Received: 1 May 2020 | Revised: 14 September 2020 | Accepted: 15 October 2020 | Published on: 3 November 2020

DOI: 10.1002/joc.6898

RESEARCH ARTICLE

International Journal of Climatology RMetS

Assessing potential of sparse-input reanalyses for centennial-scale land surface air temperature homogenisation

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

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

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

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

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

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

Funding information
Climate Program Office; NOAA Physical Sciences Laboratory

Abstract

Observations from the historical meteorological observing network contain many artefacts of non-climatic origin which must be accounted for prior to using these data in climate applications. State-of-the-art homogenisation approaches use various flavours of pairwise comparison between a target station and candidate neighbour station series. Such approaches require an adequate number of neighbours of sufficient quality and comparability – a condition that is met for most station series since the mid-20th Century. However, pairwise approaches have challenges where suitable neighbouring stations are sparse, as remains the case in vast regions of the globe and is common almost everywhere prior to the early 20th Century. Modern sparse-input centennial reanalysis products continue to improve and offer a potential alternative to pairwise comparison, particularly where and when observations are sparse. They do not directly ingest or use land-based surface temperature observations, so they are a formally independent estimate. This may be particularly helpful in cases where structurally similar changes exist across broad networks, which challenges current techniques in the absence of metadata. They also potentially offer a valuable methodologically distinct method, which would help explore structural uncertainty in homogenisation techniques. The present study compares the potential of spatially-interpolated sparse-input reanalysis products to neighbour-based approaches to perform homogenisation of global monthly land surface air temperature records back to 1850 based upon the statistical properties of station-minus-reanalysis and station-minus-neighbour series. This shows that neighbour-based approaches likely remain preferable in data dense regions and epochs. However, the most recent reanalysis product, NOAA-CIRES-DOE 20CRv3, is potentially preferable in cases where insufficient neighbours are available. This may in particular affect long-term global average estimates where a small number of long-term stations in data sparse regions will make substantial contributions to global estimates and may contain missed data artefacts in present homogenisation approaches.

KEYWORDS
homogeneity, reanalyses, surface temperatures

1 | INTRODUCTION

Scientists have been collecting and analysing global land surface air temperature records for a very long time. Callender put together the first truly global collection of temperature estimates in 1938, collating by hand a number of global station records and concluding that carbon dioxide from the burning of fossil fuels was responsible for the warming of the climate that he had assessed over the previous 50 years (Callender, 1938; Hawkins and Jones, 2013). Since that pioneering work, there have been significant advances and very many global, regional, and national surface temperature datasets have been created, curated, and analysed. These have become progressively more complete, and the methods used in their creation have become more advanced. Notable current global and regional works include, but are not limited to, CRUTEM now at version 5 (Osborn et al., submitted), GHCN-M now at version 4 (Menne *et al*., 2018), E-OBS (Cornes *et al*., 2018), and the Berkely Earth dataset (Rohde *et al*., 2013). 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 (Sanchez-Lugo *et al*., 2019).

Despite numerous advances, 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. 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, urbanization effects, instrument changes, land cover changes, and observation practice changes amongst others (Parker, 1993; Peterson *et al*., 1998; Changnon and Kunkel, 2005; Trewin, 2010). 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 36,000 individual station records (although many are short-period records). It is the most extensive global collection of original instrumental land surface air temperature series produced thus far. It increases by approximately three-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 homogenized to create both a new version of the Global Historical Climatology Network Monthly product – GHCNv4 (Menne *et al*., 2018), and an estimate of Diurnal Temperature Range changes (Thorne *et al*., 2016). Both of these have utilized 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 homogenized 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*., 2005).

Such broadened 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 homogenize radiosonde temperatures (Haimberger *et al*., 2012).

More recently, surface-only sparse-input reanalysis products that extend back to the 19th Century have been produced (Compo *et al*., 2011; Poli *et al*., 2016; Laloyaux

et al., 2018; Slivinski *et al*., 2019). Most specify fields of homogenized sea surface temperatures and sea ice concentration as a lower boundary condition (Rayner *et al*., 2005; Titchner and Rayner, 2014). 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 (Ferguson and Villarini, 2012; Jones *et al*., 2012; Compo *et al*., 2013; Parker, 2016; Wang *et al*., 2018) 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). 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.

This paper sets out to assess whether using the latest generation of sparse-input reanalysis products may plausibly constitute an alternative approach to homogenize 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 homogenisation 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 paper is structured as follows. Section 2 considers the options of constructing a comparator series for homogenisation and outlines the role that sparse-input reanalysis products could play. Section 3 highlights work to post-process the ISTI databank to remove identified duplicates. Section 4 details the interpolation method employed to arrive at a reanalysis-based comparator record. Section 5 examines the relative performance of state-of-the-art pairwise comparison versus sparse-input reanalysis as a tool for homogenisation based upon the statistical properties of station-minus-neighbour and station-minus-reanalysis timeseries. Section 6 discusses the key remaining issues. Finally, conclusions are given in Section 7.

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 (e.g., Menne and Williams, 2009; Venema *et al*., 2012). Homogenisation of a candidate station record thus requires some comparator series. Acquiring a suitable and robust series is 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 as arising from data artefacts. Fundamentally, a comparator series needs to be as highly correlated with the target series, and with as low noise, 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 and Williams, 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.

Some researchers have used sections of the record of the station under examination that they have high

confidence in to homogenize suspect sections of the same record (Peterson *et al*., 1998; Mamara *et al*., 2012). But most 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 also be both of sufficient length and overlap substantially. 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 (Figure 1).

As a novel alternative, series from reanalysis products offer the potential to circumvent many of these issues. Sparse-input reanalysis products (Compo *et al*., 2011; Poli *et al*., 2016; Laloyaux *et al*., 2018; Slivinski *et al*., 2019) 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 homogenize 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.

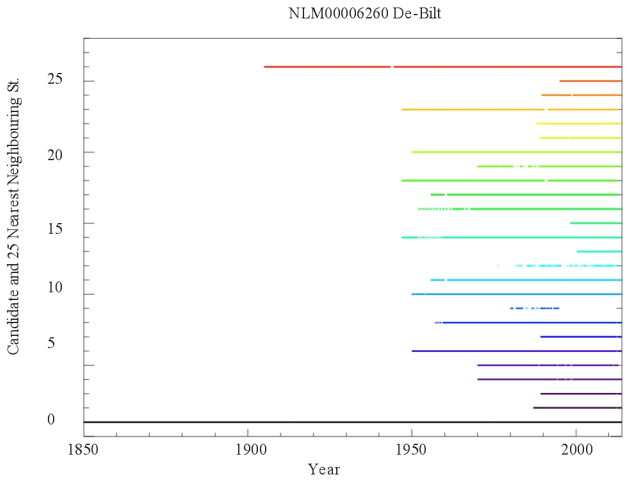


FIGURE 1 Summary of neighbour station data availability for De Bilt since 1850 (the series extends to the 1700s but for the present study the interest is in the period since the 1850s driven by availability of sparse-input reanalysis products and globally representative observations). This series is a centennial station series with almost continuous availability (bottom black) since the 1850s, although prior to 1897 data arise 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 1850–1900 period. There is one potential neighbour for the period of 1900–1945, after which there are several possible neighbours for pairwise homogenisation. Effectively pairwise homogenisation techniques are not possible for the period of 1850–1945 without expanding the neighbour search radius due to a lack of suitable neighbours

3 | REMOVAL OF DUPLICATED DATA FROM THE ISTI DATABANK HOLDINGS

The ISTI databank consists of a hierarchical merge of records of original monthly-averaged temperature observations from more than 70 underlying data sources (Rennie *et al*., 2014) of hugely varying volume (both station number and period of record) and provenance. Many stations have been shared broadly such that they exist in many of the sources used to create the merged holdings. To further confound matters, data from different sources may differ in coordinate precision, station naming, application of Quality Control, and even in a small number of cases homogenisation. Furthermore, some sources may have performed merges which will have been invisible to the databank creators. The ISTI databank construction process is automated and uses a mix of geographical metadata and data similarity to make a decision to either: (a) merge series; (b) create a new series; or (c) withhold the series. Despite significant efforts, it is recognized that there will have been inevitable incorrect choices made by the algorithm.

