

## The international surface temperature initiative global land surface databank: monthly temperature data release description and methods

J. J. Rennie<sup>1\*</sup>, J. H. Lawrimore<sup>2</sup>, B. E. Gleason<sup>2</sup>, P. W. Thorne<sup>1,3</sup>, C. P. Morice<sup>4</sup>, M. J. Menne<sup>2</sup>, C. N. Williams<sup>2</sup>, W. Gambi de Almeida<sup>5</sup>, J.R. Christy<sup>6</sup>, M. Flannery<sup>7</sup>, M. Ishihara<sup>8</sup>, K. Kamiguchi<sup>8</sup>, A. M. G. Klein-Tank<sup>9</sup>, A. Mhanda<sup>10</sup>, D. H. Lister<sup>11</sup>, V. Razuvaev<sup>12</sup>, M. Renom<sup>13</sup>, M. Rusticucci<sup>14</sup>, J. Tandy<sup>4</sup>, S. J. Worley<sup>15</sup>, V. Venema<sup>16</sup>, W. Angel<sup>2</sup>, M. Brunet<sup>11,17</sup>, B. Dattore<sup>15</sup>, H. Diamond<sup>2</sup>, M. A. Lazzara<sup>18</sup>, F. Le Blancq<sup>19</sup>, J. Luterbacher<sup>20</sup>, H. Mächel<sup>21</sup>, J. Revadekar<sup>22</sup>, R. S. Vose<sup>2</sup> and X. Yin<sup>23</sup>

<sup>1</sup>Cooperative Institute for Climate and Satellites – NC, North Carolina State University and NOAA's National Climatic Data Center, Asheville, NC, USA

<sup>2</sup>NOAA's National Climatic Data Center, Asheville, NC, USA

<sup>3</sup>Nansen Environmental and Remote Sensing Center, Bergen, Norway

<sup>4</sup>Met Office Hadley Centre, Exeter, UK

<sup>5</sup>Instituto Nacional de Pesquisas Espaciais, Centro de Previsão de Tempo e Estudos Climáticos, Cachoeira Paulista/SP, Brazil

<sup>6</sup>University of Alabama-Huntsville, Huntsville, AL, USA

<sup>7</sup>Bureau of Meteorology, Melbourne, Australia

<sup>8</sup>Japan Meteorological Agency, Tokyo, Japan

<sup>9</sup>Royal Netherlands Meteorological Institute (KNMI), De Bilt, the Netherlands

<sup>10</sup>African Centre of Meteorological Applications for Development, Niamey, Niger

<sup>11</sup>Climatic Research Unit, University of East Anglia, Norwich, UK

<sup>12</sup>Russian Research Institute of Hydrometeorological Information, Obninsk, Russia

<sup>13</sup>Universidad de la Republica, Montevideo, Uruguay

<sup>14</sup>University of Buenos Aires, Buenos Aires, Argentina

<sup>15</sup>National Center for Atmospheric Research, Boulder, CO, USA

<sup>16</sup>Meteorologisches Institut Universität Bonn, Bonn, Germany

<sup>17</sup>Centre for Climate Change, University Rovira i Virgili, Tarragona, Spain

<sup>18</sup>Antarctic Meteorological Research Center, Space Science and Engineering Center, University of Wisconsin-Madison, Madison, WI, USA

<sup>19</sup>States of Jersey Meteorological Department, Jersey, UK

<sup>20</sup>Department of Geography, Justus-Liebig University, Giessen, Germany

<sup>21</sup>Deutscher Wetterdienst (DWD), Offenbach am Main, Germany

<sup>22</sup>Centre for Climate Change Research, Indian Institute of Tropical Meteorology, Pune, India

<sup>23</sup>ERT, Inc., Asheville, NC, USA

\*Correspondence: J. J. Rennie, Cooperative Institute for Climate and Satellites – NC, North Carolina State University and NOAA's National Climatic Data Center, Asheville, NC, USA, E-mail: jared.rennie@noaa.gov

[Correction added 2 January 2015 after original online publication: the affiliation for W. Angel, H. Diamond and R. S. Vose has been corrected to 'NOAA's National Climatic Data Center, Asheville, NC, USA'.]

Described herein is the first version release of monthly temperature holdings of a new Global Land Surface Meteorological Databank. Organized under the auspices of the International Surface Temperature Initiative (ISTI), an international group of scientists have spent three years collating and merging data from numerous sources to create a merged holding. This release in its recommended form consists of over 30 000 individual station records, some of which extend over the past 300 years. This article describes the sources, the chosen merge methodology, and the resulting databank characteristics. Several variants of the databank have also been released that reflect the structural uncertainty in merging datasets. Variants differ in, for example, the order in which sources are considered and the degree of congruence required in station geolocation for consideration as a merged or unique record. Also described is a version control protocol that will be applied in the event of updates. Future updates are envisaged with the addition of new data sources, and with changes in processing, where public feedback is always welcomed. Major updates, when necessary, will always be accompanied by a new journal paper. This databank release forms the foundation for the construction of new global land surface air temperature analyses by the global research community and their assessment by the ISTI's benchmarking and assessment working group.

*Geosci. Data J.* 1: 75–102 (2014), doi: 10.1002/gdj3.8

Received: 16 April 2013, revised: 26 September 2013, accepted: 3 October 2013

Key words: global climate, surface land temperature, statistical analysis

## Dataset

Identifier: doi:10.7289/V5PK0D3G

Creator: International Surface Temperature Initiative (ISTI)

Title: Global Land Surface Meteorological Databank

Publisher: Global Observing Systems Information Center (GOSIC)

Publication year: 2014

## Introduction and rationale

Since the 17th century, starting with a few sites in Europe (Camuffo & Bertolin, 2012 and references therein), efforts have been established to continuously and systematically measure land surface air temperature through instrumental means. These long-term records provide an insight into the temperature variations of the Earth. Major efforts were made in the 1980s and 1990s to collect these observations from around the world and create a consolidated monthly-timescale database on a global scale. Climatologists from the National Oceanic and Atmospheric Administration's (NOAA's) National Climatic Data Center (NCDC) and the Carbon Dioxide Information Analysis Center (CDIAC) produced the Global Historical Climatology Network – Monthly (GHCN-M) dataset in 1992, which contained more than 6000 stations (Vose *et al.*, 1992). A second version of GHCN-M, containing 7280 stations with monthly mean, maximum and minimum temperature ( $T_{AVG}$ ,  $T_{MAX}$ , and  $T_{MIN}$  respectively), was released in 1997 (Peterson & Vose, 1997). More recently in 2011, a third version of GHCN-M updated the quality control (QC) procedures, as well as the algorithm used to identify and account for inhomogeneities (Lawrimore *et al.*, 2011). Routine updates for about 2000 stations are made on a daily basis. Since version 2, GHCN-M has been a major component in the development of NASA's GISS dataset (Hansen *et al.*, 1999, 2010), which contains 6000 stations. An independent effort in the United Kingdom produced a first release of its CRUTEM product in the late 1980s. Today, a global dataset of over 6000 stations is still maintained in its fourth iteration (Jones *et al.*, 2012). Although methodologies for all three datasets differ, they all exhibit close agreement with respect to large scale changes in global land surface air temperatures. Such series can be reasonably calculated from a point in the mid-to-late 19th century when there is sufficient global station coverage.

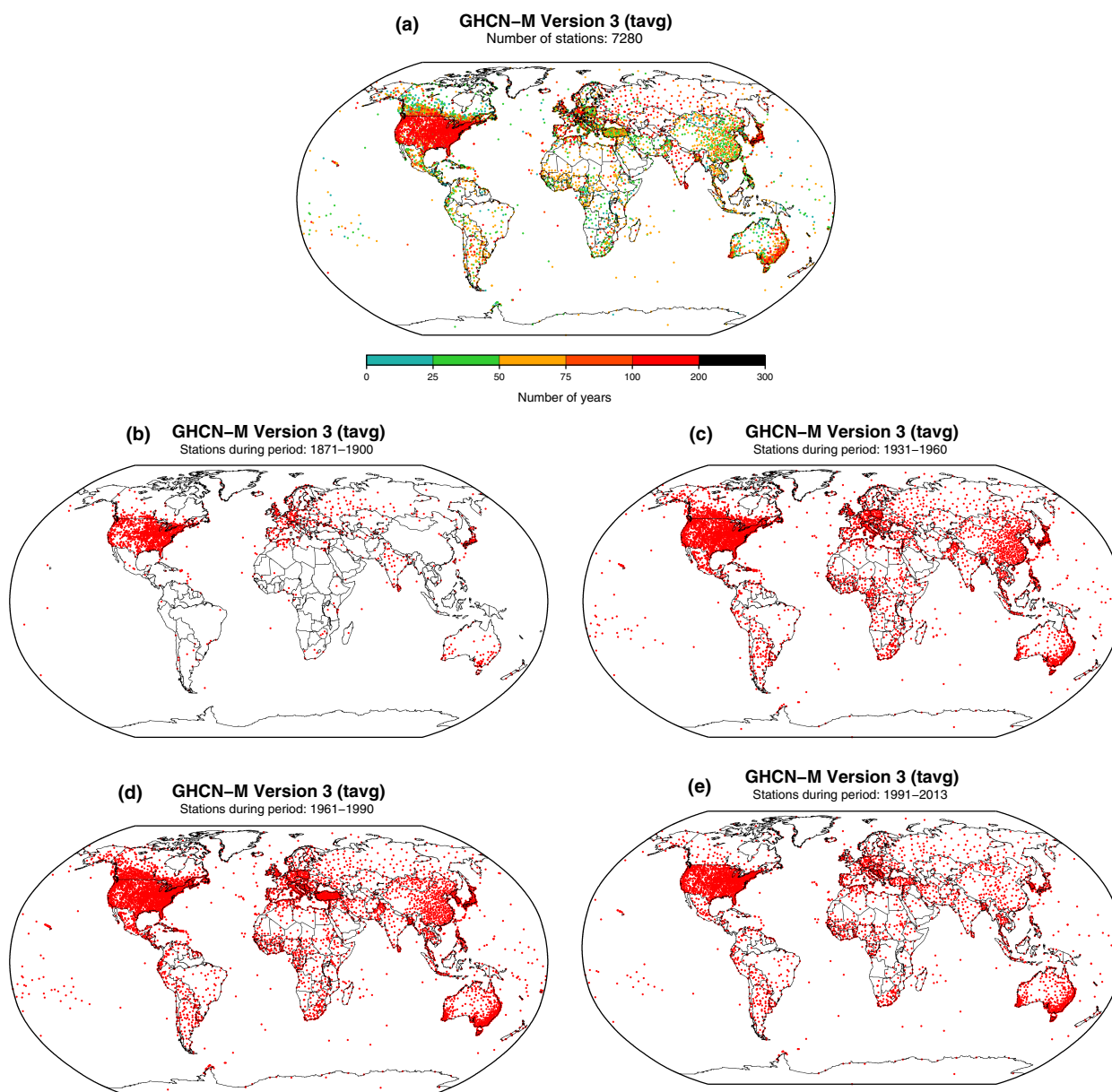
Land surface air temperature products have been essential for monitoring the evolution of the climate system, and are included in reports such as annual State of the Climate (Blunden & Arndt, 2013), national assessments (USGCRP, 2014), and the Intergovernmental Panel on Climate Change (IPCC, 2013). Taken together with sea surface temperatures, they repre-

sent the longest continuous direct measurement record by which to monitor and understand climate variability and change.

