1900 to 2005-2014
	1851-2014	1900-2014	1950-2014	1980-2014	
Default settings					
Neighbour _{double-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 to 1990 period					
Neighbour _{double-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-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

Table 5.10. Global trend analysis sensitivity assessment using three versions. Top set is those used in Section 5.5. The middle set is using the same settings but restricted solely to the subset of stations for which a 1961-90 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.

Finally, for an operational product, a means to update the series in a timely manner would be required. Although this could be achieved by simply appending new data reports, over time the risk of new breakpoints which had gone undetected would greatly increase. To enable operational updates would require as a pre-requisite either the provision of a sparse-input reanalysis product which was updated regularly or the use of a full-input reanalysis product to enable timely reassessments of recent station series homogeneity on an ongoing basis.

Following operationalisation, efforts to create a product version that undertook interpolation over reasonable distances would be advisable. There exist a range of such approaches now (Cowtan and Way, 2014, Rohde and Hausfather, 2020, Kadow et al., 2020). It would be necessary to test which, if any of these, worked for the present products. Naively, given the retention of as many stations as possible, all published approaches should be applicable. To take this work yet further it could then be merged with an existing SST or NMAT data product to create a truly global surface temperature estimate.

5.8 Conclusion

Using the 20CRv3 sparse-input reanalysis product an assessment of the ISTI databank has been undertaken to produce 4 novel exploratory 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 verify on this basis as potentially reasonable estimates. When compared to the full-range of published estimates of LSAT change the series broadly agree for recent changes but increasingly diverge for earlier records 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 land surface temperature series, further work is required to investigate why the divergence occurs between these estimates and the other land surface datasets. In addition, further work upon uncertainty quantification and a means to provide regular updates would be required prior to the ability to use such estimates in an operational context in future.

Chapter 6 Summary and discussion

6.1 Context

Typical state-of-the-art land surface air temperature products result from some form of neighbour-based comparisons to identify and adjust for non-climatic effects. This leaves open the possibility of common biases arising from a similarity in methodological basis. Neighbour-based approaches are not, however, the only possible solution. Haimberger (2006) introduced the use of innovations derived from radiosonde observations minus background forecasts from the data assimilation process used in the development of ERA-40 as a reference series for the homogenisation of radiosonde observations (RAOBCORE). Haimberger et al. (2012) expanded on this previous work to produce a radiosonde homogenised data set using combined comparison with reanalysis background series and neighbouring station series (RICH). This thesis set out to build on the work of Haimberger by investigating if the use of the most recent generation of sparse-input reanalysis products, which are independent of any land surface temperature observations and extend back as far as the 19th Century can be used as a reference series for the homogenisation of monthly land surface air temperatures. The thesis benefits from the International Surface Temperature Initiative's recently released databank which contains many more station records than any prior global data collection.

6.2 Key Findings

6.2.1 Assessment of the ISTI Databank

Chapter 3 undertook a substantive investigation of the suitability of the ISTI databank for the task. Initial examination of the ISTI databank uncovered an issue of duplication of data across neighbouring stations. Contact with NCEI revealed that they had uncovered 269 duplicate files during the development of GHCNMv4, a blacklist which they shared and was applied. Despite this, additional duplications were identified. This meant a full assessment of the databank was required before work could continue. This resulted in the identification of three main forms of duplication:

1. Full or almost full duplication of data within two or more stations with different names and coordinates requiring the full removal of one or more series.
2. Significant duplication of data, that could be repaired by the removal of one or more sources that made up the station series or by merging the stations.
3. Stations with different names but with identical coordinates. These stations usually contained distinct data. The series that contained the most observations was retained and the others discarded. The issue of identical coordinates with different names is predominantly associated with Canada.

In total an additional 563 stations were removed; 181 stations were merged into other series and 579 stations had one or more source data stream removed from the file resulting in a shortened, but now unique, series.

Researchers using the ISTI databank must be aware of these issues. Specifically, this problem is likely to be somewhat more extensive than was addressed here due to the limited criteria employed and time limitations.

6.2.2 Suitability of sparse-input reanalyses for reference series construction

State of the art homogenisation methods make use of pairwise comparisons with neighbouring series. Good pairs will have a high correlation between the individual neighbour and the candidate station and low standard deviation of the difference series (candidate-neighbour). However, the availability of suitable neighbouring stations for pairwise comparison is limited for many sparse locations and, in early records, almost everywhere. Modern sparse input reanalysis products continue to improve in quality as each new product builds on the experience and knowledge of past versions. Chapter 4 (also published as Gillespie et al. (2020)) found that the most recent 20CRv3 product offers the best opportunity for the homogenisation of land surface air temperature. While neighbour-based comparisons clearly remain preferable in data rich areas and periods, in the early epochs and in regions where stations are separated by 700km or more 20CRv3 is clearly potentially preferable to these traditional techniques. The 20CRv3 product shows less variation in performance between well-sampled and poorly-sampled regions than state-of-the-art pairwise approaches would do, which may also yield desirable properties in any resulting products.

Chapter 4 further considered the relative value of sparse-input reanalysis ensemble means versus individual ensemble members. The analysis of 20CRv2c suggests that the ensemble mean is better correlated and yielded the lowest standard deviation of the difference series. This is as would be expected given the ensemble design and theoretical expectations. However, this does not diminish the potential value of individual ensemble members for further analysis in future work or in other scientific contexts.

6.2.3 Homogenisation using sparse-input reanalysis products

Chapter 5 went on to apply 20CRv3 to perform an initial exploratory homogenisation of the ISTI databank. The breakpoint detection step involved passing the SNHT test through a 20CRv3-candidate station series difference series (where 20CRv3 had been interpolated to the station location) using a critical value of 16. Four adjustment options were then considered to adjust for these identified breaks. These borrowed heavily from the precursor work of Haimberger et al. (2012) and used 20CRv3 (in two variants) and neighbour segment characteristics (also in two variants) to apply adjustments:

- 20CRv3_{long} which used the entire 20CRv3-candidate series segments to adjust irrespective of length
- 20CRv3_{short} which was identical but truncated segments at 5 years if they were longer.
- Neighbour_{segments} which used the difference series to all available apparently homogeneous segments from the 250 nearest stations by great arc distance.
- Neighbour_{double-diffs} that instead uses the differences in 20CRv3-station differences (conceptually similar to double differencing techniques that have widespread usage in operational meteorological settings) but is otherwise identical to neighbour segments.

For the two neighbour-based approaches in the c.10% of cases where an insufficient set of neighbour-based estimates were available the adjustment estimate for 20CRv3_{short} was used rather than to leave the identified break be unadjusted.

The series were compared to the station series resulting from GHCNMv4 (Menne et al., 2018) which used the same ISTI databank source but applied a pairwise

homogenisation procedure. Comparisons were undertaken from the level of station series through gridbox anomalies and trends to various regional and global aggregations. Overall the series exhibited reasonable behaviour, but GHCNMv4 diverged and was cooler in the early period of record than the newly produced estimates at the global aggregate. This behaviour was dominated by a divergence in Northern Hemisphere series which arises principally in the Indian subcontinent and North America.

The new analyses produced here suggest less long-term warming globally than all existing published estimates except, perhaps, the recently published CLSAT product from China. Over the longest-term 1851-2014 period, when only a subset of published estimates are available, the effect is a 15-40% reduction depending upon the choice of diagnostic and comparator series. The differences rapidly diminish such that trends from 1900-2014 are broadly similar and since 1950 trends are indistinguishable from the family of existing published estimates.

6.3 Implications

The present analysis introduces a substantive methodological degree of freedom into the characterisation of global LSAT. The results in this thesis broadly reaffirm the findings from other studies of global LSAT changes that the world has warmed and that the overall rate of warming has accelerated since the mid-20th Century. This serves to overall increase confidence in the findings by IPCC and numerous other scientific bodies of unequivocal warming over the instrumental era.