For this study, several steps have been taken to address potential issues in the holdings of the ISTI

databank. NCEI has undertaken their own blacklisting of issues they found in the ISTI databank in GHCNMv4 construction, (Rennie, pers. comm), and these have been applied herein as a first step. Following identification of additional potential issues, two distinct analysis steps were undertaken herein to detect and remedy apparent residual duplication of records in the database prior to using it. Between them, these additional steps removed about 3.5% of the remaining stations in entirety and, in addition, removed questionable sources from 340 series leading to a reduction in their station series length.

The first issue pertains almost exclusively to data rich regions and is most prevalent for 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, quality controlled, and/or homogenized. The end result is that for several stations, either exact matches or near exact matches that, that is, repeat a seasonal cycle, exist across multiple nearby locations in the ISTI databank. To identify such cases, each station of more than 10 years record duration was compared with its 25 nearest neighbours. Cases with either strings of zero differences or annually repeating differences were identified. All such cases with greater than 10% prevalence were then examined manually to confirm the presence of duplication, and the longest available series was retained with all other series discarded (Figure 2). A total of 973 such cases of erroneous data duplications have been identified. In many cases, these had multiple duplications that necessitated dropping more than one series to be satisfactorily resolved.

The second issue arises when considering neighbour distances. There are just over 300 cases where 2 or more stations in the ISTI databank have exactly matched geographical coordinates recorded. Such a result is possible, particularly where the coordinate resolution is coarse (e.g., a 0.1° resolution coordinate, which is not uncommon in many of the ISTI databank sources, has a c. 10 km radius in which the true location may exist). The issue is most prevalent in Canada and the United States, where the ISTI databank is most dense, accounting for in excess of 75% of the identified cases. Out of an abundance of caution, a single series is retained at a given unique coordinate. Where records overlap, only the longest record has been retained. Where stations did not overlap in time, they have been merged.

Between the two steps outlined above, a total of 1,467 additional station series have been removed (sometimes by merging), beyond the NCEI originated blacklist. A map of these removed stations is given in Figure 3. All

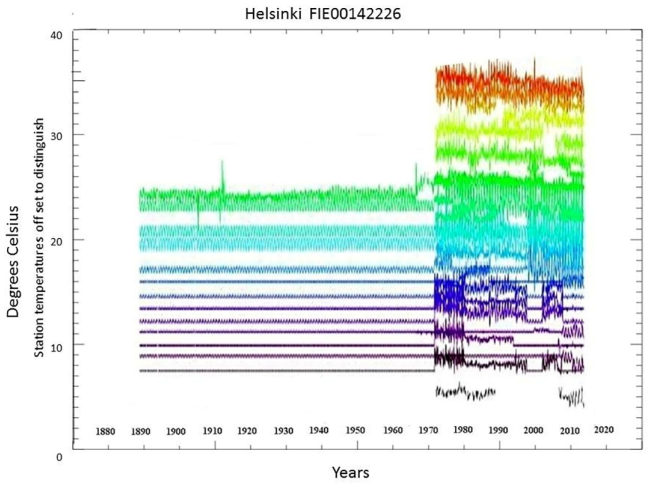


FIGURE 2 Example from Helsinki, Finland, where neighbouring stations records exhibit substantial spurious periods of data overlap due to ghosting of a single long station into the regional network on multiple occasions. Each series shows the candidate station minus the neighbour (offset by 2C and coloured distinctly for clarity). In reality, there is one very long-term record in Helsinki, but it has been shared on multiple occasions. Further analysis shows that the issue for this station arises in large part due to upstream source merge decisions, but there are also some cases where the ISTI databank merge procedures erroneously merged this single long series from several sources into nearby stations. Mis-merged series contain both repeat zero differences and annually repeating values implying one or more of the mis-merged series has been homogenized, further complicating matters

the decisions have been communicated back to NOAA NCEI, who have integrated this additional blacklist into their operational GHCNMv4 processing (Rennie, pers. comm), and are reported in Data S1. 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).

4 | INTERPOLATION OF REANALYSIS GRIDDED SERIES TO STATION LOCATIONS

Sparse-input (or 20th Century) reanalyses 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:

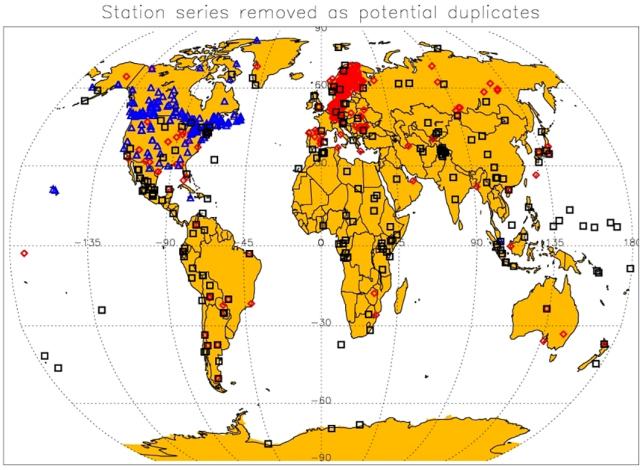


FIGURE 3 Stations removed from the ISTI databank in this work, in addition to the NOAA NCEI provided blacklist (black boxes). The stations in red diamonds, predominantly in Europe, are stations that had exactly duplicated data or seasonally repeating offsets. In all such cases, only the longest record was retained. Stations marked in blue triangles, predominantly in North America, are where two or more stations had the same coordinates but not necessarily the same data. If two or more stations had the same coordinates but different data, the station with the longest time series was retained and the shorter time series was deleted. If the two stations had the same coordinates with time series over distinct periods a merge was undertaken

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 (Compo *et al.*, 2011; 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 20CR. 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) 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 10-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 2-m 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).

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 km for 20CRv2c decreasing significantly as it nears the poles. The other reanalysis products (ERA-20C, 20CRv3 and CERA-20C) are under half these distances. Two methods of inverse distance weighting were considered: (a) inverse distance weighting and (b) inverse distance squared weighting. Weighting was applied to absolute values and the resulting series was then anomalised after matching to target station data availability. Using a station period of record climatology at this stage maximizes the retained pool of station records. Results were compared individually in anomaly space for a selection of global stations and evaluated using Pearson correlations and standard deviations. On an individual station basis, there were only negligible differences (Table 1) with, in general, a very marginal performance advantage using inverse squared distance weightings. Given the very marginal differences, subsequent sections consider only the inverse distance squared weighting approach.