More recently attention has turned to the construction of global daily land surface temperature datasets. This was in recognition of the need to characterize sub-monthly variability, and in particular climatic extremes. The Global Historical Climatology Network – Daily (GHCN-D) dataset was a result of these efforts (Menne *et al.*, 2012). Today, GHCN-D provides daily maximum and minimum temperature for nearly 30 000 stations. Although more stations exist on the daily scale, records are generally shorter in length than the monthly mean temperature records in GHCN-M. GHCN-D data for most stations outside the United States do not begin until the middle of the 20<sup>th</sup> Century. Nevertheless this provides the world's most complete record of daily variations and extremes not available in the monthly climate record. There also exist numerous other national and regional daily holdings as well as more complete indices datasets where the station data have not been shared, but derived indices are (Caesar *et al.*, 2006; Allan *et al.*, 2011; Compo *et al.*, 2011; Skansi *et al.*, 2012). This reflects some of the real challenges over data sharing and provision between rights holders.

Although there have been tremendous advances in the understanding of climate change provided by these data collection efforts, analyses, and resulting datasets, there remain substantive spatial and temporal gaps due to deficiencies in global collections of data. These deficiencies have deleterious impacts on our collective ability to monitor and characterize climate (Figure 1). There is limited spatial coverage in many parts of the world, especially in the 1800s and earlier. Additional sources of data exist, often in original manuscript form or scanned images. These forms are often housed in designated data centres such as NCDC, however lack of resources and funding have prevented efforts to convert the data into digital formats for use in modern datasets.

In addition there has been limited success at completely documenting the provenance and implementing version control from the point measurement through dissemination and data sharing pathways, QC, bias correction, and archive and access. More can



**Figure 1.** Station locations with at least 1 month of data in GHCN-M version 3 (a). The colour corresponds to the number of years of data available for each station. Station locations during the periods 1871–1900 (b), 1931–1960 (c), 1961–1990 (d), and 1991–2013 (e) are also shown.

be done to improve practices to ensure full openness, transparency, and availability of data and the details associated with each processing step. By putting in place such practices the wider community will have the opportunity to more fully engage in the process of improving data practices.

To address these issues scientists from the climate, statistics, and geoscience communities have come together to establish the International Surface Temperature Initiative (ISTI) (Thorne *et al.*, 2011). The first goal of the initiative is the establishment of the global land surface databank described herein, which will provide the most comprehensive possible set of data holdings brought together in a consistent and traceable manner.

The ISTI Steering Committee was formed and they convened a Databank Working Group (DWG) to oversee the development and management of the databank. The process builds on past efforts to construct a new global land surface dataset, paying special attention to ensure users can fully understand the provenance of the data in the merged holding to the extent that it is known. Its openness and transparency will be made certain by the release of all data, meta-data, and software code for public access.

The remainder of this article provides a detailed explanation regarding the creation of the databank. Section 2 describes the overall design of the databank. The merge method is highlighted in Section 3. Section 4 summarizes the resulting dataset characteristics

from the recommended merge and several other variants. Section 5 describes the version control protocols. Conclusions are presented in Section 6, along with avenues for increased expert participation.

## 1. Databank architecture

This section provides information about the design of the global databank and includes a description of all the stages, as well as the implementation of provenance tracking flags.

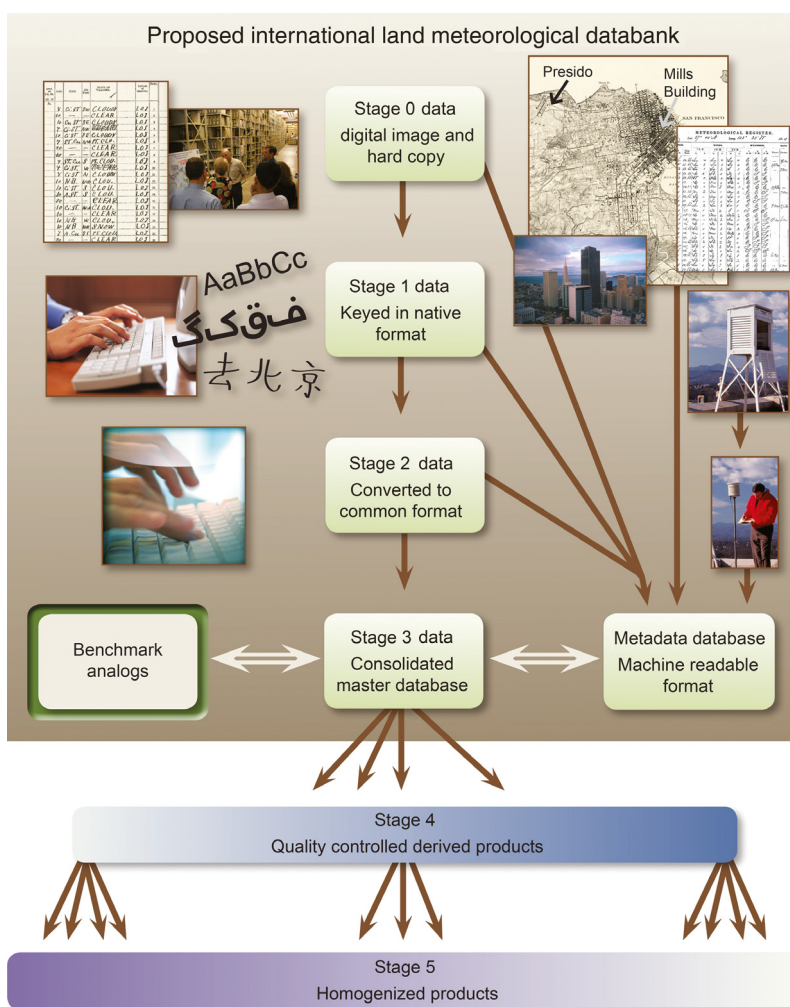
### 1.1. Databank stages

The databank design includes 6 data stages, starting from the original observation to the final quality controlled and bias corrected products (Figure 2). The initial focus is on the collection of temperature data on the sub-daily, daily, and monthly timescale, although other elements and timescales will be added to the databank as they become available.

Stage 0 consists of digital observations in their original form. The historical record consists primarily of

observations recorded on paper. In recent years, there have been efforts to automate weather and climate networks or develop capabilities to digitally report observations, however, many networks continue to rely on manual observations and paper records. Over the past decade, programs such as NOAA's Climate Database Modernization Program (CDMP) and the International Environmental Data Rescue Organization (IEDRO) have converted these records to photographic or scanned images. Such images are essential to ensuring the preservation of the original observations. Because of their importance, these images are hosted on the databank server where secure third party hosting is not possible. The original paper forms or images are otherwise archived at National Meteorological Services and other designated repositories. In many cases, however, the location of the original form remains unknown.

Stage 1 contains digitized data, in their native format, provided by the contributor. No effort is required on their part to convert the data into any other format. This reduces the possibility that errors could occur during translation and permits easy retranslation



**Figure 2.** Schematic summary of the current structure of the global land surface temperature databank and its relationship to engaged metadata holdings and analogs used to perform performance benchmarking activities. Figure courtesy NCDRC graphics team.



in the event such issues occur. Because a large percentage of stage 0 data are missing or unknown, stage 1 data will be the first data level for many sources.

Once the data are submitted as stage 1, all data are converted into a common stage 2 format. In addition, data provenance flags are added to every observation to provide a history of that particular value. stage 2 files are maintained in ASCII format and each data source is held in a separate subdirectory. An inventory file is produced for each source dataset containing available metadata. At a minimum this typically consists of a station id, name, latitude, longitude, elevation, and beginning and ending year of data. The code to convert all the sources to stage 2 is written in the Perl scripting language, and is made available on the databank server. This provision documents the data translation and provides a means for other researchers to evaluate the process and if errors are detected later they can easily be addressed.

At the time of writing, there are 58 sources that have been collected and converted to stage 2. Table 1 summarizes all the sources currently in the databank. Some sources only provide data on monthly time-scales, while others have daily or even sub-daily data. Most station records include maximum and minimum temperature, but in some cases only average temperatures are available. The databank policy recommends data to be submitted in stage 0 and stage 1 form, prior to any QC or bias correction. If no other forms exist, then quality controlled and bias corrected data are accepted. The status of the original source data is documented with provenance fields on each record and they have an important role in setting the hierarchy for inclusion during the data merging process.

Given the historical nature of data creation, sharing, and rescue, there are many cases where a single station exists in multiple data sources. Figure 3 summarizes the current geographical distribution and record length of all the stations in stage 2. Because of the possibility of duplicate station records, these stations are considered non-unique. In addition, due to different collection and reprocessing techniques, the duplicate records do not necessarily have identical temperature values for the same station even though they are based upon the same fundamental measurements. Figure 4 provides different values of daily average temperature for the same station using multiple calculating methods.

Sources with daily data have been converted to monthly averages and added as additional sources in the monthly holding. The methodology for creating a monthly average follows that of the World Meteorological Organization (WMO). For any given month, there must be no more than five missing days, and there must be no more than three consecutive days missing (WMO, 2011). If one or both of the criteria fail, the monthly value is set to missing.

Once all the sources have been collected and formatted, the data are then merged into a single,

comprehensive stage 3 dataset. The algorithm that performs the merging is described in Section 3. Development of the merged dataset is followed by QC and homogeneity adjustments (stage 4 and 5 respectively). Development of these last two stages is not the responsibility of the DWG, and is not discussed here beyond to say that it is hoped and strongly encouraged that multiple independent groups will undertake such efforts (Thorne *et al.*, 2011).

## 1.2. Data provenance

To provide a traceable record that documents as much of the histories of each observation as possible, data provenance tracking flags are added to each observation, beginning with stage 2. Flags are a three digit numeric value that represents unique information regarding each observation. Currently there are seven data provenance tracking flags (Tables 2–8), with the opportunity to add additional flags as needed. All flags are extendable for future situations using additional three digit numbers not previously assigned.