The differences, particularly in the 19th Century, are marked with all four estimates showing warmer global temperatures than existing estimates for much, if not all, of the period 1851-1900. This result would imply, were the new estimates to prove reliable following further evaluation, that estimates of Global Mean Surface Temperature (GMST) change over the longest term may be biased toward excessive long-term warming. However, even with the improved coverage afforded by the ISTI databank, marine data coverage remains predominant prior to 1850 and thus the impact will not be as substantial on GMST estimates as it is on LSAT estimates shown in the present thesis. This has potential implications for numerous diagnostics

such as estimates of threshold crossing times and remaining carbon budgets that depend upon observed changes to date directly or indirectly.

The new series created herein also show systematically distinct warming patterns in certain regions and over certain periods. These differences are most marked prior to 1950 and over North America and the Indian sub-continent. In both these regions the sign and significance of trends is distinct between GHCNMv4 and some or all of the adjusted series. Whereas in other regions such as Eurasia the series are all in much closer agreement. The regions and epochs with large uncertainty are those which are data sparse. Whether this relates to limitations in state-of-the-art approaches such as GHCNMv4 or issues in the use of 20CRv3 to perform homogenisation remains uncertain and requires further investigation.

6.4 Limitations and possible future work

The present analysis was undertaken as a PhD project and was not intended to be exhaustive. It has offered an exploratory analysis of the potential for, and the possible implications of, the use of sparse-input reanalysis products to create an independent LSAT product or family of products. Significant further work including the further evaluation and the development of robust estimates of uncertainty would be required before the products developed could be considered operational and suitable for application in a policy context.

The use of reanalysis products in climate studies is well established (Chapter 2) and full-input reanalysis fields have been successfully applied by Haimberger (2006) and by Haimberger et al. (2012) in the homogenisation of radiosonde data. Building on previous work this analysis implies that modern sparse-input reanalysis products have the potential to augment other well-established homogenisation processes and to provide for the inclusion of stations in remote locales and early epochs that may otherwise be either excluded or unable to be adequately assessed for the presence of inhomogeneities for the lack of well correlated neighbours.

However, conversely, reanalysis products are vulnerable to inhomogeneities arising from incorrectly prescribed boundary conditions and time-varying assimilation streams. Unlike pairwise homogenisation, there is not any dilution benefit to be

accrued from other sources in such an event. While 20CRv3 exhibits correspondence at various aggregations (Chapter 4) it would be valuable to have more sparse-input reanalysis estimates produced independently and similarly spanning from the early 1800s to assess sensitivity to the choice of sparse-input reanalysis system. Although Chapter 4 showed that the sparse-input ensemble mean fields used in Chapter 5 had preferable statistical properties to underlying ensemble members, in the absence of alternative products it would be desirable to repeat the analysis using the 80-member ensemble. This may help to quantify uncertainty particularly locally. However, given the nature of the ensemble, it will not be able to address any systematic errors (that is any errors that may be persistent resulting from choices made when producing the ensemble) that may exist in the 20CRv3 product.

Parametric uncertainty could be quantified via an ensemble as discussed in Section 5.7 through perturbing the main parameters identified in Table 5.9. This ensemble could be produced by varying the identified parameters within reasonable ranges in multiple combinations so that any inter-dependencies could be fully expressed. In the absence of such an ensemble Table, 5.10 summarises adjustment of just two factors: i) creating estimates of the 11,314 stations that were able to be normalised to 1961-1990 normals without recourse to a 20CRv3 proxy climatology and ii) adopting a SNHT critical value of 12. The first shows little sensitivity, whereas the second led to 10-15% changes implying that a full exploration of uncertainty may be considerably larger and partially or fully reconcile estimates developed with published estimates and their respective uncertainties.

The present intercomparison to existing products was deliberately non-exhaustive. There exists considerable potential to further compare the 4 new estimates produced with existing products across a range of spatial and temporal aggregations. Such an intensive and systematic comparison would undoubtedly yield additional valuable insights. It is, however, a very substantial undertaking which would require considerable time and due care and attention and as such was deemed to be beyond the scope of the present thesis.

To be useful for monitoring and operational applications a means by which to update the products would be required. This should include the ability to detect and adjust

for more recent data inhomogeneities. This would require either the availability of a sparse-input reanalysis product updated in a timely fashion or to use e.g. ERA5 to assess recent series homogeneity.

Most modern products undertake some degree of spatial infilling over reasonable distances. The lower values in recent years in the estimates produced herein may relate to not undertaking such an interpolation as it is well documented (e.g. Simmons et al., 2010) that sampling effects can bias recent estimates, and recent updates to HadCRUT5 (Morice et al., 2021) have confirmed this to be the case for HadCRUT5. Published interpolation approaches would, in theory, be applicable to the data products here and applying such interpolation approaches in future work may reduce the discrepancies identified in Section 5.6.

The approaches developed herein are in principle applicable to a broader range of surface meteorological parameters than surface temperatures. Applying these techniques to other parameters may yield additional insights and new and novel products which may further inform our understanding of in particular pre-satellite era climate. Such an analysis may also serve to provide a suite of first guess bias-corrections which may prove important if future sparse-input reanalysis products were to ingest additional surface parameters than surface pressure.

The 20CRv3 product can provide daily and sub-daily comparisons to station data records. Future work could explore the potential application to assessments of homogeneity and the application of adjustments at such scales. The data volumes involved and the documented challenges in daily and sub-daily homogeneity assessments precluded such an effort here.

REFERENCES

- Aguilar, E., Auer, I., Brunet, M., Peterson, T. & Wieinga, J. 2003. Guidelines on climate metadata and homogenization. *World Meteorological Organisation*.
- Alexandersson, H. 1986. A homogeneity test applied to precipitation data. *International Journal of Climatology*, 6, 661-675.
- Alexandersson, H. & Moberg, A. 1996. Homogenization of Swedish temperature data part 1. Homogeneity test for linear trends. *International Journal of Climatology*, 17, 25-34.
- Allan, R., Brohan, P., Compo, G., Stone, R., Luterbacher, J. & Brönnimann, S. 2011. The international atmospheric circulation reconstructions over the earth (ACRE) initiative. *Bulletin of the American Meteorological Society*, 92, 1421-1425.
- Allison, L., Hawkins, E. & Woollings, T. 2014. An event-based approach to understanding decadal fluctuations in the Atlantic meridional overturning circulation. *Climate Dynamics*, 44, 163-190.
- Bell, S., Cornford, D. & Bastin, L. 2013. The state of automated amateur weather observations. *Weather*, 68, 36-41.
- Bell, S., Cornford, D. & Bastin, L. 2015. How good are citizen weather stations addressing a biased opinion. *Weather Observations*, 70, 75-84.
- Bessemoulin, P., Gutmann, N., Ikeoghan, I., Malone, L. & Sterin, A. 2018. Guide to climatological practice Third Edition. In: WMO (ed.). WMO Library: World Meteorological Organization.
- Blunden, J. & Arndt, D. 2019. State of the climate in 2018. *Bulletin of The American Meteorological Society*, 100, Si-S306.
- Bosilovich, M., Kennedy, J. J., Dee, D. & O, N., A. 2012. On the reprocessing and reanalysis of observations for climate. In: Asrar, G. & Hurrell, J. (eds.) *Climate Science for Serving Society*. European Center for Medium Range Forecasting Reading UK: Springer.
- Bowker, D. 2011. Meteorology and the ancient Greeks. *Weather*, 66, 249-251.
- Brandsma, T. & Van Der Meulen, J. 2008. Thermometer screen intercomparison in De Bilt (the Netherlands)—Part II: description and modelling of mean temperature differences and extremes. *International Journal of Climatology*, 28, 389-400.