The 20CRv2c data set comes as an ensemble mean and 56 underlying ensemble members. To determine whether the ensemble members may provide additional valuable context, the similarity to the target station series

TABLE 1 Comparison of interpolated reanalysis minus station difference series correlation and standard deviation using the 20CRv2c ensemble mean product between inverse linear distance and inverse linear squared distance for interpolation for selected stations (Section 5.1)

			Inverse distance weighing interpolation		Inverse distance squared weighted interpolation		Differences	
Station	**Name**	**Country**	**Sigma**	***r***	**Sigma**	***r***	**Sigma**	***r***
AR000087828	Trelew Aero	Argentina	1.096	0.628	1.096	0.628	0.000	0.000
ASXLT209646	Hobarttasmanwas_949700	Australia	0.857	0.510	0.857	0.510	0.000	0.000
ASXLT263670	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
AYXLT563342	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 Ils)	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	Findland	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
MZXLT405557	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
PKXLT983863	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
TZXLT095229	Dar_Es_Salaam_Tanzania_Beafr	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

Note: The difference between the methods on both an individual basis and an aggregate basis is 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).

of estimates derived from individual ensemble members was compared with that from the ensemble mean for a selected set of stations (Section 5.1). Consistent with Kalman Filter theory, each ensemble member contains the analysed geophysical signal plus noise whose standard deviation is the uncertainty of the ensemble mean. Therefore it is expected that the ensemble mean reanalysis values should yield an estimate that may be better correlated and have lower noise than the individual members. For the selected stations analysed, this holds true in the majority of cases (e.g., Figure 4). 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 instigation of the analysis

only the ensemble mean product was available. Initial inspection confirms that the 20CRv2C ensemble member versus ensemble average behaviour found herein holds also for 20CRv3, as would be expected for the reasons discussed above.

5 | RESULTS

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, 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.

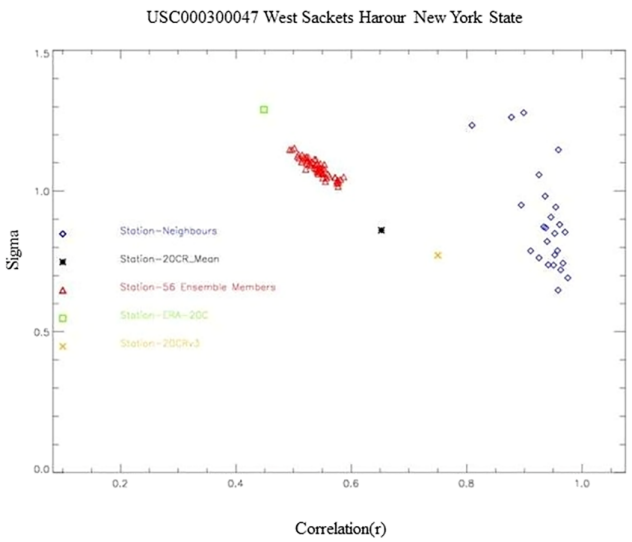


FIGURE 4 An example analysis (Adams central, West Sackets Harour New York state at 43.9° N, 76.0667°W) of correlation (r) and standard deviation (sigma) of the 56 ensemble members of 20CRv2c to determine if the ensemble mean or individual ensemble members are most suitable for further comparison to pairwise homogenisation. The correlation and 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. The study station is in the densely sampled US. The 20CRv2c ensemble mean (black asterisk) is clearly preferable to the underlying ensemble members (red triangles) in this case, and the 20CRv3 ensemble mean (yellow x) is a significant improvement on 20CRv2c mean. For this station, its nearby neighbours still represent a generally preferable estimator based on these similarity metrics

5.1 | Case study stations analysis

A set of 29 carefully-chosen stations was selected in an attempt to ensure a representative sampling that considers the inclusion of urban stations, rural stations, coastal stations, desert stations, high altitude stations, island stations, and densely and sparsely sampled regions. 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 summarized in Table 2 and their locations are shown in Figure 5. To retain the maximum number of stations for use in this part of the analysis, the climatology of the full period of station availability over 1850–2014 was used and not normalized to 1961–1990 or any other rigid 30-year climatology.

In the literature, there are several different ways of selecting candidates as a reference series for pairwise homogenisation (Menne and Williams, 2009; Mamara *et al.*, 2012; Wang *et al.*, 2018). These methods often select stations based upon correlation, spatial representativeness, or both. Two selection choices are used here. Firstly, for simplicity, the 25 nearest neighbours are used herein to compare relative performance of reanalysis- and neighbour-based approaches (see Figure 5). While less optimized than many state-of-the-art techniques, it mitigates against a metric-based selection of a neighbour

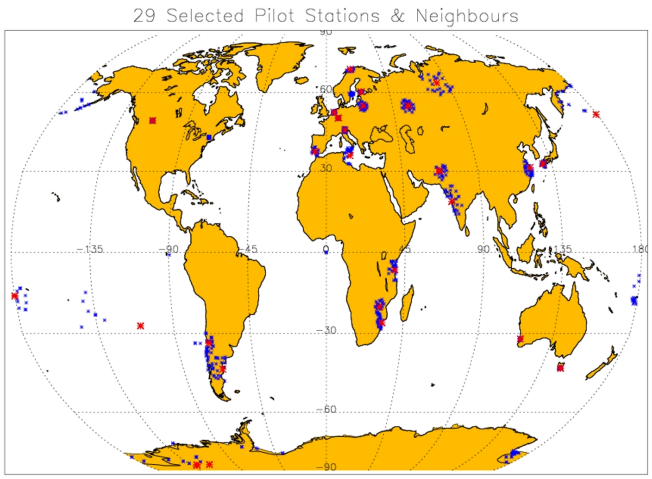


FIGURE 5 Locations of the 29 stations chosen for the case study (red crosses). For pairwise comparison of these stations, their 25 nearest neighbour stations were also selected (blue stars). Note that for some case study stations, particularly in the Southern Hemisphere, there are common neighbours

TABLE 2 A summary of the initial subset of 29 case study stations used herein including their identifier, name, country, number of valid observations, geolocation, regional characteristics and local environment

Station number	Station name	Country	No. Obs	Latitude	Lon	Alt	Regions	Environment
AR000087828	TRELEW_AERO	Argentina	1367	−43.2	−65.266	43	Mid Lat	Desert
ASXLT209646	HOBARTTASMANWAS_949700	Australia	1732	−42.9	147.5	54	Mid Lat	Complex
ASXLT263670	PERTHAUSWAS_946080	Australia	1028	−32	115.9	60	S. Tropic	Complex
AYM00089314	THERESA	Antarctic	201	−84.6	−115.817	1463	Polar	Ice
AYXLT563342	ERIN	Antarctic	146	-84.6	-128.82	1006	Polar	Ice
CA003031400	CARWAY	Canada	1025	49	−113.383	1354	Mid Lat	Crop/Grass
CHM00058362	SHANGHAI	China	1703	31.4	121.467	4	S. Tropic	Complex
CI000085469	ISLA_DE_PASCUA	Chile (Easter Ils)	862	−27.167	−109.433	69	S. Tropic	Crop/Grass
CIXLT967829	SANTIAGOWAS_855770	Chile	1765	−33.5	−70.7	520	Mid Lat	Complex
FIE00142226	HELSINKI_KUMPULA	Findland	1961	60.2028	24.9642	24	Polar	Complex
FJ000091652	UDU_POINT_AWS	Fiji	664	−16.133	−179.983	63	Tropic	Costal
GMM00010628	GEISENHEIM	Germany	1566	49.983	7.95	123	Mid Lat	Crop/Grass
INM00043057	BOMBAY_COLABA	India	1628	18.9	72.8167	11	Tropic	Costal
ITE00115588	PADOVA	Italy	1888	45.3983	11.8803	12	Mid Lat	Complex
JA000047817	NAGASAKI	Japan	1681	32.732	129.866	35	S. Tropic	Complex
LH000026730	VILNIUS	Lithuania	1935	54.6331	25.1	156	Mid Lat	Crop/Grass
MT000016597	LUQA	Malta	1862	35.85	14.4831	91	Mid Lat	Complex
MZXLT405557	LOURENCO_MARQUES	Mozambique	1031	−26	32.6	64	S. Tropic	Costal
NLM00006260	DE_BILT_1	Netherlands	1968	52.1014	6.1867	2	Mid Lat	Costal
NOE00134898	TROMSOLANGNES	Norway	1908	69.6767	18.9131	8	Polar	Costal
PKXLT983863	QUETTASHEIKH_MANDA	Pakistan	1138	30.18	66.95	1803	S. Trope	Desert
RSM0O023662	TOLKA	Russia	800	63.98	82.08	31	Mid Lat	Forest/Jungle
RSM00028722	UFA	Russia	1486	54.7167	55.8831	104	Mid let	Forest/Jungle
SPE00120143	HUELVA_RONDA_DEL_ESTE	Spain	1335	37.28	−6.9	19	Mid Lat	Complex
SWE00136129	VARTAN	Sweden	1950	59.35	18.1	20	Mid Lat	Complex
TZXLT095229	DAR_ES_SALAAM_TANZANIA_BEAFR	Tanzania	803	−6.5	39.29999	58	Tropic	Costal
USC00300047	ALBANY	United States	1826	42.6461	−73.7472	13	Mid Lat	Complex
USC00500252	AMCHITKA	USA	136	51.3833	179.2833	69	Mid Lat	Complex
ZI000067975	MASVINGO	Zimbabwe	1019	−20.067	30.867	1095	Tropic	Crop/Grass