The first two flags (Tables 2 and 3) describe the source of stage 0 and stage 1 data respectively. The stage 1 source may differ from the stage 0 providers, or it may provide additional formation such as the name of the host's dataset from which the data originated. The next flag (Table 4) indicates if the data provided by the host were in its raw form or whether it had been previously quality controlled or bias corrected. Although the preference is to have data as raw as possible, there are times where such data do not exist, or have not been provided to the databank. Therefore pre-processed data are accepted and the appropriate flag is assigned to the observations. Table 5 describes the method and location of data digitization, if available. Two flags describe the calculation of data on the daily and monthly timescale (Tables 6 and 7 respectively). The daily calculation flag (Table 6) depicts how the observation was generated if the source contains sub-daily values. Depending on the type of station, observations can be made and reported on an hourly basis. However some stations, such as many of those providing synoptic observations, only report observations four times a day. Within the databank, a daily value is calculated during the conversion to stage 2, and a flag is added to inform the user how many observations were available prior to the conversion. Similarly, a monthly calculation flag (Table 7) informs the user how many days in a month were used to create a monthly average. The final flag (Table 8) describes the process used to transfer the data to the databank servers.

## 2. Merging methodology

This section covers, at a high level, the process in which individual stage 2 sources are combined to form a comprehensive stage 3 dataset. At the time of writing, only monthly sources, as well as monthly

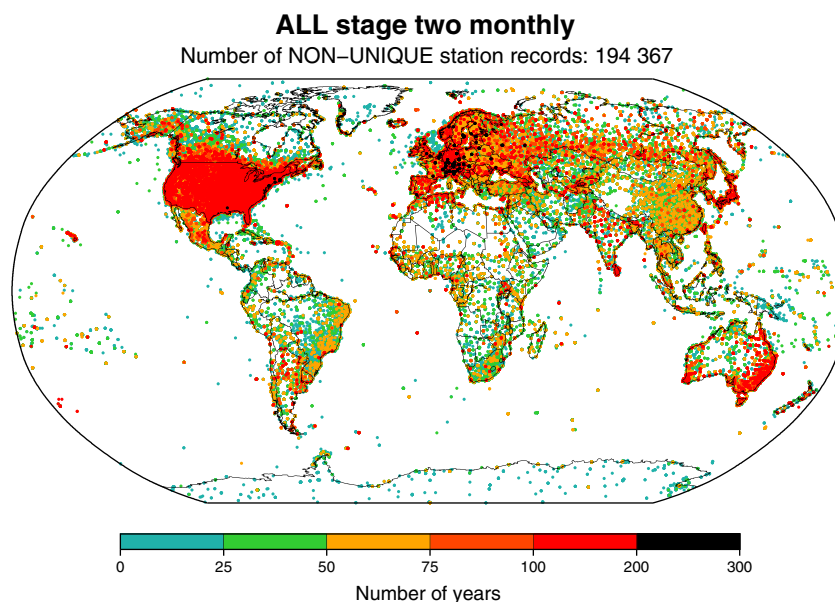
**Table 1.** Summary of sources that have been converted to Stage Two data ordered.

Name	Source	Time scale	Raw/QC/ Homogenized	TMAX	TMIN	TAVG
Antarctica	SCAR Reader Project	MONTHLY	Raw	N	N	Y
Antarctica (AWS)	Antarctic Meteorological Research Center	DAILY	Raw	Y	Y	N
Antarctica (Palmer Station)	Antarctic Meteorological Research Center	DAILY	Raw	Y	Y	Y
Antarctica (South Pole Station)	Antarctic Meteorological Research Center	MONTHLY	Raw	Y	Y	Y
Arctic	IARC/Univ of Alaska Fairbanks	MONTHLY	Homogenized	N	N	Y
Argentina	National Institute of Agricultural Technology (INTA)	DAILY	Raw	Y	Y	N
Australia	Australia Bureau of Meteorology	DAILY	Homogenized	Y	Y	Y
Brazil	INPE, Nat. Institute for Space Research	DAILY	Raw	Y	Y	N
Brazil-Inmet	INMET	DAILY	Raw	Y	Y	N
Canada	Environment Canada	MONTHLY	Homogenized	Y	Y	Y
Canada	Environment Canada	MONTHLY	Raw	Y	Y	Y
Central Asia	NSIDC	MONTHLY	Homogenized	Y	Y	Y
Channel Islands	States of Jersey Met	DAILY	Raw	Y	Y	N
Colonial Era Archives	Griffith	MONTHLY	Raw	Y	Y	N
CRUTEM4	UKMO	MONTHLY	Homogenized	N	N	Y
East Africa	Univ. of Alabama Huntsville	MONTHLY	Raw	Y	Y	Y
Ecuador	Inst. Nacional De Met E Hidrologia	DAILY	Raw	Y	Y	N
Europe/N. Africa	European Climate Assessment (Daily, Non-Blended)	DAILY	Raw	Y	Y	Y
Europe/N. Africa	European Climate Assessment (Daily, Blended)	DAILY	QC	Y	Y	Y
Europe/N. Africa	European Climate Assessment (Monthly)	MONTHLY	QC	Y	Y	Y
Germany	DWD- Germany	MONTHLY	Raw	N	N	Y
GHCN-Daily	NCDC	DAILY	QC	Y	Y	N
GHCN-M v2	NCDC	MONTHLY	QC	Y	Y	Y
GHCN-M v2 Source	NCDC	MONTHLY	Raw	N	N	Y
Giessen	University of Giessen	DAILY	Raw	Y	Y	N
Global Summary of the Day	NCDC	DAILY	Raw	Y	Y	N
Greater Alpine Region	Histalp/ZAMG	MONTHLY	Homogenized	N	N	Y
Greenland	NCAR	DAILY	Raw	Y	Y	N
HadISD	UKMO	DAILY	QC	Y	Y	N
India	India Meteorological Department	DAILY	Raw	Y	Y	N
Japan	JMA	DAILY	QC	Y	Y	Y
Max/Min Stations from R. Vose	NCDC	MONTHLY	Raw	Y	Y	N
Mexico	CDMP	DAILY	Raw	Y	Y	N
Mon. Clim Data of World (MCDW)	NCDC	MONTHLY	Raw	N	N	Y
MCDW (Completed, unpublished)	NCDC	MONTHLY	Raw	N	N	Y
Mon. Surf. Station Clim. (WMSSC)	NCAR	MONTHLY	Raw	N	N	Y
Norway	Norwegian Meteorological Institute	MONTHLY	QC	Y	Y	Y
Pitcairn Island	Met Service of New Zealand	DAILY	Raw	Y	Y	N
Polar	ISPD	DAILY	Raw	N	N	Y
Preliminary CLIMAT	NCDC	MONTHLY	Raw	Y	Y	Y
Russia	Roshydromet	DAILY	QC	Y	Y	Y
Southeast Asia	Southeast Asia Climate Assessment (Non-Blended)	DAILY	Raw	Y	Y	Y
Southeast Asia	Southeast Asia Climate Assessment (Blended)	DAILY	QC	Y	Y	Y
Spain	Univ. Rovira I Virgili	DAILY	QC	Y	Y	Y
Sweden	GCOS Surface Network	DAILY	Raw	Y	Y	Y
Switzerland	ISPD	DAILY	Raw	N	N	Y

(continued)

Table 1. (continued).

Name	Source	Time scale	Raw/QC/			
			Homogenized	TMAX	TMIN	TAVG
Switzerland	Digihom/MetoSwiss/IAC-ETH	DAILY	QC	Y	Y	Y
Sydney	ISPD	DAILY	Raw	N	N	Y
Tunisia/Morocco	ISPD	DAILY	QC	Y	Y	Y
Uganda	Univ. of Alabama Huntsville	MONTHLY	Raw	Y	Y	Y
UK CLIMAT	UKMO	MONTHLY	Raw	Y	Y	Y
UK met office historical	UKMO	MONTHLY	QC	Y	Y	N
Uruguay	Universidad de la Republica, Montevideo, Uruguay	DAILY	QC	Y	Y	N
Uruguay	Inst. Nacional de Invest Agropecuaria	DAILY	QC	Y	Y	Y
US CLIMAT	NCDC	MONTHLY	Raw	Y	Y	Y
US Forts	CDMP	DAILY	Raw	Y	Y	N
Vietnam	CDMP	DAILY	Raw	Y	Y	N
World weather records	WMO	MONTHLY	Raw	Y	Y	Y



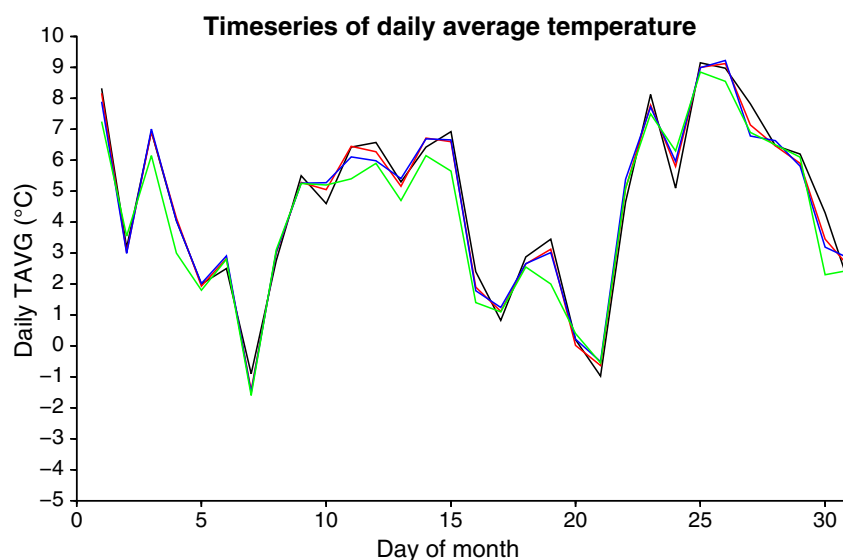
**Figure 3.** Location of all stations in the stage 2 components of the databank. The colour corresponds to the number of years of data available for each station. Stations with longer periods of record mask stations with shorter periods of record when they are in approximate identical locations.

averages calculated from daily sources, are considered. A similar daily merged product under the auspices of the ISTI is planned for the future. Using a quasi-probabilistic approach, the algorithm attempts to mimic the decisions an expert analyst would make manually. Because many sources may contain records for the same station it is necessary to create a process for identifying and removing duplicate stations, merging some sources to produce a longer station record, and in other cases determining when a station should be brought in as a new and distinct record. It is also necessary to place aside (withhold) some records where it remains unclear whether to

merge the record or create a new unique record. The following methods describe the decisions used to create a merged product recommended and endorsed by the ISTI (results are discussed in Section 4.1). However, to characterize uncertainty, changes to these methods are also investigated and provided for public access (Section 4.2).