- Brohan, P., Kennedy, J. J., Harris, I., Tett, S. & Jones, P. 2006. Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *Journal of Geophysical Research*, 111.
- Bronnimann, S., Martius, O., Franke, J., Stickler, A. & Auchmann, R. 2013. Historical weather extremes in the " Twentieth Century Reanalysis". In: Bronnimann, S. & Martius, O. (eds.) *Weather extremes during the past 140 years*. 2013 ed. University of Bern Switzerland: Geographica Bernensia.
- Bronnimann, S. 2015. Climatic changes since 1700. *Climatic changes since 1700*. Switzerland: Springer International Publishing.
- 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., Westcott, N. E., Williamson, F. & Wyszyński, P. 2019. Unlocking pre-1850 instrumental meteorological records: A global inventory. *Bulletin of the American Meteorological Society*, 100, ES389-ES413.
- Brunet, M., Saladié, O., Jones, P., Sigró, J., Aguilar, E., Moberg, A., Lister, D., Walther, A., Lopez, D. & Almarza, C. 2006. The development of a new dataset of Spanish daily adjusted temperature series (SDATS) (1850–2003). *International Journal of Climatology*, 26, 1777-1802.
- Brunet, M. & Jones, P. 2011. Data rescue initiatives: bringing historical climate data into the 21st century. *Climate Research*, 47, 29-40.
- Butler, M. 2018. Personal weather stations and sharing weather data via the Internet. *Weather*, 74, 22-29.
- Callendar, G. 1938. The artificial production of carbon dioxide and its influence on temperature. *Quarterly journal of the Royal Meteorological Society*, 64, 223-240.
- Callendar, G. 1961. Temperature fluctuations and trends over the earth. *Quarterly Journal of the Royal Meteorological Society*, 87, 1-12.
- Camuffo, D. & Jones, P. 2002. Errors in early temperature series arising from changes in style of measuring time, sampling schedule and number of observations. *Climatic Change*, 52, 331-352.

- Camuffo, D. & Bertolin, C. 2011. The earliest temperature observations in the world: the Medici network (1654–1670). *Climatic Change*, 111, 335-363.
- Caussinus, H. & Mestre, O. 2004. Detection and correction of artificial shifts in climate series. *Applies Statistic*, 53, 405-425.
- Changnon, S. & Kunkel, K. 2005. Changes in instruments and sites affecting historical weather records. A case study. *Journal of Atmospheric and Oceanic Technology*, 23, 825-828.
- Childs, C. 2004. Interpolating surface in arcgis spatial analysis. *ArcUser ERSI Education Service*, 32-35.
- Chun, B. & Guhathakurta, S. 2016. Daytime and nighttime urban heat islands statistical models for Atlanta. *Environment and Planning B: Urban Analytics and City Science*, 44, 308-327.
- Compo, G., Whitaker, J. & Sardeshkukh, P. 2006. Feasibility of a 100 year reanalysis using only surface pressure data. *American Meteorological Society*, 87, 175-190.
- Compo, G., Whitaker, J. & Sardeshmukh, P. 2008. 20th century reanalysis project. *Proceedings of the Third WCRP International Conference on Reanalysis*. The University of Tokyo, Japan. : University of Colorado/CIRES Climate Diagnostics Center.
- 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., Mauerer, M., Mok, H. Y., Nordli, Ø., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D. & Worley, S. J. 2011. The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, 137, 1-28.
- Compo, G., Sardeshmukh, P., Whitaker, J., Brohan, P., Jones, J. & 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., Sardeshmukh, P., Giese, B., Brohan, P. & Slivinski, L. 2016. 20 th century reanalysis version "2c" (1851-2012) and prospects for 200 years of reanalysis. *American Geophysical Union fall meeting 2016*. Washington , DC: American Geophysical Union.
- Conrad, V. 1946. *Methods in climatology*, United States of America, Harvard University Press.

- Cornes, R., Jones, P., Briffa, K. & Osborn, T. 2012. A daily series of mean sea-level pressure for London, 1692-2007. *International Journal of Climatology*, 32, 641-656.
- Cornes, R., Van Der Schrier, G., Van Den Besselaar, E. & Jones, P. 2018. An ensemble version of the E-OBS temperature and precipitation data sets. *Journal of Geophysical Research: Atmospheres*, 123, 9391-9409.
- Costa, A. & Soares, A. 2008. Homogenization of climate data: review and new perspectives using geostatistics. *Mathematical Geosciences*, 41, 291-305.
- Cowan, K. & Way, R. 2014. Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quarterly Journal of the Royal Meteorological Society*, 140, 1935-1944.
- Cowan, K., Jacobs, P., Thorne, P. & Wilkinson, R. 2018. Statistical analysis of coverage error in simple global temperature estimators. *Dynamics and Statistics of the Climate System*, 3, 1-18.
- Cram, T., Compo, G., Yin, X., Allen, R., Mccoll, C., Vose, R., Whitaker, J., Matsui, N., Ashcroft, L., Auchmann, R., Bessemoulin, P., Brandsma, T., Brohan, P., Brunet, M., Comeaux, J., Crouthamel, R., Brandsma, T., Brohan, P., Brunet, M., Comeaux, J., Crouthamel, R., Gleason, V., Groisman, P., Hersbach, H., Jones, P., Jonssen, T., Jourdain, S., Kelly, G., Knapp, K., Kruger, A., Kubota, H., Lentini, G., Lorrey, A., Lott, N., Lubker, S., Luterbacher, J., Marshall, G., Maugeri, M., Mock, J., Mok, O., Nordi, M., Rodwell, M., Ross, T., Schuster, D., Srncic, L., Valente, M., Vizi, Z., Wan, X., Westcott, N., Wollen, J. & Worley, S. 2015. The international surface pressure databank version 2. *Geoscience Data Journal*, 2, 31-46.
- Curry, J. 2011. Reasoning about climate uncertainty. *Climatic Change*, 108, 723-732.
- Dai, A., Wang, J., Thorne, P. W., Parker, D. E., Haimberger, L. & Wang, X. L. 2011. A new approach to homogenize daily radiosonde humidity data. *Journal of Climate*, 24, 965-991.
- Dee, D., Källén, E., Simmons, A. & Haimberger, L. 2011a. Comments on "reanalyses suitable for characterizing long-term trends". *Bulletin of the American Meteorological Society*, 92, 65-70.
- Dee, D., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A., Van De Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, I., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B., Morcrette, J., Park, B. K., Peubey, C., De Rosnay, P., Tavolato, C., Thépaut,

- J. & Vitart, F. 2011b. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137, 553-597.
- Degaetano, A. 2005. Attributes of several methods for detecting discontinuities in mean temperature series. *Journal of Climate*, 19, 838-853.
- Demaree, G., Lachaert, P., Verhoeve, T. & Thoen, E. 2002. The long-term daily central Belgium temperature series 1767-1998 and early instrumental meteorological observations in Belgium. *Climatic Change*, 53, 269-293.
- Diamond, H., Karl, T., Palecki, M., Baker, C., Bell, J., Leeper, R., Easterling, D., Lawrimore, J. H., Meyers, T., Helfert, M., Goodge, G. & Thorne, P. W. 2013. U.S. climate reference network after one decade of operations: status and assessment. *Bulletin of the American Meteorological Society*, 94, 485-498.
- Domonkos, P. & Stepanak, P. 2009. Statistical characteristics of detectable inhomogeneities in observed meteorological time series. *Stud Geophysica Geodaetica*, 53, 239-260.
- Domonkos, P. 2011. Efficiency evaluation for detecting inhomogeneities by objective homogenisation methods. *Theoretical and Applied Climatology*, 105, 455-467.
- Domonkos, P., Venema, V. & Mestre, O. Efficiencies of homogenisation methods: our present knowledge and its limitations. 7th Seminar for homogenization and quality control in climatological databases, 2012 Budapest Hungary.
- Domonkos, P. & Coll, J. 2017. Homogenisation of temperature and precipitation time series with ACMANT3: method description and efficiency tests. *International Journal of Climatology*, 37, 1910-1921.
- Ducré-Robitaille, J.-F., Vincent, L. A. & Boulet, G. 2003. Comparison of techniques for detection of discontinuities in temperature series. *International Journal of Climatology*, 23, 1087-1101.
- Durre, I., Menne, M. J., Gleason, B. E., Houston, T. G. & Vose, R. S. 2010. Comprehensive automated quality assurance of daily surface observations. *Journal of Applied Meteorology and Climatology*, 49, 1615-1633.
- Eden, P. 2009. Traditional weather observing in the UK :An historical overview. *Weather*, 64, 239-245.
- Fenner, D., Meier, F., Bechtel, B., Otto, M. & Scherer, D. 2017. Intra and inter 'local climate zone' variability of air temperature as observed by crowdsourced