station-set that may inadvertently contain data issues of similar structure to those present in the candidate station. Secondly, for completeness, the 25 closest neighbours with at least a 50% data overlap were also retained. Including the 50% data overlap criteria results in an increase in distance between the candidate station and the overall pool of selected neighbours that becomes larger for the longest station records. This increase in the distance has a knock-on effect in decreasing the correlation between the pairs and an increase in the standard deviation in the difference series. In well sampled regions such as North America, these effects are small. However, in the poorly sampled regions of the globe, the increased separation between the pairs (Figure 6) results in a significant cost to the correlation and standard deviation of the difference series, such that the selection of neighbours based on the 50% observation overlap criterion disadvantages pairwise comparison compared to homogenisation by sparse-input reanalysis (compare Tables 3 and 4).

For each of the 29 case-study stations, the performance of the reanalysis-estimate and neighbour-based difference series was compared. Figure 7 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, and is available quasi-continuously through to the present. The difference series to reanalysis (Figure 7 top panel) shows a marked break in the series at around the turn of the twentieth Century relative to both 20CRv2 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 Utretch to De Bilt in 1897. This move is coincident with the break that would be identified as a 'potential' break if only reanalysis data were available. 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 7 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 1897 break. Extending the neighbour search to include stations with >50% overlap (Figure 7 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.

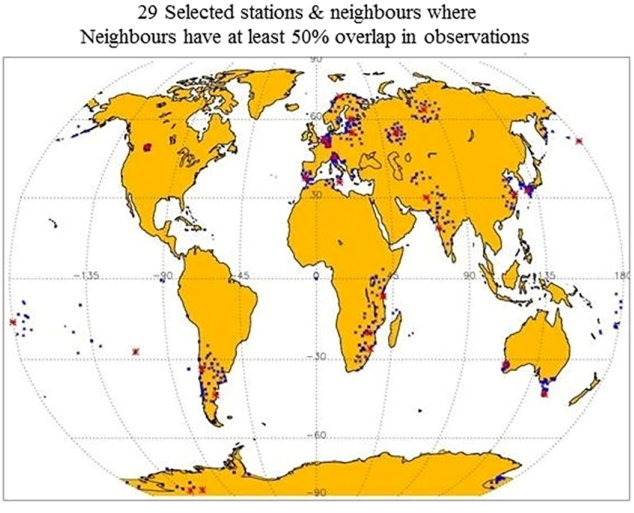


FIGURE 6 As in Figure 5 but now showing the 25 nearest neighbour stations with a minimum of 50% observation overlap (blue stars). Note the increased distance between candidate station and neighbours, particularly in less well sampled regions

The results shown in Figure 7 and confirmed with KNMI metadata 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 5 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).

TABLE 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)

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	ASXLT209646	0.510256	0.553	0.550	0.349	0.881	0.857	0.828	0.807	1.055	0.419	380
Australia	*ASXLT263670*	*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	**AYXLT563342**	**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**	**0353**	**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	**MZXLT405557**	**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	PKXLT983863	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*
Tarzania	TZXLT095229	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**

Individual correlations between the case study stations and their nearest neighbours vary from near 1 to values of less than 0.1 (Table 4). 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 ranges 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 Ferandez 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. This is returned to in Section 5.3.

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 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 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 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 limit of the current analysis to postulate 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.

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, 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 8 illustrates the global distribution of ISTI databank stations with greater than 120 months of observations into these regions. This illustrates the uneven distribution of meteorological stations, with fully 68.2% of the ISTI databank stations 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.

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 9). This abrupt shift is much reduced in both 20CRv3 (Figure 10) and ERA-20C (Figure 11). Overall, 20CRv3 shows the best agreement with aggregated station series across most regions of the globe (Figure 12) and this performance extends far further back in time compared to the prior generation of sparse-input reanalysis products (compare Figure 10 to Figures 9 and 11). 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 (e.g., 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).

5.3 | Comparison between densely and sparsely sampled regions

The Giorgi region analysis in Section 5.2 highlighted the fact that the vast majority of available monthly-mean

TABLE 4 Comparison of standard deviation and correlation (as in Table 1) based upon the selection of neighbours based on two criteria (1) the selection by distance only, and (2) distance with 50% overlap in observations, which will expand the search region and include, on average, stations further away

Station	25 nearest neighbours by distance in KM only			25 nearest neighbours with (1) at least 50% of time series overlap (2) distance in KM					
	Median neighbour distance	Median sigma	Median *r*	Median neighbour distance	Median sigma	Median *r*	Difference in sigma	Difference in *r*	Difference in distance (K.M)
AR000087828	496	1.10322	0.650854	768	1.17775	0.50599	0.075	−0.145	272
ASXLT209646	49	0.418938	0.880665	538	0.823665	0.614597	0.405	−0.266	489
ASXLT263670	26	0.520081	0.888683	105	0.567301	0.863795	0.047	−0.025	79
AYM00089314	1276	2.50478	0.596172	1285	2.61629	0.58712	0.112	−0.009	9
AYXLT563342	1145	2.22507	0.667931	1149	2.38308	0.667931	0.158	0.000	4
CA003031400	35	1.01877	0.932628	91	1.04338	0.936018	0.025	0.003	56
CHM00058362	264	0.637924	0.868736	633	1.04105	0.673613	0.403	−0.195	369
CI000085469	3534	1.52033	0.0373619	3645	1.41506	0.0129261	-0.105	−0.024	111
CIXLT967829	245	1.14816	0.447765	759	1.32908	0.396188	0.181	−0.052	514
FIE00142226	43	0.598645	0.973082	399	1.11627	0.894389	0.518	−0.079	356
FJ000091652	351	0.643598	0.484389	751	0.662648	0.662648	0.019	0.178	400
GMM00010628	44	0.464403	0.966137	75	0.458728	0.969095	-0.006	0.003	31
INM00043057	434	0.850291	0.536156	826	0.94336	0.427932	0.093	−0.108	392
ITE00115588	85	0.793006	0.861037	125	0.781861	0.870261	-0.011	0.009	40
JA000047817	89	0.420418	0.937734	169	0.554082	0.896932	0.134	−0.041	80
LH000026730	167	0.729963	0.951429	406	1.05975	0.901483	0.330	−0.050	239
MT000016597	272	0.814277	0.800403	823	1.22322	0.583578	0.409	−0.217	551
MZXLT405557	197	1.8771	0.224805	729	1.48552	0.274845	-0.392	0.050	532
NLM00006260	61	0.507578	0.955779	222	0.820983	0.900369	0.313	−0.055	161
NOE00134898	71	1.1854	0.79446	533	1.66031	0.752392	0.475	−0.042	462
PKXLT983863	273	1.51831	0.598139	858	1.63385	0.438383	0.116	−0.160	585
RSM00023662	439	1.70664	0.892128	477	1.76759	0.883339	0.061	−0.009	38
RSM00028722	199	0.939461	0.94119	351	1.12412	0.915512	0.185	−0.026	152
SPE00120143	131	0.62613	0.877993	250	0.803909	0.823566	0.178	−0.054	119
TZXLT095229	277	0.901559	0.2575	734	1.21402	0.162887	0.312	−0.095	457
USC00500252	1170	2.08791	0.613247	1325	1.79609	0.530162	-0.292	−0.083	155