## 2.1. Prioritizing the sources in stage 2

Before a merge is performed, a hierarchy of all the source datasets within the databank is created. Sources with higher priority take precedence over



**Figure 4.** Time series of daily average temperature for December 2011, taken from hourly observations from HadISD station number 014150-99999 (STAVANGER/SOLA). Averages were calculated using the main standard times for surface synoptic observations (black), main and intermediate standard times (red), all hours available (blue), and the average of maximum and minimum (green).

**Table 2.** List of sources for stage 0 files.

Flag	Description
101	Paper, NCDC
102	Paper, JMA
103	Paper, Australian BOM
104	Paper, Met Service of New Zealand
105	Paper, Royal Netherlands Meteorological Institute (KNMI)
201	Images, University Rovira i Virgili, Centre for Climate Change
301	Images, databank stage 0 FTP Site
302	Images, EDADS website, NCDC
303	Images, NOAA Library Website
999	Missing/Unknown/Not applicable

lower priority sources when more than one record for the same station and same period of time exists. The priority that one source may have over another is based on a number of criteria. Sources that have better data provenance, extensive metadata, come from a national weather or hydrological service, or have long and consistent periods of record are the most desirable and are assigned higher priority.

Because of the emphasis the ISTI places on data provenance, the stage 3 databank holdings are envisaged to constitute as close to the raw data as possible. Ideally data should be tracked as far back as the hard copy form on which the observation was first recorded. Sources for which such records are available are given higher priority during the merge process. In addition, data rescued recently when the importance of such provenance has been explicitly recognized are

given high priority. Other data given high preference during merging include sources with monthly mean maximum and minimum temperatures. This is preferred over monthly mean temperature because they can be directly used to calculate the monthly mean and because there is compelling evidence that many data artefacts affect maximum and minimum differently (Williams *et al.*, 2012).

Using those principles (see Appendix A for a more detailed description), 49 of 58 stage 2 sources were prioritized (Table 9). Note that not all sources (Table 1) are used, due to either their inclusion in GHCN-D, or issues that arose over their data quality. GHCN-D was selected to be the highest priority, or target dataset, and the monthly dataset derived from it is the starting point for the merge process. GHCN-D is regularly reconstructed, usually every weekend, from its 25-plus data sources to ensure it is generally in sync with its growing list of constituent sources. Many of these sources are provided directly by national holdings and include comprehensive contributions of daily holdings from the United States, Canada, and Australia as well as a number of other large regional collections. (Menne *et al.*, 2012). Furthermore, since GHCN-D provides a backbone of daily maximum and minimum temperatures, it is also a logical dataset to develop a new daily merged dataset in the future. This coherency between monthly and daily datasets developed in the ISTI is important as finer temporal research questions are considered.

It is important to note that even though there is a chosen source hierarchy, based on a set of priorities, other plausible hierarchies can be created. Some experiments with different priority settings have been run and are described in Section 4.2. Other decisions



**Table 3.** List of sources for stage 1 files.

Flag	Description
100	NCDC International Collection
101	High Plains Regional Climate Center, USA
102	NCDC DSI-3200
103	NCDC DSI-3206
104	University Rovira i Virgili, Centre for Climate Change, Spain
105	NCDC CDMP Digital Archive
106	Japan Meteorological Agency
107	Met Service of New Zealand
108	European Climate Assessment & Data Project
109	University of Alabama: Huntsville
110	Antarctic Meteorological Research Center
111	Meteo France
112	National Institute for Space Research (INPE), Brazil
113	MeteoSwiss, Switzerland
114	Nicholas Copernicus University IPY Collection, Poland
115	University of Melbourne, Australia
116	Met Office, United Kingdom
117	INIA (Instituto Nacional de Investigacion Agropecuaria), Uruguay
118	Australian BOM
119	Environment Canada
120	International Arctic Research Center: University of Alaska, Fairbanks
121	Central Institute for Meteorology and Geodynamics (ZAMG), Austria
122	National Snow and Ice Data Center (NSIDC), USA
123	Instituto Nacional de Meteorologia e Hidrologia, Ecuador
124	Scientific Committee on Atmospheric Research
125	Databank stage 2 Daily source (converted to monthlies)
126	Databank stage 3 Daily source (converted to monthlies)
127	Databank stage 4 Daily source (converted to monthlies)
128	States of Jersey Meteorological Department
129	Meteo Russia (RIHMI-WDC)
130	University of Giessen, Department of Geography, Germany
131	CISL Research Data Archive, USA
132	INTA (National Institute of Agricultural Technology), Argentina
133	India Meteorological Department
134	DWD (Deutscher Wetterdienst): Germany
135	Universidad de la Republica, Montevideo, Uruguay
136	Norwegian Meteorological Institute
999	Missing/Unknown/Not applicable

may lead to a different hierarchy and can alter the result. Since the databank aims to be open and transparent, others are encouraged to evaluate the hierarchies we have tried, establish their own list of preferred sources, run the source code available online, and to make results available for comparison.

**Table 4.** List of type of data sent by source.

Flag	Description
101	Raw
102	Quality controlled by originator
103	Homogenized by originator
999	Missing/Unknown/Not applicable

**Table 5.** List of methods of digitization of data.

Flag	Description
101	Keyed, Source Corp
102	Keyed, CDMP
103	Keyed, CDMP Forts Project
104	Keyed, Local originator
000	Auto collect
999	Missing/Unknown

**Table 6.** List of methods of calculating hourly data into daily.

Flag	Description
101	Data values original
102	Daily value calculated from main standard synoptic observations (00,06,12,18 UTC)
103	Daily value calculated from main and intermediate synoptic observations (00,03,06,09,12,15,18,21 UTC)
104	Daily value calculated from other sub-daily observations (at least 3 obs available)
105	Daily value calculated from other sub-daily observations (at least 20 obs available)
999	Missing/Unknown/Not applicable

**Table 7.** List of methods for calculating daily data into monthly.

Flag	Description
000	Data values original
001–031	Monthly value calculated from daily average (number indicates number of days available)
999	Missing/Unknown/Not Applicable

## 2.2. Overall description of merge program

The merge process is accomplished in an iterative fashion, starting from the highest priority data source (target) and running progressively through the other sources (candidates). The merge process is designed to be based upon metadata matching and data equivalence criteria. A general overview work flow chart of the algorithm can be found in Figure 5. The algorithm was written so that the

**Table 8.** List of methods of transmission of data to ISTI.

Flag	Description
101	Mail
102	E-Mail
103	FTP
104	SRRS FTP
105	NOAA port
106	NMHS web service
107	Telephone modem
108	Direct Datalogger download/PDA
109	Other satellite
999	Unknown

entire program is fully automated. The highest source is read in and compared to the source of next lower priority. Every candidate station is compared to all target stations, and one of three possible decisions is made. First, when a station match is found the candidate station is merged with the target station. Second, if the candidate station is determined to be unique it is added to the target dataset as a new station. Third, the available information is insufficient, conflicting, or ambiguous, and the candidate station is withheld. After all stations in the two sources are tested and combined into a new merged source dataset, the process then applies the same tests using the next lower ranked source in the priority hierarchy. Due to heavy computational requirements required for the algorithm to run, the code was written in FORTRAN 95. As with other software used to develop the databank, it is made available on the databank ftp site, along with access to a free compiler.

### 2.3. Station error checking

When a candidate station is first read in, it is checked against a list of stations with known issues in its data and metadata. If this step is not performed, stations with erroneous information may tamper results, or even be incorrectly recognized as unique. For example, a station having correct metadata in one source but having the sign of latitude accidentally flipped in another may be classified as two distinct stations. When a candidate station is found in this list, one of two options can occur. Either the metadata is changed to reflect the correct information (whitelisting), or the station is withheld (blacklisting).

To generate this list, three distinct analyses were run to test the validity of a station's metadata and data. First, a running decadal statistical test looked for large shifts in the candidate stations variance. There were times where a station reported in Fahrenheit during one period, and Celsius in another, creating an unnatural shift. If detected, the station is blacklisted and withheld from the algorithm. Second,

a distance *versus* correlation check on all stations found cases where collocated stations (less than 10 km) had high correlations (greater than 0.90), plausibly creating duplicate stations. Conversely, if stations were highly correlated, but with large distances (greater than 1000 km), it was possibly the same station, but with incorrect metadata, creating duplication. A third test included a land sea mask, whereby using a high-resolution gridded dataset, land-based stations whose metadata placed it in the middle of the ocean were flagged. For the last two tests, stations were whitelisted to reflect the correct metadata using information from either NCDC or the WMO.

### 2.4. Metadata comparisons ( $T_{MAX}$ and $T_{MIN}$ )

Once error checking is completed, each candidate station is run through all the target stations and four metadata criteria calculated as the first test to identify matching stations. This process takes into account the likelihood that the same station from two sources may have different precision values for longitude, latitude, and elevation between sources. The station names may also differ, particularly for countries that were once colonial and have subsequently gained independence, or the phonetic spelling of names that may differ by source.

Using the latitudes and longitudes, the geographical distance between the two stations is computed. The distance is then fitted to an exponential decay function (which decays to nearly zero at 100 km distance), and a metric between the two stations is determined, where 0 corresponds to no match and 1 represents a perfect match. Next, the same approach is performed using the height difference between two stations (here the exponential decays to nearly zero at 500 m height difference). Third, a comparison of when the data record began is made. Although not always the case, there is a higher chance the candidate station matches with a target station if they start at or near the same year. Therefore an exponential decay function is applied if the start years fall within 10 years of each other. Finally, the similarity of the station name is considered. This is done using the Jaccard Index (JI) (Jaccard, 1901), which is defined as the intersection divided by the union of two sample sets, A and B:

$$JI = \frac{|A \cap B|}{|A \cup B|}$$

In other words, JI will look for cases in which certain letters exist in both station names, as well as the number of times letters occur in one name, but not in the other. Once the ratio is known, a probability is calculated. One drawback to JI is that it does not take into account the position of the character within the word. Therefore anagrams (i.e. TOKYO and KYOTO) would have a perfect JI of 1.