citizen weather stations in Berlin, Germany. *Meteorologische Zeitschrift*, 26, 525-547.

- Ferguson, C. & Villarini, G. 2012. Detecting inhomogeneities in the twentieth century reanalysis over the central United States. *Journal of Geophysical Research*, 117, 1-11.
- Folland, C. K., Rayner, N. A., Brown, S. J., Smith, T. M., Shen, S. S. P., Parker, D. E., Macadam, I., Jones, P. D., Jones, R. N., Nicholls, N. & Sexton, D. M. H. 2001. Global temperature change and its uncertainties since 1861. *Geophysical Research Letters*, 28, 2621-2624.
- Freitas, L., Pereira, M., Caramelo, L., Mendes, M. & Nunes, L. 2013. Homogeneity of monthly air temperature in Portugal with HOMOR and MASH. *Quarterly Journal of the Hungarian Meteorological Service*, 117, 69-90.
- Fujibe, F. 2009. Detection of urban warming in recent temperature trends in Japan. *International Journal of Climatology*, 29, 1811-1822.
- Gelaro, R., Mccarty, W., Suarez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., Da Silva, A., Gu, W., Kim, G. K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M. & Zhao, B. 2017. The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of Climate*, 30, 5419-5454.
- Giese, B. S., Seidel, H. F., Compo, G. P. & Sardesmukh, P. D. 2016. An ensemble of ocean reanalyses for 1815–2013 with sparse input. *Journal of Geophysical Research: Oceans*, 121, 6891-6910.
- Gillespie, I. M., Haimberger, L., Compo, G. & 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.
- Giorgi, F. & Francisco, R. 2000. Evaluating uncertainties in the prediction of regional Climate. *Geophysical Research letters*, 27, 1295-1298.
- Gubler, S., Hunziker, S., Begert, M., Croci-Maspoli, M., Konzelmann, T., Brönnimann, S., Schwierz, C., Oria, C. & Rosas, G. 2017. The influence of station density on climate data homogenization. *International Journal of Climatology*, 37, 4670-4683.

- Guijarro, J. A. 2014. Climatol-guide. A R contributed package for homogenization of climatological series. State Meteorological Agency Spain: State Meteorological Agency(AEMET).
- Guttman, N. 1998. Homogeneity, data adjustments and climatic normals. Asheville United States of America: National Climatic Data Center
- Haimberger, L. 2006. Homogenization of radiosonde temperature time series using innovation statistics. *Journal of Climate*, 20, 1377-1403.
- Haimberger, L., Tavalato, C. & Sperka, S. 2008. Toward elimination of the warm bias in historic radiosonde temperature records : Some new results from a comprehensive intercomparison of upper-air data. *Journal of Climate*, 21, 4587-4606.
- Haimberger, L., Tavalato, C. & 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.
- Hansen, J. & Lebedeff, S. 1987. Global trends of measured surface air temperature. *Journal of Geophysical Research*, 92, 13,345-13,372.
- Hansen, J., Ruedy, R., Glascoe, J. & Sato, M. 1999. GISS analysis of surface temperature change. *Journal of Geophysical Research: Atmospheres*, 104, 30997-31022.
- Hansen, J., Ruedy, R., Sato, M. & Lo, K. 2010. Global surface temperature change. *Reviews of Geophysics*, 1-52.
- Hansen-Bauer, I. & Forland, E. J. 1994. Homogenizing long Norwegian precipitation series. *Journal of Climate*, 7, 1001-1013.
- Hartmann, D., Klein Tank, A., M.G, Rusticucci, M., Alexander, L. V., Bronnimann, S., Charabi, Y. a.-R., Dentener, F. J., Dlugokencky, E. J., Easterling, D. R., Kaplan, A., Soden, B., Thorne, P. W., Wild, M. & Zhai, P. 2013. Observations: atmosphere and surface in climate change: The physical science basis. Contribution of working group 1 to the fifth assessment report of the Intergovernmental Panel on Climate Change. *In: Stocker, T. F., Qin, D., G, K., Plattner,, Tignor, M., S.K. Allen, S. K., Boschung, j., Nauels, a., Xia, Y., Bex & Midgley, P. M. (eds.)*. Cambridge University Press UK & USA.
- Hausfather, Z., Cowtan, K., Menne, M. & 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.

- Hawkins, E. & Jones, P. D. 2013. On increasing global temperatures: 75 years after Callendar. *Quarterly Journal of the Royal Meteorological Society*, 139, 1961-1963.
- Hawkins, E. & Sutton, R. 2016. Connecting climate model projections of global temperature change with the real world. *Bulletin of the American Meteorological Society*, 97, 963-980.
- Hawkins, E., Ortega, P., Suckling, E., Schurer, A., Hegerl, G., Jones, P., Joshi, M., Osborn, T. J., Masson-Delmotte, V., Mignot, J., Thorne, P. & Van Oldenborgh, G. J. 2017. Estimating changes in global temperature since the preindustrial period. *Bulletin of the American Meteorological Society*, 98, 1841-1856.
- Hawkins, E., Burt, S., Brohan, P., Lockwood, M., Richardson, H., Roy, M. & Thomas, S. 2019. Hourly weather observations from the Scottish highlands (1883–1904) rescued by volunteer citizen scientists. *Geoscience Data Journal*, 6, 160-173.
- Hellmann, G. 1908. The Dawn of Meteorology. *Nature*, 79, 173-176.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horanyi, A., Sabater, J. M., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J. E., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Holm, E., Janiskova, M., Keeley, S., Laloyaux, P., Lopez, P., Radnoti, G., De Rosnay, P., Rozum, I., Vamborg, F., Villaume, S. & Thepaut, J.-N. 2020. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146, 1999-2049.
- Hirahara, S., Ishii, M. & Fukuda, Y. 2014. Centennial-Scale Sea Surface Temperature Analysis and Its Uncertainty. *Journal of Climate*, 27, 57-75.
- Hoppen, K., Theodore 1976. The nature of the early Royal Society. Part I. *The British Journal for the History of Science*, 9, 1_24.
- 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. & Brönnimann, S. 2017. Identifying, attributing, and overcoming common data quality issues of manned station observations. *International Journal of Climatology*, 37, 4131-4145.
- Jones, P., Wigley, T. M. L. & Kelly, P. M. 1982. Variations in surface air temperatures. Part 1 Northern Hemisphere, 1881-1980. *Monthly Weather Review*, 110, 59-70.