TABLE 4 (Continued)

Station	25 nearest neighbours by distance in KM only			25 nearest neighbours with (1) at least 50% of time series overlap (2) distance in KM					
	Median neighbour distance	Median sigma	Median r	Median neighbour distance	Median sigma	Median r	Difference in sigma	Difference in r	Difference in distance (K.M)
ZI000067975	265	0.870369	0.693615	653	1.06254	0.540329	0.192	−0.153	388
Summary									
Minimum		0.41894	0.03736	*Minimum*	0.45873	0.01293			
Maximum		2.50478	0.97308	*Maximum*	2.61629	0.96910			
Mean		1.06046	0.71593	*Mean*	1.20613	0.65505			

Note: The comparison of the differences in sigma (°C), correlation, and distance between the two selection methods are shown in the final three columns.

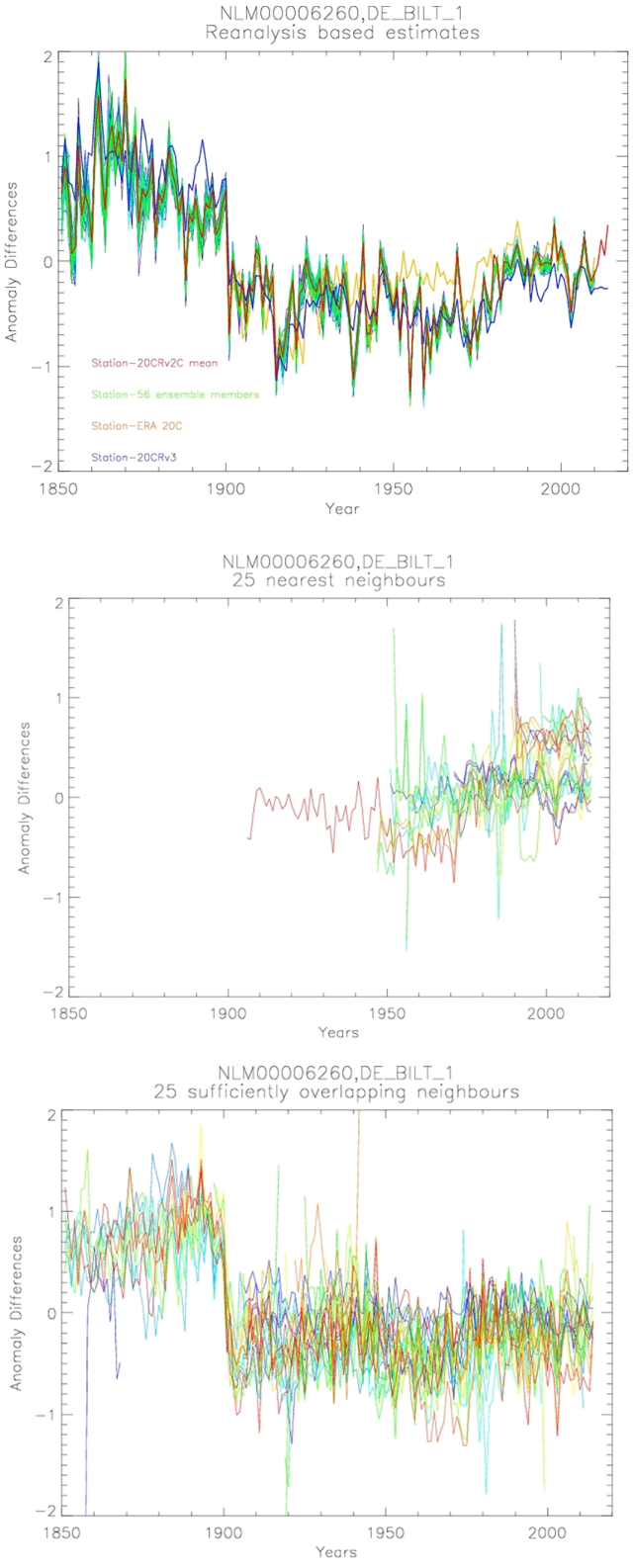
station records in the ISTI databank sample only a small percentage of the globe. This is of particular concern because global mean surface temperatures are calculated by area-weighting regional temperature records from a combination of land and marine sources. The influence of an individual station 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 13). Given that the station series are the basic 'raw' 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 correlation, lower standard deviation). Conversely, in sparse regions, the 20CRv3 estimates likely have more power. The cross-over between the two occurs somewhere around the 600–800 km 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.

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 homogenisation of the longest records. This is particularly acute in sparsely sampled regions. All current state-of-the-art homogenisation 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 of 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 United States 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 temperature 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; Parker, 2016; Simmons *et al.*, 2017; Zhou *et al.*, 2018), provides the evidence basis that the quality of 20CRv3 is likely sufficient.

A substantial further advantage in the use of these reanalyses spanning more than a century is the

FIGURE 7 Top panel: Anomaly difference series between the long-running De Bilt series in the Netherlands (although note caveats identified in Figure 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). Each neighbour difference series is a different colour for illustrative purposes in the middle and lower panels

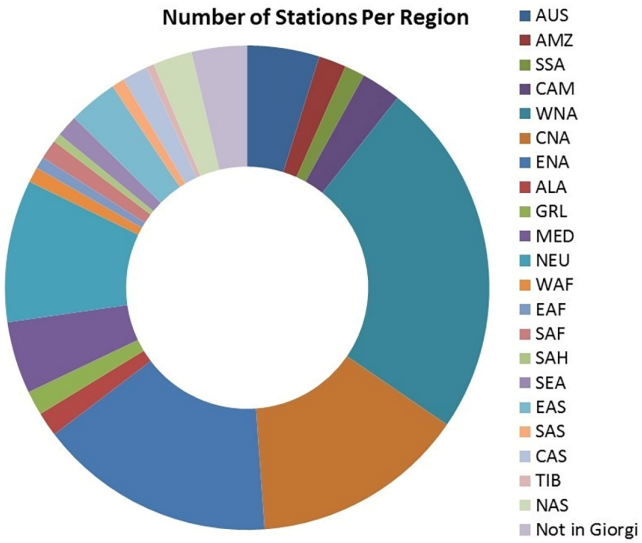


FIGURE 8 The 27,639 long-term stations in the ISTI dataset, following removal of questionable series as detailed in Section 3, split out into Giorgi region groupings. Note the extra grouping of 'Not in Giorgi' which captures the Antarctic, remote island, and some Arctic sites not included in the original 21 Giorgi regions