**Table 9.** Summary of Stage Two sources, in prioritized form, used for the recommended version of the merge program.

Priority	Name	TMAX	TMIN	TAVG	Primary	Secondary
1	ghcnd	Y	Y	N	27300	0
2	Mexico	Y	Y	N	18	2
3	Vietnam	Y	Y	N	15	8
4	Usforts	Y	Y	N	0	38
5	Channel-islands	Y	Y	N	1	1
6	Ecuador	Y	Y	N	1	0
7	Pitcairnisland	Y	Y	N	0	1
8	Giessen	Y	Y	N	13	0
9	Brazil-inmet	Y	Y	N	174	81
10	Brazil	Y	Y	N	4	0
11	Argentina	Y	Y	N	24	8
12	Greenland	Y	Y	N	5	0
13	India	Y	Y	N	15	0
14	gsn-Sweden	Y	Y	Y	0	0
15	Canada-raw	Y	Y	Y	70	431
16	wwr	Y	Y	Y	435	608
17	Colonialera	Y	Y	N	397	178
18	East-Africa	Y	Y	Y	61	39
19	Uganda	Y	Y	Y	23	2
20	Antarctica-aws	Y	Y	N	64	2
21	Antarctica-palmer	Y	Y	Y	1	0
22	Antarctica-southpole	Y	Y	Y	0	0
23	ispd-Swiss	N	N	Y	0	0
24	ispd-ipy	N	N	Y	0	0
25	ispd-Sydney	N	N	Y	0	0
26	Antarctica-scar-reader	N	N	Y	3	21
27	mcdw	N	N	Y	44	458
28	Spain	Y	Y	Y	0	24
29	Uruguay-Inia	Y	Y	Y	1	2
30	Uruguay	Y	Y	N	0	8
31	Swiss-digihom	Y	Y	Y	0	0
32	ispd-tunisia-Morocco	Y	Y	Y	2	3
33	sacad_non-blended	Y	Y	Y	0	0
34	Japan	Y	Y	Y	3	151
35	ukmet-hist	Y	Y	N	19	10
36	knmi	Y	Y	Y	414	1159
37	eklima	Y	Y	Y	215	117
38	russsource	Y	Y	N	1525	2677
39	Germany	N	N	Y	0	53
40	ghcnsorce	N	N	Y	535	2043
41	wmssc	N	N	Y	57	311
42	Central-Asia	Y	Y	Y	80	65
43	Arctic	N	N	Y	3	51
44	histalp	N	N	Y	26	78
45	crutem4	N	N	Y	94	1792
46	hadisd	Y	Y	N	450	1260
47	climat-uk	Y	Y	Y	1	0
48	climat-ncdc	Y	Y	Y	49	0
49	mcdw-unpublished	N	N	Y	0	0

Several sources in Table 1 are not included due to either gross-duplication, inclusion in GHCN-D, or perceived issues over data quality which adversely impacted the merged product in this version 1 release.

Each individual metadata criteria has a value from 0 to 1, which are then combined to form a posterior

metric of possible station match, known as the *metadata metric*.

$$\text{metadata metric} = \frac{(9 \times \text{dist}) + (1 \times \text{height}) + (2.5 \times \text{year}_{T_{\text{MAX}}}) + (2.5 \times \text{year}_{T_{\text{MIN}}}) + (5 \times \text{JI})}{20}$$

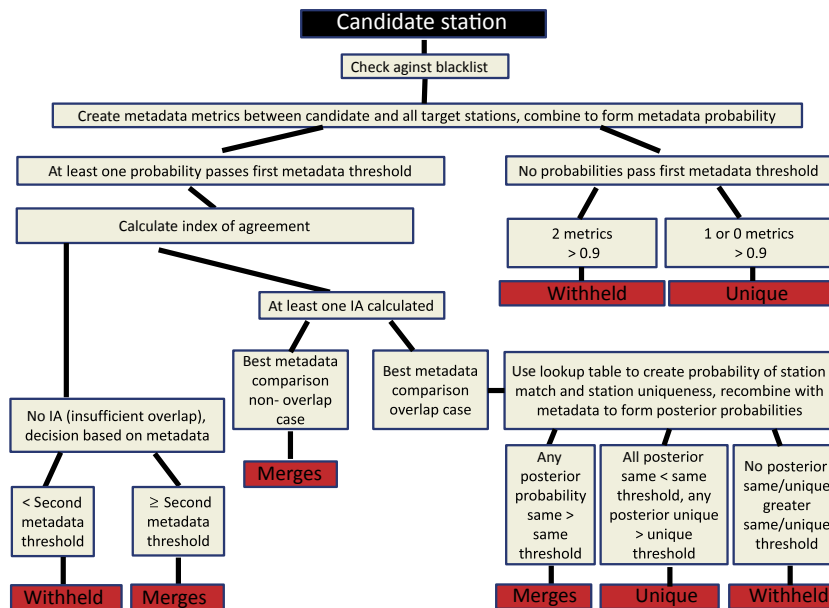


Figure 5. Work flow chart of the merge algorithm.

Weights are given to each criterion based on the reliability of each. Since the latitude and longitude are most likely to not change unless there has been a station relocation, it is given the highest weight. The height of the station is more often misleading, inaccurate, or missing entirely, so it is given the lowest weight. If the *metadata metric* surpasses a threshold of 0.50, and the ID's do not match, then an evaluation based on data comparisons is then made. If the ID's exhibit an exact match, they are then chosen to merge. The threshold to move on to data comparisons is set relatively low to account for possible errors in the metadata. If any of the criteria are missing, the equation is re-adjusted, with the exception of missing latitude and longitude, where the candidate station is withheld.

If none of the comparisons between the candidate station and all the target stations pass the metadata threshold, a review of each metadata criteria is performed. If two of the values are greater than 0.90, then there is the possibility that incorrect metadata within the candidate station has corrupted the overall *metadata metric*. When this occurs, the candidate station is withheld. If this is not the case, it is determined that the candidate station is unique and it is added to the target dataset without any further tests being performed.

## 2.5. Data comparisons ( $T_{MAX}$ and $T_{MIN}$ )

For any of the stations that pass the metadata threshold, a data comparison is made between that target station and candidate station. To have a reliable data comparison, there is a minimum overlap threshold between the two stations of 60 months. If this threshold is met, the data comparison is performed using the index of agreement (IA) (Willmott, 1981).

IA is a 'goodness-of-fit' measure and is defined as the ratio between the mean square error and the potential error. It was designed to overcome issues of correlation measures such as the coefficient of determination. These methods are insensitive to differences in both mean and variance between the target and candidate station, and the presence of outliers would lead to higher values due to the squaring of terms. A modified version of IA (Willmott *et al.*, 1985; Legates & McCabe, 1999) is used where the squared term is removed, and is the equation used during the data comparison stage of the merge program:

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

Where  $T_i$  and  $C_i$  are corresponding monthly values for the target and candidate stations (respectively) and  $\bar{T}$  is the mean of the target station. Note that the mean of the candidate station is not used. Between a candidate and target station, IA is calculated first to the overlapping  $T_{MAX}$  and then the overlapping  $T_{MIN}$ . Resulting values range between 0 and 1. Although these are considered a 'goodness-of-fit' comparison, IA does not take into account the number of months ( $n$ ) of overlap. Although the minimum requirement is 5 years, there could be 50 or more years of overlap. This may lead to a bias, with higher IA occurring for longer periods of overlap.

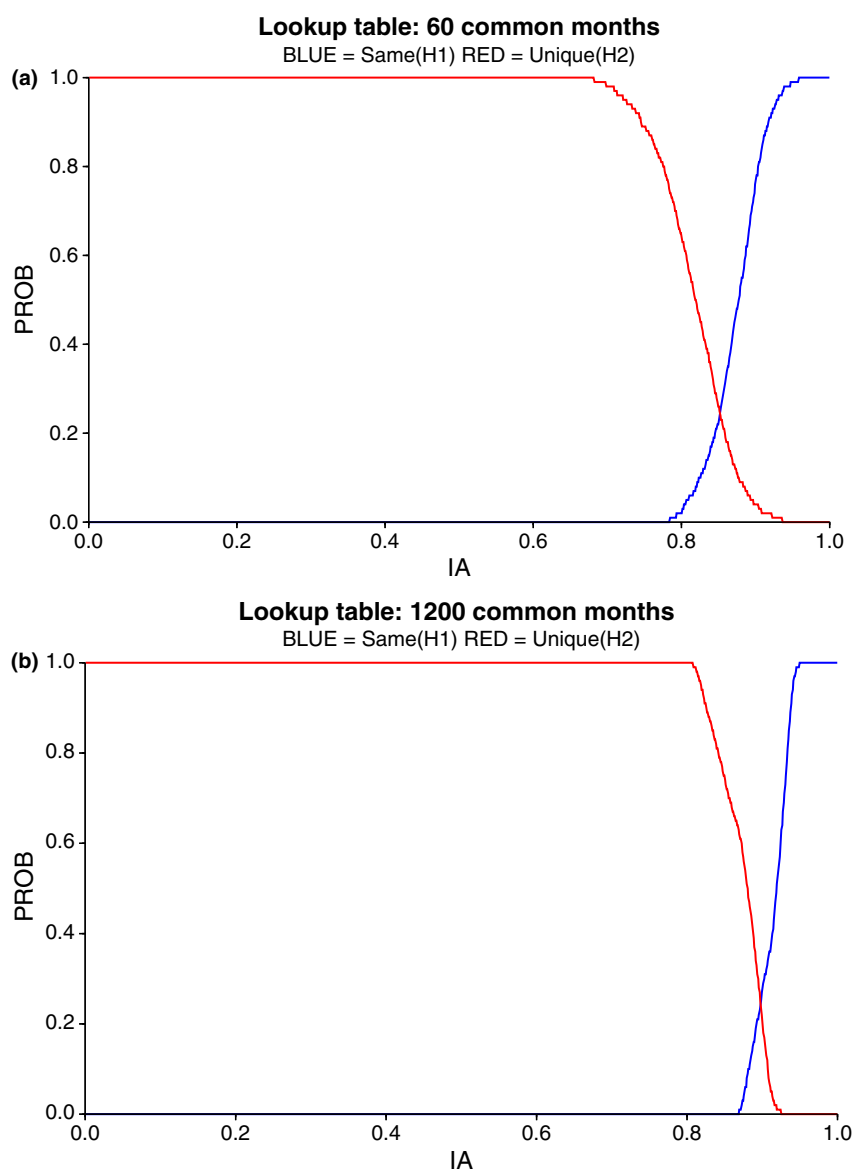
To account for this, a lookup table was generated to provide a probability of station match ( $H1$ ), as well as station uniqueness ( $H2$ ). Shifts in mean and variance were simulated between station records by drawing sequences of random numbers from a normal distribution with specified mean and variance, and then calculating IA. This was applied 1000 times using periods



of record of various lengths. To create this table for *H1*, shifts in overlapping data were applied using a station with a long period of record. For our purposes, the station from De Bilt, The Netherlands was used, since continuous data are available since 1706 for  $T_{AVG}$  (1901 for  $T_{MAX}$  and  $T_{MIN}$ ). For *H2*, statistics were derived from stations within 50 km of a number of target stations within densely sampled regions of GHCN-D, and these were used to derive reasonable expectations of how neighbouring stations may be expected to differ on a month-to-month basis. Using these results, a cumulative distribution function is calculated for each contingency (same station and unique station) and stratified overlap periods of various lengths. The greater the overlap period, the closer to 1.0 IA needs to be to be considered a station match (Figure 6).