- Jones, P., Raper, S. C. B., Bradley, R. S., Diaz, H. F., Kelly, P. M. T. M. & Wigley, L. 1985a. Northern Hemisphere surface air temperature variations 1851 to 1984. *Journal of Applied Meteorology and Climatology*, 25 161-179.
- Jones, P., Raper, S. C. B., Santer, B., Bradley, R. S. & Diaz, H. F. 1985b. A grid-point-surface air temperature data set for Northern Hemisphere. Office of Energy Research Carbon Dioxide Research Division Washington D.C. 20545: United States Department of Energy.
- Jones, P., Raper, S. C. B., Goodess, C. M. & Wigley, B. S. G. 1986. A grid point surface air temperature data set for the Southern Hemisphere 1851-1984. USA: US Dept of Energy.
- Jones, P. & Briffa, K. 1992. Global surface air temperature variations during the twentieth century Part 1, spatial, temporal and seasonal details. *The Holocene*, 2, 165-179.
- Jones, P., Osborn, T. J. & Briffa, K. R. 1997. Estimating sampling errors in large scale temperature averages. *Journal of Climate*, 10, 2548-2568.
- Jones, P., Lister, D. H. & Li, Q. 2008. Urbanization effects in large-scale temperature records, with an emphasis on China. *Journal of Geophysical Research*, 113, 1-12.
- Jones, P., Lister, D. H., Osborn, T. J., Harpam, C., Salmon, M. & Morice, C. P. 2012. Hemispheric and large scale land surface air temperatures: An extensive revision and an update to 2010. *Journal of Geophysical Research*, 117, 1-29.
- Kadow, C., Hall, D. M. & Ulbrich, U. 2020. Artificial intelligence reconstructs missing climate information. *Nature Geoscience*, 13, 408-413.
- Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., M, I., Saha, S., White, G. H., Woolen, G., Zhu, Z., Chelliah, M., Ebisuzaki, M., Higgins, W., Janowlak, J., Mo, K. C., Ropelewski, C., J, W., Leetmaa, A., Reynolds, R. W., Jenne, R. & Joseph, D. 1995. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77, 437-471.
- Kanamitsu, M., Alpert, J. C., Campana, K. A., Deaven, D. G., Iredell, M., Katz, B., Pan, H., L, Sela, J. & White, G. H. 1991. Recent changes implemented into the global forecast system at NMC. *Weather and Forecasting*, 6, 425-435.
- Karl, T. R. & Williams, C., Jr 1987. An approach to adjusting climatological time series for discontinuous inhomogeneities. *Journal of Climate and Applied Meteorology*, 26, 1744 -1763.

- Kincer, J. B. 1933. Is Our Climate Changeing_A Study of Long-Time Tempature Trends. *Monthly Weater Review*, 61, 251-259.
- Kistler, R., Collins, W., Saha, S., White, G., Woollen, J., Kalnay, E., Chelliah, M., Ebisuzaki, W., Kanamitsu, M., Kousky, V., Van Den Dool, H., Jenne, R. & Fiorino, M. 2001. The NCEP–NCAR 50–year reanalysis: monthly means CD–ROM and documentation. *Bulletin of the American Meteorological Society*, 82, 247-267.
- Klingbjjer, P. & Moberg, A. 2003. A composite monthly temperature record from Tornedalen in northern Sweden, 1802-2002. *International Journal of Climatology*, 23, 1465-1494.
- Knight, J. R., Folland, C. K. & Scaife, A. A. 2006. Climate impacts of the Atlantic Multidecadal Oscillation. *Geophysical Research Letters*, 33, 1-4.
- Kobayashi, C., Endo, H., Ota, Y., Kobayashi, S., Onoda, H., Harada, Y., Onogi, K. & Kamahori, H. 2014. Preliminary results of the JRA-55C, an atmospheric reanalysis assimilating conventional observations only. *Sola*, 10, 78-82.
- Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K. & Takahashi, K. 2015. The JRA-55 reanalysis: general specifications and basic characteristics. *Journal of the Meteorological Society of Japan. Ser. II*, 93, 5-48.
- Kunkel, K., Liang, X.-Z., Zhu, J. & Lin, Y. 2005. Can CGCMs simulate the twentieth-century “Warming Hole” in the central United States. *Journal of Climate*, 19, 4137-4153.
- Lackner, B. C., Steiner, A. K., Hegerl, G. C. & Kirchengast, G. 2011. Atmospheric climate change detection by radio occultation data using a fingerprinting method. *Journal of Climate*, 24, 5275-5291.
- Ladstadter, F., Steiner, A. K., Foelsche, U., Haimberger, L., Tavolato, C. & Kirchengast, G. 2011. An assessment of differences in lower stratospheric temperature records from (A)MSU, radiosondes, and GPS radio occultation. *Atmospheric Measurement Techniques Discussions*, 4, 2127-2159.
- Lagouvardos, K., Kotroni, V., Bezes, A., Koletsis, I., Kopania, T., Lykoudis, S., Mazarakis, N., Papagiannaki, K. & Vougioukas, S. 2017. The automatic weather stations NOANN network of the National Observatory of Athens: operation and database. *Geoscience Data Journal*, 4, 4-16.
- Laloyaux, P., De Boiieson, E., Magdalena, B., Jean-Raymond, B., Stefan, B., Roberto, B., Per, D., Dee, D., Leopold, H., Hans, H., Yuki, K., Matthew, M., Paul, P., Nick, R., Elke, R. & Dinan, S. 2018. CERA 20C A coupled

reanalysis of the twentieth century. *Journal of Advances in Modelling Earth Systems*, 10, 1172-1195.

- Lanzante, J. & Klein, S. 2003. Temporal homogenization of monthly radiosonde temperature data. Part II trends. *Journal of Climate*, 16, 241-262.
- Lawrimore, J., Menne, M. J., Gleason, B. E., Williams, C. N., Wertz, D. B., Vose, R. S. & Rennie, J. 2011. An overview of the global historical climatology network monthly mean temperature data set, version 3. *Journal of Geophysical Research*, 116, 1-18.
- Lawrimore, J., Rennie, J. & Thorne, P. W. 2015. Responding to the need for better global temperature data. *Earth and Space science news*, 94, 61-62.
- Le Treut, H. R., Solomon, S., D, Qin, M., Manning, Z., Marquis, M., Averyt, K., B, Tignor, M. & Miller, H., L 2007. Historical overview of climate change. in: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group 1 to the Fourth assessment report of the Intergovernmental Panel on climate Change*. New York.
- Lee-Di 1978. On the invention of the meteorological instruments in ancient China. *Chinese Journal of Atmospheric Science*, 2, 85-88.
- Lenssen, N. J. L., Schmidt, G. A., Hansen, J. E., Menne, M., Persson, A., Ruedy, R. & Zyss, D. 2019. Improvements in the GISTEMP uncertainty model. *JGR Atmospheres*, 124, 6307-6326.
- Liang, C., Lixin, G., Guiping, F., Xuerui, W., Yang, Z. & Yu, Z. 2018. Comparison of the Arctic Upper-air temperatures from radiosonde and radio occultation observations. *Acta Oceanol Sin*, 37, 30-39.
- Mamara, A., Argiriou, A. A. & Anadranistakis, M. 2012. Homogenization of mean monthly temperature time series of Greece. *International Journal of Climatology*, 33, 2649-2666.
- Manley, G. 1974. Central England temperatures monthly means 1659 to 1973. *Quarterly Journal of the Royal Meteorological Society*, 100, 389-405.
- Mascioli, N. R., Previdi, M., Fiore, A. M. & Ting, M. 2017. Timing and seasonality of the United States 'warming hole'. *Environmental Research Letters*, 12, 034008.
- Mccoll, C. & Compo, G. 2021. Timeseries. TMP2m 1981-2000. University of Colorado/CIRES and NOAA physical Science Laboratory: University of Colorado/CIRES and NOAA physical Science Laboratory.