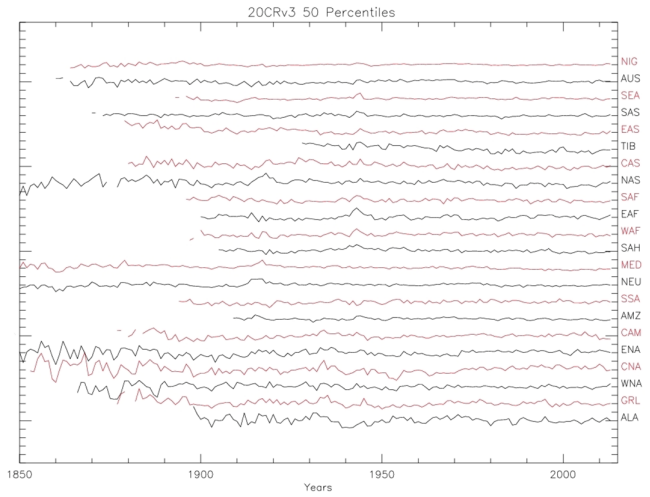


FIGURE 10 As Figure 9 but for 20CRv3. The 20CRv3 product shows better performance than either ERA-20C or 20CRv2c across all regions with stable behaviour back to at least 1900 across all regions. The mid-20th century is much more stable than either of the other sparse-input reanalysis products

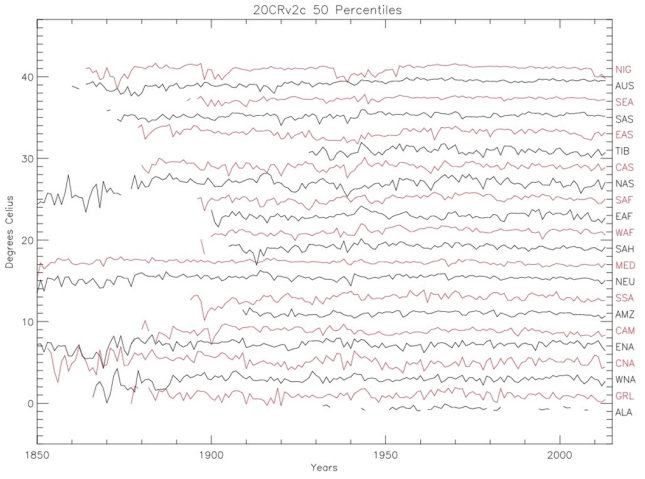


FIGURE 9 The median value (50th percentile) of the regionally-aggregated differences series between 20CRv2c ensemble mean and the station anomalies at each timestep aggregated over the Giorgi regions. Each series is vertically offset for clarity. There is a marked degradation in apparent performance over many regions in the mid-20th century. For region definitions see main text

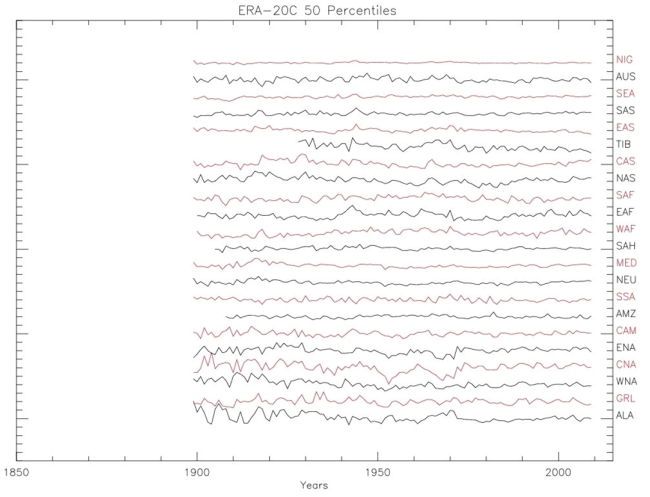


FIGURE 11 As Figure 9 but for ERA-20C which starts in only 1900. This is a clear limitation on the use of ERA-20C compared to the two NOAA sparse-input reanalysis products. Although ERA-20C contains apparent decadal variations in the mid-20th century, the degradation in this case is much less marked than for 20CRv2c in most regions (c.f. Figure 9)

availability of a comparator series at each and every month for which a temperature value is available in the target station series. Conversely, Figure 14 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

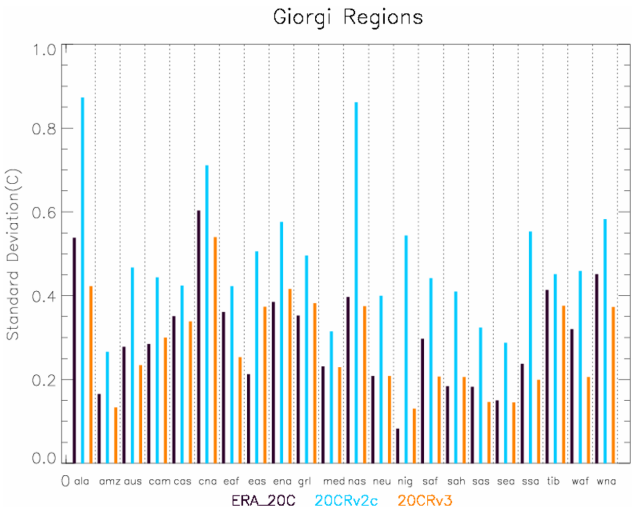


FIGURE 12 For each of the 22 Giorgi regions, the bars summarize the Standard Deviation of timeseries shown in Figures 9–11 for the three reanalysis products. The 20CRv3 product exhibits the lowest standard deviation for almost all regions (except CAM, EAS, GRL, Not in Giorgi, SAH)

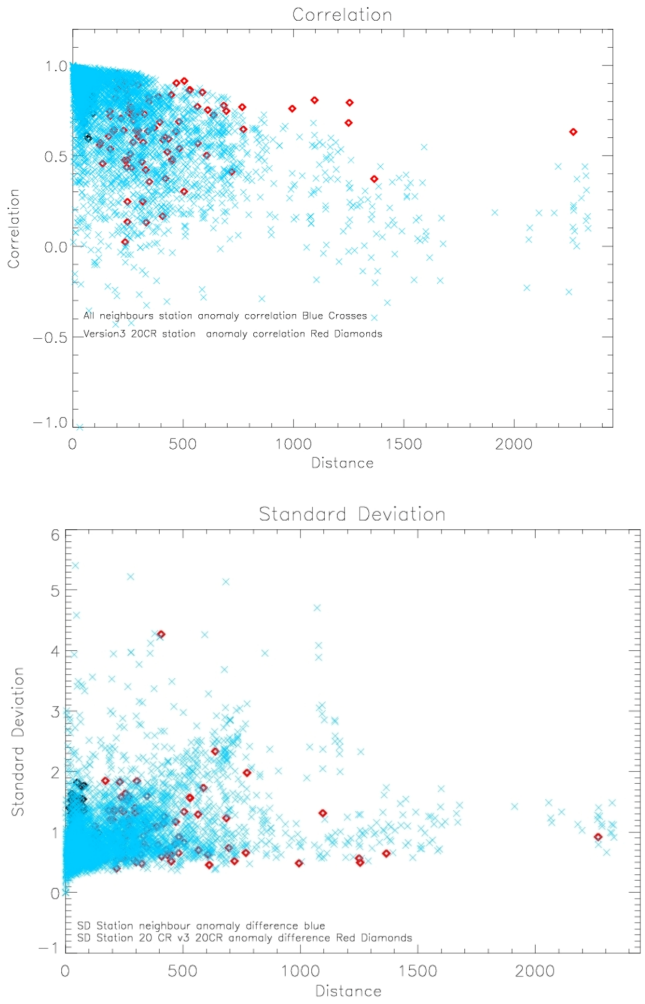


FIGURE 13 Neighbour comparisons show a clear tendency for the correlation to decrease and standard deviation to increase with distance, at least over the first 100 km or so between stations, although with noticeable spread. Top Panel is a pooled comparison of the correlations (r) (Blue x's) between each station and its 25 nearest neighbours across 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 test station to its neighbours is less than 350 km and, conversely 20CRv3 reanalysis performs better when the distances are c. 700 km or greater