This data comparison is applied to all the target stations that could match with the candidate station according to the metadata test. There are three distinct possibilities when attempting to perform a data comparison: (1) No data comparisons were possible because of insufficient overlap, (2) Some comparisons were possible, but some did not include those targets with the highest *metadata metrics* because of insufficient overlap, and (3) Data overlap comparisons were possible for at least the highest *metadata metric* cases.

If there was insufficient overlap, the final decision is based solely upon the *metadata metric*. Because of this the metadata comparisons need to be closer to perfection, so the *metadata metric* threshold is increased from 0.50 to 0.90. If the highest metadata comparison with a target station received a *metadata*



**Figure 6.** Lookup table used to determine probability of station match (blue) and probability of station uniqueness (red) between a target and candidate station based upon their index of agreement with 5 years of overlap (a) and 100 years of overlap (b).

*metric* larger than this new threshold, then the candidate station merges with that station. Otherwise it is withheld.

There are also cases where data comparisons were made, but the *metadata metric* of a non-overlapping station was higher than for any of the stations that had a data overlap. This can occur in areas with a dense network of stations. If this is found to be true, then that candidate station is merged with the non-overlapping target station.

Otherwise there are five resulting metrics, one *metadata metric*, and four data metrics (tests for station match and uniqueness, for both  $T_{MAX}$  and  $T_{MIN}$ ). These prior metrics are then recombined to form two new posterior metrics, one of station match, and one of station uniqueness. The unique equation was structured so it favours a lower *metadata metric* (near 0.50), and because it is not weighted, this value can range between 0 and 2.50.

$$\text{posterior metric same}_{T_{MAX}/T_{MIN}} = \frac{\text{metadata metric} \times H1_{t_{max}} \times H1_{t_{min}}}{3}$$

$$\text{posterior metric unique}_{T_{MAX}/T_{MIN}} = (1 - \text{metadata metric}) + H2_{t_{max}} + H2_{t_{min}}$$

Once these posterior metrics are made for all possible comparisons between a candidate station and its target stations, thresholds are set for station match and uniqueness (0.50 and 1.30 respectively) to determine the final fate of the candidate station. If any of the values returned for *posterior metric same* exceed the same threshold of 0.50, then the candidate station is merged with the target station with the highest *pos-*

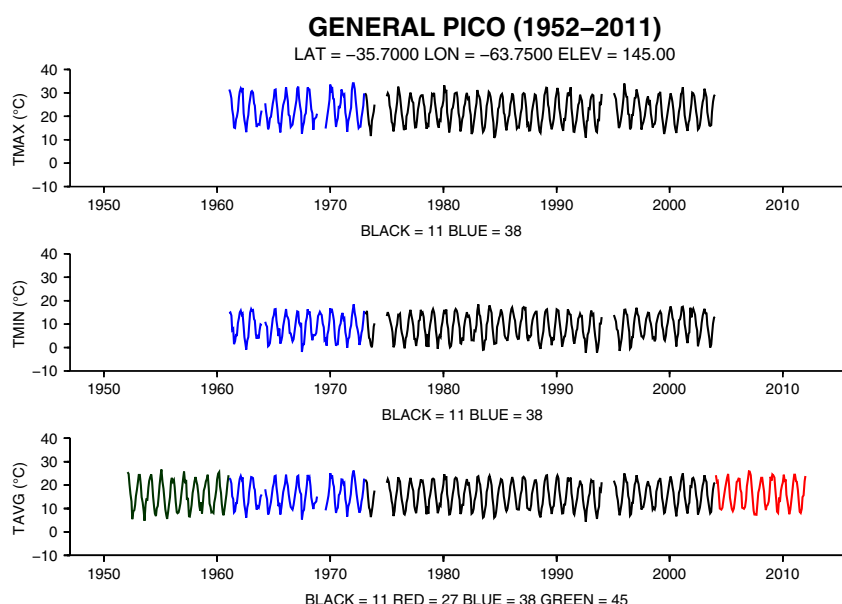
*terior metric same*. If none of the stations exceed that threshold, but one of the *posterior metric unique* values exceeds the unique threshold, then the candidate station becomes unique and is added to the target dataset. If no metrics pass either threshold, then the station is withheld.

If merging of data is performed, only data from the candidate station not already in the target station record are added to create the new merged record. If data occur for both the candidate station and the target station, preference is always given to the target, since it contains data that were higher in the prioritized list. The merging appends data from the candidate to the target to create a single, extended, record. No candidate data are inserted into the middle of the target series unless they could fill a string of at least five consecutive years of missing data. This is done to better ensure sufficient record length for detecting inhomogeneities that may result from com-

binning data from different sources. Data segments can be added to a single station from multiple sources through the iterations across sources (Figure 7).

## 2.6. $T_{AVG}$ comparisons

Using the above metadata and data criteria, all of the sources are subjected to the merge algorithm looking



**Figure 7.** Station series of  $T_{MAX}$ ,  $T_{MIN}$ , and  $T_{AVG}$  data for General Pico, Argentina. Four data sources consist of the final merged product including data from Argentina's National Institute of Agriculture (source number 11 in the recommended merge, black), mcdw (source number 27, red), russsource (source number 38, blue), and CRUTEM4 (source number 45, green).

**Table 10.** List of flags associated with data in the withheld folder that did not get merged.

Flag	Description
101	Missing metadata
102	Poor metadata (2 thresholds >0.9, but <i>metadata_metric</i> still <0.5)
103	No data comparison made, best station does not reach second metadata threshold
104	Data comparison made, no station had <i>posterior_metric_same</i> or <i>posterior_metric_unique</i> above the same/unique threshold
105	Final target station has less than 12 months of data
106	<i>metadata_metric</i> ≥0.9, however data comparisons were so poor that station would have become unique
107	Candidate station blacklisted during error checking

only at stations that have  $T_{MAX}$  and  $T_{MIN}$ . Afterwards,  $T_{AVG}$  is generated from the merged dataset (by averaging  $T_{MAX}$  and  $T_{MIN}$ ). Since there are numerous ways to calculate average temperature (Trewin, 2010), it is important to keep the calculation consistent. Following this, the sources are checked a second time, only looking at sources which have  $T_{AVG}$ . Data comparisons of  $T_{AVG}$  are similar to those of  $T_{MAX}$  and  $T_{MIN}$ , with the exception that the *metadata metric* and final posterior metrics are modified. This is because there is only one temperature variable ( $T_{AVG}$ ) instead of two ( $T_{MAX}$  and  $T_{MIN}$ ):

After  $T_{AVG}$  comparisons are applied, target stations with less than 12 months of data are withheld and the

observation in the USA and were chosen to maximize the time of observation bias in the corresponding monthly means generated from these daily maximums and minimums (Karl *et al.*, 1986; Menne *et al.*, 2009). Nearly all of the stations with these generated daily maximums and minimums are also represented in GHCN-D, but with observational days ending at local midnight (when the time of observation bias is by definition zero). By comparing monthly mean maximum and minimum temperatures with the extremes of time of observation biases to values generated from the local midnight standard, it is possible to quantify the skill of the merging procedure in a relatively high density station network using a target and candidate dataset that are similar, but not exact matches. Because GHCN-D is the first priority in the source hierarchy, the dataset generated from hourly values is considered the candidate source. An example of the validation experiment is provided in Table 11. More generally, of the 1952 stations already represented in GHCN-D, 1668 (85.45%) were correctly identified as merged candidates, 5 (0.26%) became unique stations, and 279 (14.29%) were withheld. Out of the 1668 that were chosen as merge candidates, 1556 (93.29%) were merged with the correct GHCN-D stations.

### 3. Results of stage 3 dataset

The following section highlights the results of the merging algorithm. To address the sensitivity of the merge process to source priority and various thresholds, several variants of the merge were produced

$$\text{metadata metric} = \frac{(9 \times \text{dist}) + (1 \times \text{height}) + (2.5 \times \text{year}_{T_{AVG}}) + (5 \times \text{JI})}{17.5}$$

$$\text{posterior metric same}_{T_{AVG}} = \frac{\text{metadata metric} \times H1_{T_{AVG}}}{2}$$

$$\text{posterior metric unique}_{T_{AVG}} = (1 - \text{metadata metric}) + H2_{T_{AVG}}$$

result is the final, merged stage 3 dataset. Stations that have been withheld are placed in a separate directory, along with specific flags indicating the reason they were not included in the final merged product (Table 10).

## 2.7. Validation

All the decisions made in the previous sections were tested against an independent dataset. The dataset was generated from hourly data for US stations available in the Integrated Surface Dataset ('ISD-Lite'; Smith *et al.*, 2011). Daily maximum and minimum temperatures were generated by taking the highest and lowest of 24-hourly values for observational days ending at 12 and 22 UTC. These times correspond to local morning and late afternoon/evening times of

(Section 4.2) along with the recommended merge endorsed by ISTI (highlighted in section 4.1).

### 3.1. Recommended merge

Using the source hierarchy (Table 9) and thresholds (Table 12) recommended by ISTI, over 32 000 unique stations were identified, over four times as many stations as GHCN-M version 3 (Figure 8). Although station coverage varies spatially and temporally, there are adequate stations with decadal and century periods of record at local, regional, and global scales. In addition, station coverage has increased where data were lacking in previous versions of GHCN-M (Figure 1), including parts of South America, Africa, and Southeast Asia.