- Meier, F., Fenner, D., Grassmann, T., Otto, M. & Scherer, D. 2017. Crowdsourcing air temperature from citizen weather stations for urban climate research. *Urban Climate*, 19, 170-191.
- Menne, M. & Williams, C. 2009. Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, 22, 1700-1717.
- Menne, M., Williams, C. & Vose, R. 2009. The US historical climatology network monthly temperature data version 2. *Bulletin of the American Meteorological Society*, 90, 993-1008.
- Menne, M., Williams, C. N. & Palecki, M. A. 2010. On the reliability of the U.S. surface temperature record. *Journal of Geophysical Research*, 115, 1-9.
- Menne, M., Durre, I., Vose, R. S., Gleason, B. E. & Houston, T. G. 2012. An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, 29, 897-910.
- Menne, M., Williams, C. N., Gleason, B. E., Rennie, J. J. & Lawrimore, J. H. 2018. The global historical climatology network monthly temperature dataset, version 4. *Journal of Climate*, 31, 9835-9854.
- Milewska, E. & Hogg, W. D. 2010. Continuity of climatological observations with automation - temperature and precipitation amounts from AWOS (Automated Weather Observing System). *Atmosphere-Ocean*, 40, 333-359.
- Miller, R. 1974. The jackknife. A Review. *Biometrika*, 61, 1 - 15.
- Mitchell, D. M., Thorne, P. W., Stott, P. A. & Gray, L. J. 2013. Revisiting the controversial issue of tropical tropospheric temperature trends. *Geophysical Research Letters*, 40, 2801-2806.
- Moberg, A. & Alexandersson, H. 1996. Homogenization of Swedish temperature data Part ii. Homogenized gridded air temperature compared with a subset of global gridded air temperature since 1861. *International Journal of Climatology*, 17, 35-54.
- Modise, W. & Mphahlele, K. 2018. Weather forecasting: From the early weather wizards to modern-day weather predictions. *Journal of Climatology & Weather Forecasting*, 06.
- Morales, C., Ortega, M., Labajo, J. & Piorno, A. 2005. Recent trends and temporal behavior of thermal variables in a region of Castilla-Leon (Spain). *Atmosfera*, 18, 71-90.

- Morice, C. P., Kennedy, J. J., Rayner, N. A. & Jones, P. D. 2012. Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research: Atmospheres*, 117, 1-22.
- 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. & Simpson, I. R. 2021. An updated assessment of near-surface temperature change from 1850: The HadCRUT5 data set. *Journal of Geophysical Research: Atmospheres*, 126, 1-28.
- Murphy, A. 1998. The Early History of Probability Forecasts Some Extensions and Clarifications. *American Meteorological Society*, 13, 5-15.
- Naylor, S. 2018. Thermometer screens and the geographies of uniformity in nineteenth-century meteorology. *Notes and Records: the Royal Society Journal of the History of Science*, 73, 203-221.
- New, M., Hulme, M. & Jones, P. 1998. Representing twentieth century space time climate Part I development of a 1961-90 mean monthly terrestrial climatology. *Journal of Climate* 12, 829-856.
- New, M., Hulme, M. & Jones, P. 1999. Representing twentieth century space time climate variability Part II. Development of 1901-96 Monthly grids of terrestrial surface climate. *Journal of Climate*, 13, 2217-2237.
- Nicholson, S. & Flohn, H. 1980. African environmental and climatic changes and the general atmospheric circulation in the late pleistocene and holocene. *Climate Research*, 2, 313-348.
- Onogi, K., Tsutsui, J., Koide, H., Sakamoto, M., Kobayashi, N., Hatsushika, H., Takahashi, K., Kadokura, S., Wada, K., Oyama, R., Ose, T., Mannoju, N. & Taira, R. 2007. The JRA-25 reanalysis. *Journal of the Meteorological Society of Japan*, 85, 369-432.
- Osborn, T. J., Jones, P. D., Lister, D. H., Simpson, I. R. & Harris, I. 2020. Land Surface Air Temperature Variations Across the Globe Updated to 2019 the CRUTEM 5 Data Set. *Journal of geophysical Research (JGR)*, 126, 1-22.
- Pan, Z., Arritt, R. W., Takle, E. S., Gutowski, W. J., Anderson, C. J. & Segal, M. 2004. Altered hydrologic feedback in a warming climate introduces a “warming hole”. *Geophysical Research Letters*, 31, 1-4.
- Parker, D., Legg, T. & Folland, C. K. 1992. A new daily central England temperature series 1772-1991. *International Journal of Climatology*, 12, 317-342.

- Parker, D. 1994. Effects of changes of exposure of thermometers at land stations. *International Journal of Climatology*, 14, 1-31.
- Parker, D. E., Gordon, M., Cullum, D. P. N., Sexton, D. M. H., Folland, C. K. & Rayner, N. 1997. A new global gridded radiosonde temperature data base and recent temperature trends. *Geophysical Research Letters*, 24, 1499-1502.
- Parker, D. E. 2004. Large scale warming is not urban. *Nature*, 432, 291-291.
- Parker, D. E. 2011. Recent land surface air temperature trends assessed using the 20th century reanalysis. *Journal of Geophysical Research*, 116, 1-6.
- Parker, W. 2016. Reanalyses and observations: what's the difference ? *Bulletin of the American Meteorological Society*, 97, 1565-1572.
- Peterson, T. & Easterling, D. R. 1994. Creation of homogeneous composite climatological reference series. *International Journal of Climatology*, 14, 671-679.
- Peterson, T. & Easterling, D. 1995. A new method for detecting undocumented discontinuities in climatological time series. *International Journal of Climatology*, 15, 369-377.
- Peterson, T. & Vose, R. 1997. An overview of the global historical climatology network temperature database. *Bulletin of the American Meteorological Society*, 78, 2837-2849.
- Peterson, T., Easterling, D. R., Karl, T. R., Groisman, P., Nicholls, N., Plummer, N., Torok, S., Auer, I., Boehm, R., Gullett, D., Vincent, L., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T., Salinger, J., Forland, E. J., Hanssen-Bauer, I., Alexanderson, H., Jones, P. & Parker, D. 1998. Homogeneity adjustment of in situ atmospheric climate data a review. *International Journal of Climatology*, 18, 1493-1517.
- Peterson, T. 2003. Assessment of urban versus rural In situ surface temperatures in the contiguous United States no difference found. *Journal of Climate*, 16, 2941-2959.
- Peterson, T. & Owen, T. W. 2005. Urban heat island assessment. Metadata are important. *Journal of Climate*, 18, 2637-2646.
- Poli, P., Hersbach, H., Dee, D. P., Berrisford, P., Simmons, A. J., Vitart, F., Laloyaux, P., Tan, D. G. H., Peubey, C., Thépaut, J.-N., Trémolet, Y., Hólm, E. V., Bonavita, M., Isaksen, L. & Fisher, M. 2016. ERA-20C: An atmospheric reanalysis of the twentieth century. *Journal of Climate*, 29, 4083-4097.