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

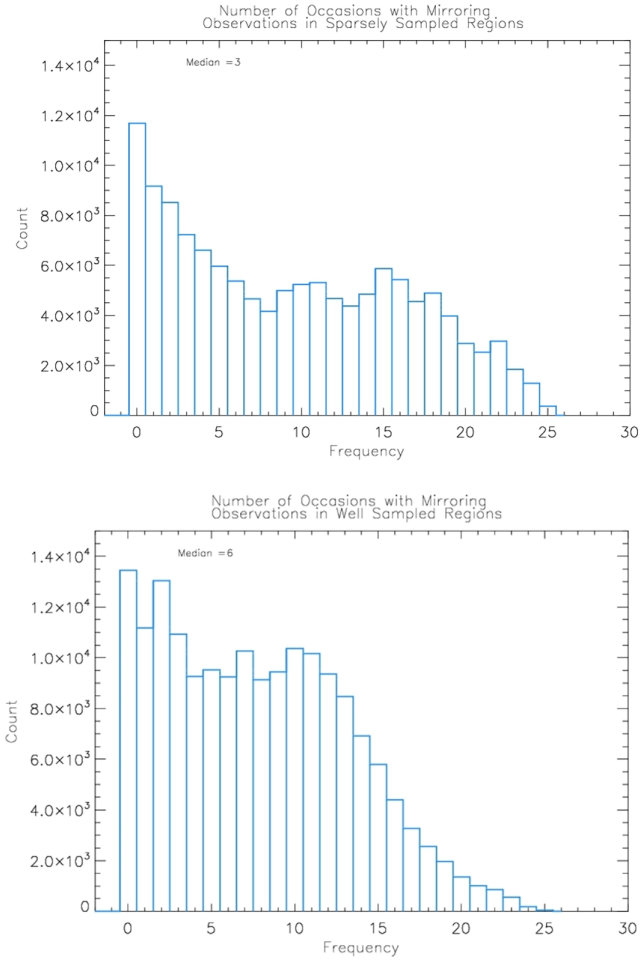


FIGURE 14 Histogram of the occurrence of the frequency of overlap between each station and its 25 nearest neighbours aggregated over the poorly sampled candidate station-neighbour pairs top panel and well sampled regions bottom panel and from 1850 to 2012. The most frequent occurrence is for no overlap occurs for both regions and the median value is six out of 25 comparisons being possible at any given timestep in well sampled regions, falling to three in poorly sampled regions

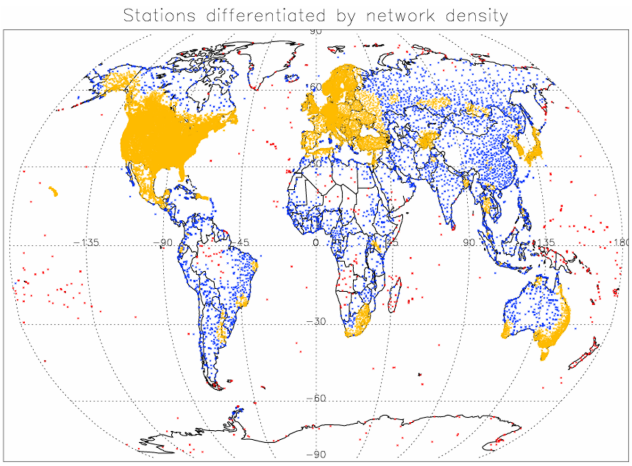


FIGURE 15 Stations with 25 or more stations within 350 km radius (yellow) for which pairwise approaches may be preferable. Stations with the 25 nearest neighbours within 700 km (blue) in which pairwise and 20CRv3 based approaches may be of comparable power according to the present analysis. Stations 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

Century, addressing previously stated concerns (Ferguson and Villarini, 2012).

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 700 km or more, while neighbour-based approaches are better at separation distances less than 350 km. This means that 20CRv3 is preferable for about 700 stations and competitive with neighbour-based approaches for a further c. 3,000 out of a total of 28,000+ stations of at least a decade duration in the ISTI databank. However, it is not the number of stations that matters, it is their spatial distribution. Figure 15 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 data set for homogenisation if the target is a global estimate of changes. Ongoing work is assessing the potential impact upon existing global surface temperature products by applying homogenisation approaches building upon Haimberger *et al.* (2012) using these series.

7 | CONCLUSION

Homogenisation of long-term records of land surface temperature is a challenging proposition. A well-characterized 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 utilize 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.

ACKNOWLEDGEMENTS

The ISTI databank was made available via NOAA NCEI. ERA-20C and CERA-20C were made available via ECMWF. Support for the Twentieth Century Reanalysis Project version 2c and version 3 datasets is provided by the U.S. DOE, Office of Science Biological Environmental Research (BER), by the NOAA Climate Program Office, and by the NOAA Physical Sciences Laboratory. The NOAA-CIRES-DOE Twentieth Century Reanalysis Project version 3 used resources of the National Energy Research Scientific Computing Center managed by Lawrence Berkeley National Laboratory which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 and used resources of NOAA's Remotely Deployed High Performance Computing Systems. GPC acknowledges support from the NOAA Physical Sciences Laboratory and NOAA Climate Program Office. IMG acknowledges PhD studentship support from the President of Maynooth University. Two anonymous reviewers are thanked for their reviews which helped sharpen the text.

ORCID

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

REFERENCES

Bronnimann, S., Martius, O., Franke, J., Stickler, A. and Auchmann, R. (2013) Historical weather extremes in the "Twentieth Century Reanalysis". *Geographica Bernensia*, G89, 7–17.

Callender, G. (1938) The artificial production of carbon dioxide and its influence on temperature. *Quarterly journal of the Royal Meteorological Society*, 64, 223–240.

Changnon, S.A. and Kunkel, K.E. (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*. Dublin, Ireland: ERSI.

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

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

Cornes, R.C., Van Der Schrier, G., Van Den Besselaar, E.J.M. and Jones, P.D. (2018) An ensemble version of the E-OBS temperature and precipitation data sets. *Journal of Geophysical Research: Atmospheres*, 123, 9391–9409.

Cowtan, K., Jacobs, P., Thorne, P. and Wilkinson, R. (2018) Statistical analysis of coverage error in simple global temperature estimators. *Dynamics and Statistics of the Climate System*, 3.

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., Luterbacker, J., Marshall, G., Maugeri, M., Mock, C., Mok, H., Nordi, O., Mark, R., Ross, T., Schuster, D., Srnec, L., Valente, M., Vizi, Z., Wanh, X., Westcott, N., Wollen, J. and Worley, S. (2015) The international surface pressure databank version 2. *Geoscience Data Journal*, 2, 31–46.