Some statistics describing the outcome of all candidate stations within the recommended merge are

**Table 11.** Example of validation using subset of GHCN-D as a candidate source.

TARGET STATIONS	META	IA_T <sub>MAX</sub>	H1_T <sub>MAX</sub>	H2_T <sub>MAX</sub>	IA_T <sub>MIN</sub>	H1_T <sub>MIN</sub>	H2_T <sub>MIN</sub>	PST_ SAME	PST_ UNIQ
CHESTNUT_HILL	0.6592	0.9287	0.9500	0.0000	0.8662	0.2000	0.2200	0.6031	0.5608
COHASSET	0.6214	0.9554	1.0000	0.0000	0.8735	0.4700	0.0900	0.6971	0.4686
HINGHAM	0.5362	0.9523	1.0000	0.0000	0.8814	0.3800	0.1300	0.6387	0.5938
JAMAICA_PLAIN	0.5614	0.9142	0.8200	0.0100	0.8730	0.2500	0.2000	0.5438	0.6486
LAKE_COCHITUATE	0.5102	0.9249	0.9600	0.0100	0.7738	0.0000	0.8200	0.4901	1.3198
LEXINGTON	0.7082	0.9410	0.9800	0.0000	0.8132	0.0600	0.5300	0.5827	0.8218
MARBLEHEAD	0.5377	0.9591	1.0000	0.0000	0.9230	0.9200	0.0100	0.8192	0.4723
MIDDLETON	0.5592	0.9322	0.9500	0.0000	0.8609	0.0200	0.3900	0.5097	0.8308
PEABODY	0.5379	0.9496	0.9900	0.0000	0.8729	0.3100	0.1500	0.6126	0.6121
READING	0.5124	0.9468	0.9900	0.0000	0.7887	0.0000	0.9500	0.5008	1.4376
SALEM_CG_AIR_STN	0.6621	0.9566	1.0000	0.0000	0.9364	0.9800	0.0000	0.8807	0.3379
SWAMPSCOTT	0.6078	0.9337	0.9600	0.0000	0.9262	0.9500	0.0100	0.8393	0.4022
WALPOLE	0.5672	0.9530	1.0000	0.0000	0.7777	0.0000	0.8600	0.5224	1.2928
WALPOLE_1_SSE	0.5912	0.9219	0.9300	0.0100	0.7915	0.0100	0.6900	0.5104	1.1088
WESTON	0.5346	0.8984	0.7200	0.0300	0.7851	0.0000	0.8100	0.4182	1.3054
NASHUA	0.5108	0.9365	0.9800	0.0000	0.7087	0.0000	1.0000	0.4969	1.4892
BEDFORD_HANSCOM_FLD	0.6388	0.9454	0.9900	0.0000	0.8133	0.0000	0.6600	0.5429	1.0212
<b>BOSTON_LOGAN_</b>	<b>0.9415</b>	<b>0.9701</b>	<b>1.0000</b>	<b>0.0000</b>	<b>0.9771</b>	<b>1.0000</b>	<b>0.0000</b>	<b>0.9805</b>	<b>0.0585</b>
<b>INTL_AP</b>									
BLUE_HILL	0.5304	0.9520	1.0000	0.0000	0.8721	0.1100	0.2600	0.5468	0.7296
SOUTH_WEYMOUTH_NAS	0.5654	0.9497	1.0000	0.0000	0.9062	0.7000	0.0300	0.7551	0.4646
NORWOOD_MEM_AP	0.5385	0.9283	0.9500	0.0000	0.7627	0.0000	0.8600	0.4962	1.3215
BEVERLY_MUNI_AP	0.5371	0.9671	1.0000	0.0000	0.8237	0.1000	0.4200	0.5457	0.8829
TAUNTON_MUNI_AP	0.5756	0.9303	0.9700	0.0000	0.9123	0.8600	0.0200	0.8019	0.4444

Candidate station (ISD ID # = 72509014739, name = BOSTON/LOGAN\_INTL) makes data comparisons with 23 target stations within GHCN-D. Matches with correct station (GHCN ID# = USW00014739, name = BOSTON\_LOGAN\_INTL\_AP). The candidate station matches with the correct station, which is bolded in this table

provided in Table 13. About 80% of stations merged with a target station, and nearly 6% of stations became unique in the recommended product. Stations were withheld primarily because they did not pass the *posterior metric same* or *posterior metric unique* thresholds after both metadata and data comparisons. These stations are not part of the recommended merge, however, are still available to the public for further analysis. Table 9 also provides information about the recommended merge by source. Overall, 85% of the target (primary) stations originate from GHCN-D, and many of the sources included stations that appended data to these target stations (secondary).

Since 1850, there are consistently more stations in the recommended merge than GHCN-M (Figure 9a). In GHCN-M version 3, there was a significant drop in stations in 1990 that is ameliorated by many of the new sources. An upward spike during the mid-1970s includes new data from the ISD, which included the introduction of data transfer through the global telecommunications system (GTS) (Smith *et al.*, 2011). The spike is the result of an increase in stations outside the United States with this addition (Figure 9(b)). A histogram of station count by record length compared to GHCN-M version 3 is shown in Figure 10. There are not only many more stations in the recommended merge but also more long series.

A comparison of grid boxes is noted in Figures 11(a)–(c). This highlights the percentage of possible coverage for the recommended merge and GHCN-M version 3. Coverage is defined by one or more stations within each 5° × 5° grid box, which contains land. Because this is a land surface dataset, grid boxes that consist solely of ocean are not considered. There is a net increase in the global station coverage for all time periods and especially so for 1990–2010 (Figure 11(a)). There is greater improvement in coverage in the Northern Hemisphere, with about a 3–5% increase in coverage from 1850–1950, increasing to 10% during the 1960s, 1970s, and during the 1980s, and then as high as 20% during the 1990s and 2000s (Figure 11(b)). Southern Hemisphere coverage varies between a 3% and 6% increase for the most part, with larger increases over the past 20 years (Figure 11(c)).

Overall, there is a dramatic increase in grid box coverage due to the increased number of stations in the databank (Figure 12). During the 19th century, stations in the western part of the United States that were not in GHCN-M version 3 have filled the gap in US coverage. Coverage also increased in parts of South America, India, and Australia. By the end of the 20th century, stations added from higher latitudes and Antarctica provides substantially more coverage than GHCN-M version 3.

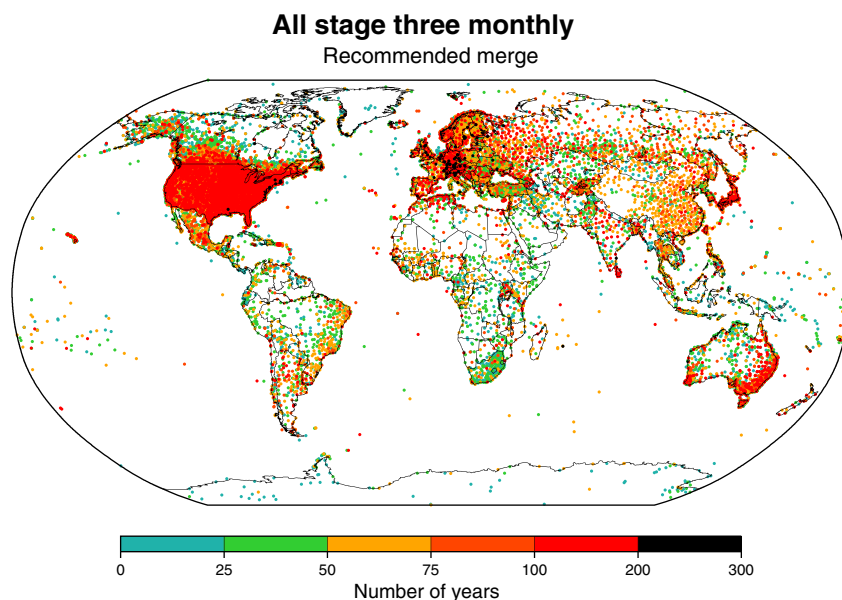
There are some grid boxes that only contain data in GHCN-M and not from the recommended merge (red



**Table 12.** List of user defined thresholds in the merge program.

Name	Description	Default	Increase?	Decrease?
Metadata Threshold	The first metadata threshold that takes into account the distance, height, and Jaccard metrics	0.50	Pull more through as unique stations	More data comparison
Metadata Threshold2	The second metadata threshold used if there is no overlap period between the target and candidate station (higher than the first metadata threshold)	0.90	Withhold more stations	More merging of stations
Posterior Threshold Same-TXN	Threshold where $T_{MAX}/T_{MIN}$ candidate station has to exceed to merge with the target station	0.50	Make stations either unique or withheld	More merging of stations
Posterior Threshold Unique-TXN	Threshold where $T_{MAX}/T_{MIN}$ candidate station has to exceed to be considered a unique station	1.30	Withhold more stations	More unique stations
Posterior Threshold Same-TVG	Threshold where $T_{AVG}$ candidate station has to exceed to merge with the target station	0.50	Make stations either unique or withheld	More merging of stations
Posterior Threshold Unique-TVG	Threshold where $T_{AVG}$ candidate station has to exceed to be considered a unique station	0.90	Withhold more stations	More unique stations
Overlap Threshold	Overlap period that must exist between the target and candidate station to calculate a data comparison via the index of agreement	60	Pull more through as unique stations	More data comparison
Gap Threshold	Gap period that must exist when merging a candidate station with the target station	60	Lower number of merges	Increase number of merges

The first metadata threshold must be lesser than the second.



**Figure 8.** Location of all stations in the recommended stage 3 component of the databank. The colour corresponds to the number of years of data available for each station. Stations with longer periods of record mask stations with shorter periods of record when they are in approximate identical locations.

dots in Figure 12). There are two reasons for this occurrence. First, in a small number of cases latitudes and longitudes in GHCN-M version 3 differ slightly from the same station in other sources (e.g. GHCN-D). Where stations are close to grid box edges a small change in

location can result in the station moving into an adjacent grid box. For example, the grid box off the very southwest tip of Australia (Figure 12(b) and (c)) contains one station only present in GHCN-M version 3. The small differences in longitudes between GHCN-M and

**Table 13.** Statistics describing percentage of candidate stations merged, unique, and withheld in the recommended program. Inner statistics also describe details about each major category.

% Merged 80.7	
No data comparisons made, rely on <i>Metadata Metric</i>	22.5
Data comparisons made, best <i>Metadata Metric</i> was non-overlap case	1.6
Data comparisons made, best station chosen through both <i>Metadata Metric</i> and <i>Posterior Metric Same</i>	61.1
Data comparisons made, ID's were a perfect match	14.8
% Unique 5.7	
Unique after metadata comparisons	81.8
Unique after metadata and data comparisons	18.2
% Withheld 13.6	
Withheld after metadata comparisons	13.1
Withheld after metadata and data comparisons	86.9

GHCN-D (Table 14) result in the station moving to a grid box completely over land, and the station is merged with a higher priority source (GHCN-D). Second, some stations that are included in GHCN-M version 3 have been placed in the withheld bin during the databank merge process, due to either bad metadata or data comparisons being unable to definitively identify whether the station should be merged or included as a unique stations (as described in previous sections).