- Quayle, R., Easterling, D., Karl, T. & Hughes, P. 1991. Effects of Recent Thermometer changes in the cooperative station network. *Bulletin of the American meteorological Society*, 72, 1718-1724.
- Quenouille, M. H. 1949. Approximate tests of correlation in time-series. *Journal of Royal Statistical Society. Series B(Methodological)*, 11, 68-84.
- Rao, Y., Liang, S. & Yu, Y. 2018. Land surface air temperature data are considerably different among BEST-LAND, CRU-TEM4v, NASA-GISS, and NOAA-NCEI. *Journal of Geophysical Research: Atmospheres*, 123, 5881-5900.
- Rayner, N., Brohan, P., Parker, D., Folland, C. K., Kennedy, J. J., Vanicek, M., Ansell, T. J. & Tett, S. F. B. 2005. Improved analyses of changes and uncertainties in sea surface temperature measured In situ since the mid-nineteenth century the HadSST2 dataset. *Journal of Climate*, 19, 446-469.
- 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., Rusticucci, M., Tandy, J., Worlry, S. J., Venma, V., W. Angel, W., Brunet, M., Dattore, B., Diamond, H., Lazzara, M. A., Blancq, F. L., Luterbacher, J., Machel, H., Revadekar, J., Vose, R. S. & 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.
- Rennie, J. J. 2015. Technical report international surface temperature initiative global land surface databank version 1.1.0. <http://www.surface temperatures.org/system/app/pages/search?scope=search-site&q=version+1.1>: National Oceanic and Atmospheric Administration.
- Ribeiro, S., Caineta, J. & Costa, A. C. 2016. Review and discussion of homogenisation methods for climate data. *Physics and Chemistry of the Earth, Parts A/B/C*, 94, 167-179.
- Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., Bosilovich, M. G., Schubert, S. D., Takacs, L., Kim, G.-K., Bloom, S., Chen, J., Collins, D., Conaty, A., Da Silva, A., Gu, W., Joiner, J., Koster, R. D., Lucchesi, R., Molod, A., Owens, T., Pawson, S., Pegion, P., Redder, C. R., Reichle, R., Robertson, F. R., Ruddick, A. G., Sienkiewicz, M. & Woollen, J. 2011. MERRA: NASA's modern-era retrospective analysis for research and applications. *Journal of Climate*, 24, 3624-3648.
- Rohde, R., A. Muller, R., Jacobsen, R., Muller, E. & Wickham, C. 2013a. A new estimate of the average earth surface land temperature spanning 1753 to 2011. *Geoinformatics & Geostatistics: An Overview*, 1.

- Rohde, R., Muller, R., Jacobsen, R., Perlmutter, S. & Mosher, S. 2013b. Berkeley earth temperature averaging process. *Geoinformatics & Geostatistics: An Overview*, 1.
- Rohde, R. A. & Hausfather, Z. 2020. The Berkeley earth land/ocean temperature record. *Earth System Science Data*, 12, 3469-3479.
- Ryan, C., Duffy, C., Broderick, C., Thorne, P. W., Curley, M., Walsh, S., Daly, C., Treanor, M. & Murphy, C. 2018. Integrating data rescue into the classroom. *Bulletin of the American Meteorological Society*, 99, 1757-1764.
- 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. & Goldberg, M. 2010. The NCEP climate forecast system reanalysis. *Bulletin of the American Meteorological Society*, 91, 1015-1058.
- Sahin, S. & Cigizoglu, H. K. 2010. Homogeneity analysis of Turkish meteorological data set. *Hydrological Processes*, 24, 981-992.
- Simmons, A. J., Willett, K. M., Jones, P. D., Thorne, P. W. & Dee, D. P. 2010. Low-frequency variations in surface atmospheric humidity, temperature, and precipitation: Inferences from reanalyses and monthly gridded observational data sets. *Journal of Geophysical Research*, 115, 1-21.
- Simmons, A. J., Berrisford, P., Dee, D. P., Hersbach, H., Hirahara, S. & Thépaut, J. N. 2017. A reassessment of temperature variations and trends from global reanalyses and monthly surface climatological datasets. *Quarterly Journal of the Royal Meteorological Society*, 143, 101-119.
- Slivinski, L. 2018. Historical reanalysis what how and why. *Journal of Advances in Modeling Earth Systems*, 10, 1736-1739.
- 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. & 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. & Wyszyński, P. 2021. An Evaluation of the Performance of the Twentieth Century Reanalysis Version 3. *Journal of Climate*, 34, 1417-1438.

Slonosky, V. C., Jones, P. D. & Davies, T. D. 1999. Homogenization techniques for european monthly mean surface pressure series. *Journal of Climate*, 12, 2658-2672.

Smith, A., Lott, N. & Vose, R. 2011. The integrated surface database: Recent developments and partnerships. *Bulletin of the American Meteorological Society*, 92, 704-708.

Sparks, W. 1972. The effects of Thermometer Screen Design on the Observed Temperature. *WMO 13*. Geneva Switzerland: World Meteorological Organisation.

Steffensen, P., Larsen, F. & Cappelen, J. 1993. Homogeneity test of climatological data. Copenhagen: Danish Meteorological Institute.

Stepanek, P., Zahradnicek, P. & Farda, A. 2013. Experiences with data quality control and homogenization of daily records of various meteorological elements in the Czech Republic in the period 1961–2010. *Quarterly Journal of the Hungarian Meteorological service*, 117, 123-141.

Thorne, P. W., Parker, D. E., Christy, J. R. & Mears, C. A. 2005a. Uncertainties in climate trends. Lessons from upper-air temperature records. *Bulletin of the American Meteorological Society*, 86, 1437-1442.

Thorne, P. W., Parker, D. E., Tett, S. F. B., Jones, P. D., Mccarthy, M. P., Coleman, H. & Brohan, P. 2005b. Revisiting radiosonde upper air temperatures from 1958 to 2002. *Journal of Geophysical Research*, 110, 1 -17

Thorne, P. W. & Vose, R. S. 2010. Reanalyses suitable for characterizing long-term trends. *Bulletin of the American Meteorological Society*, 91, 353-362.

Thorne, P. W., Willett, K. M., Allan, R. J., Bojinski, S., Christy, J. R., Fox, N., Gilbert, S., Jolliffe, I., Kennedy, J. J., Kent, E., Tank, A. K., Lawrimore, J.,

- Parker, D. E., Rayner, N., Simmons, A., Song, L., Stott, P. A. & Trewin, B. 2011. Guiding the creation of a comprehensive surface temperature resource for twenty-first-century climate science. *Bulletin of the American Meteorological Society*, 92, ES40-ES47.
- Thorne, P. W., Donat, M., Dunn, R., Williams, C., Alexanfer, L., Caerar, J., Durre, I., Harris, I., Hausfather, Z., Jones, P., Menne, M., Rohde, R., Vose, R., Davy, R., Klein-Tank, A., Lawrimore, J., Peterson, T. & Rennie, J. 2016. Reassessing changes in diurnal temperature range. *Journal of Geophysical Research: Atmospheres*, 121, 5138-5158.
- Thorne, P. W., Allan, R. J., Ashcroft, L., Brohan, P., Dunn, R. J. H., Menne, M. J., Pearce, P. R., Picas, J., Willett, K. M., Benoy, M., Bronnimann, S., Canziani, P. O., Coll, J., Crouthamel, R., Compo, G. P., Cuppett, D., Curley, M., Duffy, C., Gillespie, I., Guijarro, J., Jourdain, S., Kent, E. C., Kubota, H., Legg, T. P., Li, Q., Matsumoto, J., Murphy, C., Rayner, N. A., Rennie, J. J., Rustemeier, E., Slivinski, L. C., Slonosky, V., Squintu, A., Tinz, B., Valente, M. A., Walsh, S., Wang, X. L., Westcott, N., Wood, K., Woodruff, S. D. & Worley, S. J. 2017. Toward an integrated set of surface meteorological observations for climate science and applications. *Bulletin of the American Meteorological Society*, 98, 2689-2702.
- Thorne, P. W., Diamond, H. J., Goodison, B., Harrigan, S., Hausfather, Z., Ingleby, N. B., Jones, P. D., Lawrimore, J. H., Lister, D. H., Merlone, A., Oakley, T., Palecki, M., Peterson, T. C., De Podesta, M., Tassone, C., Venema, V. & Willett, K. M. 2018. Towards a global land surface climate fiducial reference measurements network. *International Journal of Climatology*, 38, 2760-2774.
- Tian, D., Pan, M. & Wood, E. F. 2018. Assessment of a high-resolution climate model for surface water and energy flux simulations over global land: An intercomparison with reanalyses. *Journal of Hydrometeorology*, 19, 1115-1129.
- Titchner, H. A. & Rayner, N. A. 2014. The Met office Hadley Centre sea ice and sea surface temperature data set, version 2: 1. sea ice concentrations. *Journal of Geophysical Research: Atmospheres*, 119, 2864-2889.
- Toreti, A., Kuglitsch, F. G., Xoplaki, E., Della-Marta, P. M., Aguilar, E., Prohom, M. & Luterbacher, J. 2011. A note on the use of the standard normal homogeneity test to detect inhomogeneities in climatic time series. *International Journal of Climatology*, 31, 630-632.
- Trewin, B. 2010. Exposure, instrumentation, and observing practice effects on land temperature measurements. *Wiley Interdisciplinary Reviews: Climate Change*, 1, 490-506.