Ferguson, C. and Villarini, G. (2012) Detecting inhomogeneities in the Twentieth Century Reanalysis over the central United States. *Journal of Geophysical Research*, 117, 1–11.

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. and 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. and Sardesmukh, P.D. (2016) An ensemble of ocean reanalyses for 1815–2013 with sparse.pd. *Journal of Geophysical Research: Oceans*, 121, 6891–6910.

Giorgi, F. and Francisco, R. (2000) Evaluating uncertainties in regional climate. *The International Journal Climate Dynamics*, 16, 169–182.

Haimberger, L., Tavolato, C. and Sperka, S. (2012) Homogenization of the global radiosonde temperature dataset through combined comparison with reanalysis background series and neighboring stations. *Journal of Climate*, 25, 8108–8131.

Hansen, J. and Lebedeff, S. (1987) Global trends of measured surface air temperature.pdf. *Journal of Geophysical Research*, 92, 13,345–13,372.

Hawkins, E. and Jones, P.D. (2013) On increasing global temperatures: 75 years after Callendar. *Quarterly Journal of the Royal Meteorological Society*, 139, 1961–1963.

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., Fuenten, 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. and Thepaut, J.-N. (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal. Meteorological Society*, 146, 1–10.

Jones, P.D., Lister, D.H., Osborn, T.J., Harpam, C., Salmon, M. and Morice, C.P. (2012) Hemispheric and large scale land surface air temperature: An extensive revision and an update to 2010. *Journal of Geophysical Research*, 117, 16.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woolen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowlak, J., Mo, K. C., Ropelewski, C., Wank, J., Leetmaa, A., Reynolds, R., Jenne, R. and Joseph, D. (1995) The NCEP NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77, 437–471.

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. and Fiorino, M. (2001) The NCEP–NCAR 50-year reanalysis: monthly means CD–ROM and documentation. *Bulletin of the American Meteorological Society*, 82, 247–267.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K. and Takahashi, K. (2015) The JRA-55 reanalysis: general specifications and basic characteristics. *Journal of the Meteorological Society of Japan. Ser. II*, 93, 5–48.

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. and Dinan, S. (2018) CERA 20C a coupled reanalysis of the twentieth century. *Journal of Advances in Modeling Earth Systems*, 10, 1172–1195.

Mamara, A., Argiriou, A.A. and Anadranistakis, M. (2012) Homogenization of mean monthly temperature time series of Greece. *International Journal of Climatology*, 33, 2649–2666.

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

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

New, M., Hulme, M. and Jones, P. (1999) Representing twentieth century space time climate variability part II. *Journal of Climate*, 13, 2217–2237.

Parker, D. (1993) Effects of changes of exposure of thermometers at land stations. *International Journal of Climatology*, 14, 1–31.

Parker, W. (2016) Reanalyses and observations: what's the difference? *Bulletin of the American Meteorological Society*, 97, 1565–1572.

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

Peterson, T.C., 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. and Parker, D. (1998) Homogeneity adjustment of in situ atmospheric climate data: a review. *International Journal of Climatology*, 18, 1493–1517.

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. and Fisher, M. (2016) ERA-20C: an atmospheric reanalysis of the twentieth century. *Journal of Climate*, 29, 4083–4097.

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

Rayner, N., Brohan, P., Parker, D., Folland, C.K., Kennedy, J.J., Vanicek, M., Ansell, T.J. and 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.pdf. *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., Almeida, W.G.D., Christy, J.R., Flannery, M., Ishihare, M., Ishihara, M., Kamiguchi, K., Klein-Tank, A.M.G., Mhanda, A., Lister, D.H., Razuvaev, V., Renom, M., Mrustticucci, J.T., Worlry, S.J., Venma, V., Angel, W., Brunet, M., Dattore, B., Diamond, H., Lazzara, M.A., Blancq, F.L., Luterbacher, J., Machel, H., Revadekar, J., Vose, R.S. and Yin, X. (2014) The International surface temperature initiative global land surface databank: monthly temperature data release description and methods. *Geoscience Data Journal*, 10, 75–102.

Rohde, R., Muller, R., Jacobsen, R., Perlmutter, S. and Mosher, S. (2013) Berkeley earth temperature averaging process. *Geoinformatics & Geostatistics: An Overview* 01. 1(2), 1–13.

Sanchez-Lugo, A., Berrisford, P., Morice, C. and Nicolas, J.P. (2019) Global_Surface Temperature [in "State of the Climate in 2018"]. *Bulletin of The American Meteorological Society*, 100(9), S11–S12. https://doi.org/10.1175/2019BAMSStateoftheClimate.I.

Simmons, A.J., Berrisford, P., Dee, D.P., Hersbach, H., Hirahara, S. and 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, Laura C., Compo, Gilbert P., Whitaker, Jeffrey S., Sardeshmukh, Prashant D., Giese, Benjamin S., Mccoll, Chesley, Allen, Rob, Yin, Xungang, Vose, Russell, Titcher, Holly, Kennedy, John, Spencer, Lawrence J., Ashcroft, Linden, Bronnimann, Stefan, Brunet, Manola, Camuffo, Dario, Cornes, Richard, Cram, Thomas A., Crouthamel, Richard, Dominguez-Casto, Fernando, Freeman, J. Eric, Gergis, Joëlle, Hawkins, Ed, Jones, Philip D., Jourdin, Sylvie, Kaplan, Alexey, Kubota, Hisayuki, Blancq, Frank Le, Lee, Tsz-Cheung, Lorrey, Andrew, Luterbacker, Jürg, Maugeri, Maurizio, Mock, Cary J., Moore, G. W. Kent, Przybylak, Rajmund, Pubmenzky, Christa, Reason, Chris, Slonosky, Victoria C., Smith, Catherine A., Tinz, Birger, Trewin, Blair, Valente, Maria Antónia, Wang, Xiaolan L., Wilkinson, Clive, Wood Kevin Wyszyriski, Przemysław 2019. Towards a more reliable historical reanalysis. 1_41.

Thorne, P., 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. and Rennie, J. (2016) Reassessing changes in diurnal temperature range.pdf. *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. and 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. and Willett, K.M. (2018) Towards a global land surface climate fiducial reference measurements network. *International Journal of Climatology*, 38, 2760–2774.

Thorne, P.W., Parker, D.E., Simon, F.B.T., Jones, P.D., Mccarthy, M.P., Coleman, H. and Brohan, P. (2005) Revisiting radiosonde upper air temperatures from 1958 to 2002. *Journal of Geophysical Research*, 110, 1–17.

Titchner, H.A. and 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.

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

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

Vicente-Serrano, S.M., Angel Saz- Sanchez, M. and 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.

Wang, J., Xu, C., Hu, M., Li, Q., Yan, Z. and Jones, P. (2018) Global land surface air temperature dynamics since 1880. *International Journal of Climatology*, 38, e466–e474.

Willett, K., Willuams, 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. and 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.J. and 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. and Robeson, S.M. (1995) Climatologically aided interpolation (CAI) of terrestrial air temperature. *International Journal of Climatology*, 15, 221–229.

Zhou, C., He, Y. and Wang, K. (2018) On the suitability of current atmospheric reanalyses for regional warming studies over China. *Atmospheric Chemistry and Physics*, 18, 8113–8136.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Gillespie IM, Haimberger L, Compo GP, Thorne PW. Assessing potential of sparse-input reanalyses for centennial-scale land surface air temperature homogenisation. *Int J Climatol*. 2021;41 (Suppl. 1):E3000–E3020. https://doi.org/10.1002/joc.6898