- Trewin, B. 2018. The Australian climate observations reference network - surface air temperature (ACORN-SAT) Version 2. *Australian Government Bureau of Meteorology Bureau Research Report*. Australia: Australian Bureau of Meteorology.
- Tukey, J. W. 1958. Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics*, 29, 614-623.
- Tuomenvirta, H., Alexandersson, H., Drebs, A., Frich, P. & Nordli, P. O. 2000. Trends in nordic and arctic temperature extremes and ranges. *Journal of Climate*, 13, 977-990.
- Tuomenvirta, H. 2002. Homogeneity testing and adjustment of climatic time series in Finland. *Geophysica*, 38, 15-41.
- Uppala, S., Kållberg, P., Simmons, A., Andrae, U., Da Costa Bechtold, V., Fiorino, M., Gibson, J., Haseler, J., Hernandez, A., Kelly, G., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R., Andersson, E., Arpe, K., Balmaseda, M., Beljaars, A., Van De Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B., Isaksen, L., Janssen, P., Jenne, R., McNally, A., Mahfouf, J., Morcrette, J., Rayner, N., Saunders, R., Simon, P., Sterl, A., Trenberth, K., Untch, A., Vasiljevic, D., Viterbo, P. & Woollen, J. 2005. The ERA-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society*, 131, 2961-3012.
- Van Der Meulen, J. P. & Brandsma, T. 2008. Thermometer screen intercomparison in De Bilt (The Netherlands), Part I: Understanding the weather-dependent temperature differences). *International Journal of Climatology*, 28, 371-387.
- 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., Rasul, D., Rustemier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaforte, F., Fratianni, S., Cheval, S., Klancer, M., Brunetti, M., Gruber, C., Prohom-Duran, M., Likso, T., Esteban, P. & Brandsma, T. 2012. Benchmarking homogenization algorithms for monthly data. *Climate of the Past*, 8, 89-115.
- Venema, V., Trewin, B., Wang, X., Szentimrey, T., Lakatos, M., Aguilar, E., Auer, I., Guijarro, J., Oria, C., Louamba, W. S. R. L. L. & Rasul, G. 2018. WMO_Homogenization_guidance_full_final_draft. 1. Geneva: WMO.
- Vicente-Serrano, S. M., Angel Saz- Sanchez, M. & Cuadrat, J. M. 2003. Comparative analysis of interpolation methods in the middle Ebro valley(Spain): application to annual precipitation and temperature. *Climate Research*, 24, 161-180.

- Vincent, L. A. 1998. A technique for the identification of inhomogeneities in Canadian temperature series. *Journal of Climate*, 11, 1094-1104.
- Vittori, O. & Mestitz, A. 1981. Calibration of the 'Florentine little thermometer'. *Endeavour*, 5, 113-118.
- Vose, R., Peterson, T., Schmoyer, R., Eischeid, J., Steurer, P., Heim, R. & Karl, T. 1992. The global historical climatology network long-term monthly temperature, precipitation, sea level pressure data and station pressure data. USA: Oak Ridge National Lab., TN (United States). Carbon Dioxide Information Analysis Center.
- Vose, R., Arndt, D., Banzon, V. F., Easterling, D. R., Gleason, B., Huang, B., Kearns, E., Lawrimore, J. H., Menne, M. J., Peterson, T. C., Reynolds, R. W., Smith, T. M., Williams, C. N. & Wuertz, D. B. 2012. NOAA'S merged land-ocean surface temperature analysis. *Bulletin of the American Meteorological Society*, 93, 1677-1685.
- Vose, R., Applequist, S., Squires, M., Durre, I., Menne, M. J., Williams, C. N., Fenimore, C., Gleason, K. & Arndt, D. 2014. Improved historical temperature and precipitation time series for U.S. climate divisions. *Journal of Applied Meteorology and Climatology*, 53, 1232-1251.
- Wang, J., Xu, C., Hu, M., Li, Q., Yan, Z. & Jones, P. 2018. Global land surface air temperature dynamics since 1880. *International Journal of Climatology*, 38, e466-e474.
- Wang, X. L., Wen, Q. H. & Wu, Y. 2007. Penalized maximal t-test for detecting undocumented mean change in climate data series. *Journal of Applied Meteorology and Climatology*, 46, 916-931.
- Warne, J. 1999. A preliminary investigation of temperature screen design and their impact on temperature measurement. Australia: Australian Bureau of Meteorology.
- Wendland, W. M. & Armstrong, W. 1993. Comparison of maximum–minimum resistance and liquid-in-glass thermometer records. *Journal of Atmospheric and Oceanic Technology*, 10, 233-237.
- Willett, K., Williams, C., Jolliffe, L. T., Alexander, L. V., Bronnimann, S., Vincent, L. A., Easterbrook, S., Venema, V. K. C., Berry, D., Warren, R. E., Lopardo, C., Auchmann, R., Aguilar, E., Menne, M., Gallagher, C., Hausfather, Z., Thorarindottir, T. & Thorne, P. W. 2014. A framework for benchmarking of homogenisation algorithm performance on the global scale. *Geoscientific Instrumentation Methods and Data Systems*, 3, 187-200.

- Williams, C. N., Menne, M. & Thorne, P. W. 2012. Benchmarking the performance of pairwise homogenization of surface temperatures in the United States. *Journal of Geophysical Research: Atmospheres*, 117.
- Willmott, C. J. & Robeson, S. M. 1995. Climatologically aided interpolation (CAI) of terrestrial air temperature. *International Journal of Climatology*, 15, 221-229.
- Wmo 1973. One hundred years of international cooperation in meteorology 1873 to 1973 a historical review. Geneva: World Meteorological Organization.
- Woodruff, S. D., Worley, S. J., Lubker, S. J., Ji, Z., Eric Freeman, J., Berry, D. I., Brohan, P., Kent, E. C., Reynolds, R. W., Smith, S. R. & Wilkinson, C. 2011. ICOADS release 2.5: extensions and enhancements to the surface marine meteorological archive. *International Journal of Climatology*, 31, 951-967.
- 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. & 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. & Park, J. H. 2020. A global-scale multidecadal variability driven by Atlantic multidecadal oscillation. *National Science Review*, 7, 1190-1197.
- Zen-De-Figueiredo-Neves, G., Gallardo, N. & Vecchia, F. 2017. A short critical history on the development of meteorology and climatology. *Climate*, 5, 1-17.
- Zhang, H.-M., Lawrimore, J., Huang, B., Menne, M., Yin, X., Sanchez-Lugo, A., Gleason, B., Vose, R., Arndt, D., Rennie, J. & Williams, C. 2019. Updated temperature data give a sharper view of climate trends. *Eos* 100.
- Zhou, C., He, Y. & Wang, K. 2018. On the suitability of current atmospheric reanalyses for regional warming studies over China. *Atmospheric Chemistry and Physics*, 18, 8113-8136.
- Zillman, J. 2005. The Role Of National Meteorological Services In the provision of public weather service. WMO Papers.