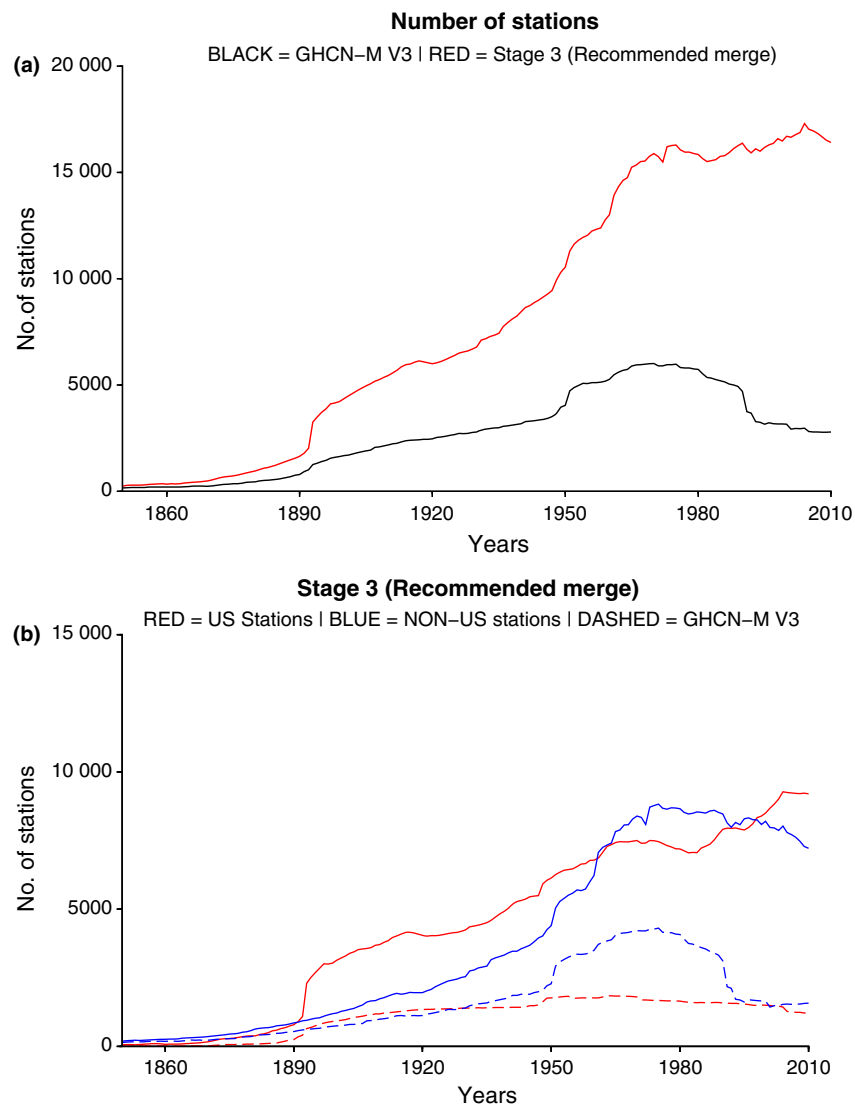
To assess the affect the large increase in stations has on global temperatures, the Climate Anomaly Method (Ropelewski *et al.*, 1984; Jones & Moberg, 2003) is used to compute global temperature anomalies (with respect to a base period from 1961 to 1990) from the recommended merge. Since the stage 3 dataset is not homogenized, it is compared to the unadjusted version of GHCN-M version 3 and stratified by annual and seasonal periods (Figure 13). Overall, anomalies are lower than unadjusted GHCN-M version 3 up to about 1950 and then equal or larger from 1990 onwards. Therefore, the merge has a larger increasing temperature trend than the global estimate based upon the unadjusted GHCN-M version 3 product (Table 15). To understand the cause of this change, Figure 14(a) plots the anomalies using unadjusted GHCN-M version 3 (black), the recommended merge (red), and the merge product again, but only containing grid boxes that existed in unadjusted GHCN-M version 3 (blue). The same is done in Figure 14(b), but for grid boxes not in unadjusted GHCN-M version 3. Using these two figures, it can be inferred that the change in the temperature trend is primarily due to a larger sampling of stations in existing grid boxes rather than the addition of entirely new grid boxes. Whether such a distinction remains after homogenization is an important and open question. This will be addressed when creating future products, such as GHCN-M version 4, which will be based upon the recommended merge variant described here, but using NCDC's QC and homogenization algorithms (Lawrimore *et al.*, 2011 and references therein). Other homogenization efforts will also add further insight into this observed trend in the unadjusted data.

### 3.2. Merge variants

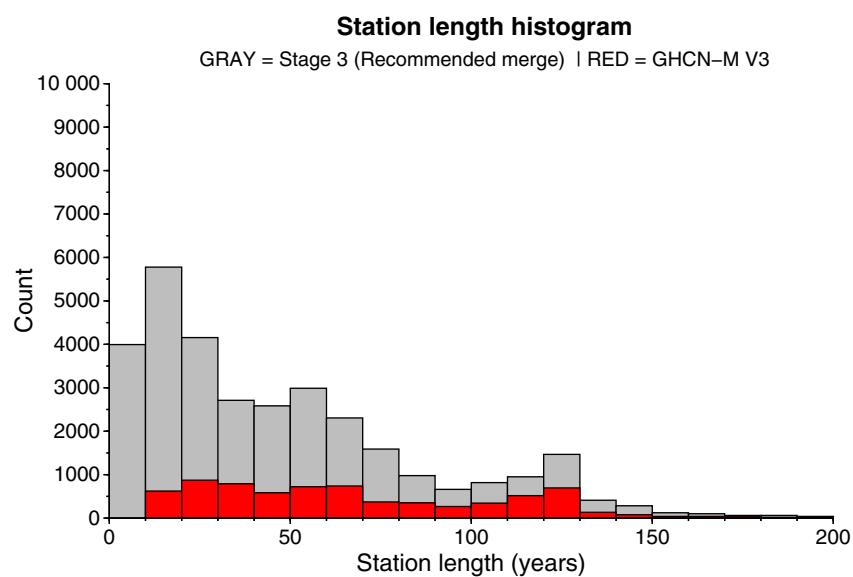
Table 12 describes eight thresholds used in performing the stage 3 merge. The table includes the definition of each, the default used for the recommended merge, and the effect of increasing or decreasing that threshold. Changing these thresholds can significantly alter the overall result of the program. Changes will also occur when the source priority hierarchy is altered. To characterize the uncertainty associated with the merge parameters, seven different variants of the stage 3 product were developed alongside the recommended merge. Members of the DWG have provided suggestions for multi-member variants and a description of each is highlighted in Table 16.

Figures 15, 16, and 17 highlight the number of stations over time, the percentage of land area in  $5^\circ \times 5^\circ$  grid boxes, and the temperature anomaly of all seven variants (red), along with the recommended merge (blue) and unadjusted GHCN-M version 3 (black). There is a large spread in the number of stations between the variants, especially after 1950. This highlights the sensitivity of this algorithm, being highly dependent on the thresholds and source hierarchy. However, it should be noted that all variants have more stations than GHCN-M version 3 from 1890 onwards. The same cannot be said for the differences in percentage of land area sample by  $5^\circ$  grid box counts. Although most variants have more spatial sampling than GHCN-M version 3, one has less sampling (variant 3). Regardless, the global temperature anomaly (Figure 17) has little spread from the recommended merge. The only outlier is variant 3, where the global anomalies were lower during the 1800s because of poor spatial sampling. Variant 3 is composed solely of sources that have maximum and minimum temperature, excluding sources with average temperature. This greatly reduces the amount of data available to form the anomaly in the early time period. If GHCN-M version 3 or the other variants are similarly sampled they would exhibit a similar behaviour.

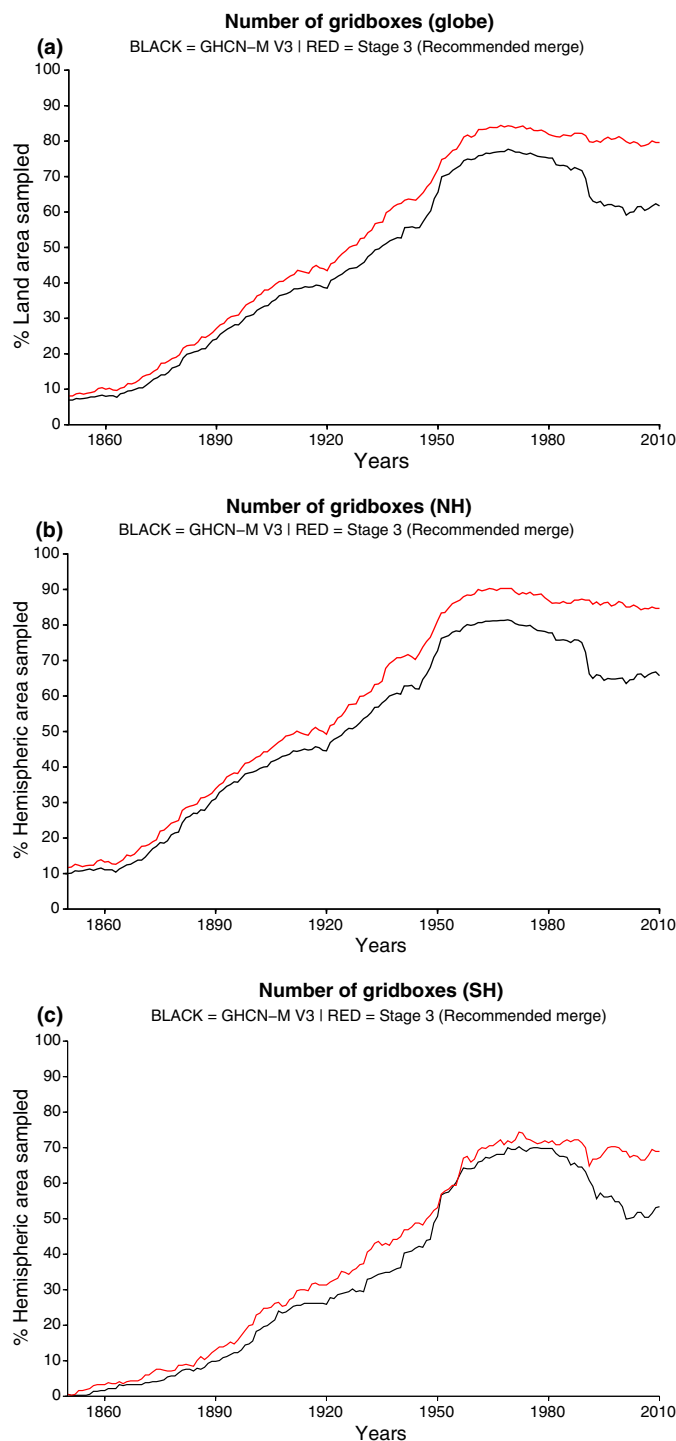
This uncertainty reflects the importance of data rescue. Although a major effort has been undertaken through this initiative, more can be done to include



**Figure 9.** Station count of recommended merge by year from 1850 to 2010. All stations (a) in red compared to GHCN-M version 3 in black, and a comparison of US in red *versus* Non-US stations (b) in blue are made *versus* GHCN-M version 3, which are dashed lines.



**Figure 10.** Histogram of station count by record length for the recommended version of the merge, compared to GHCN-M version 3.



**Figure 11.** Percentage of global (a), Northern hemispheric (b), and Southern hemispheric (c) coverage for the recommended version of the merge, compared to GHCN-M version 3.

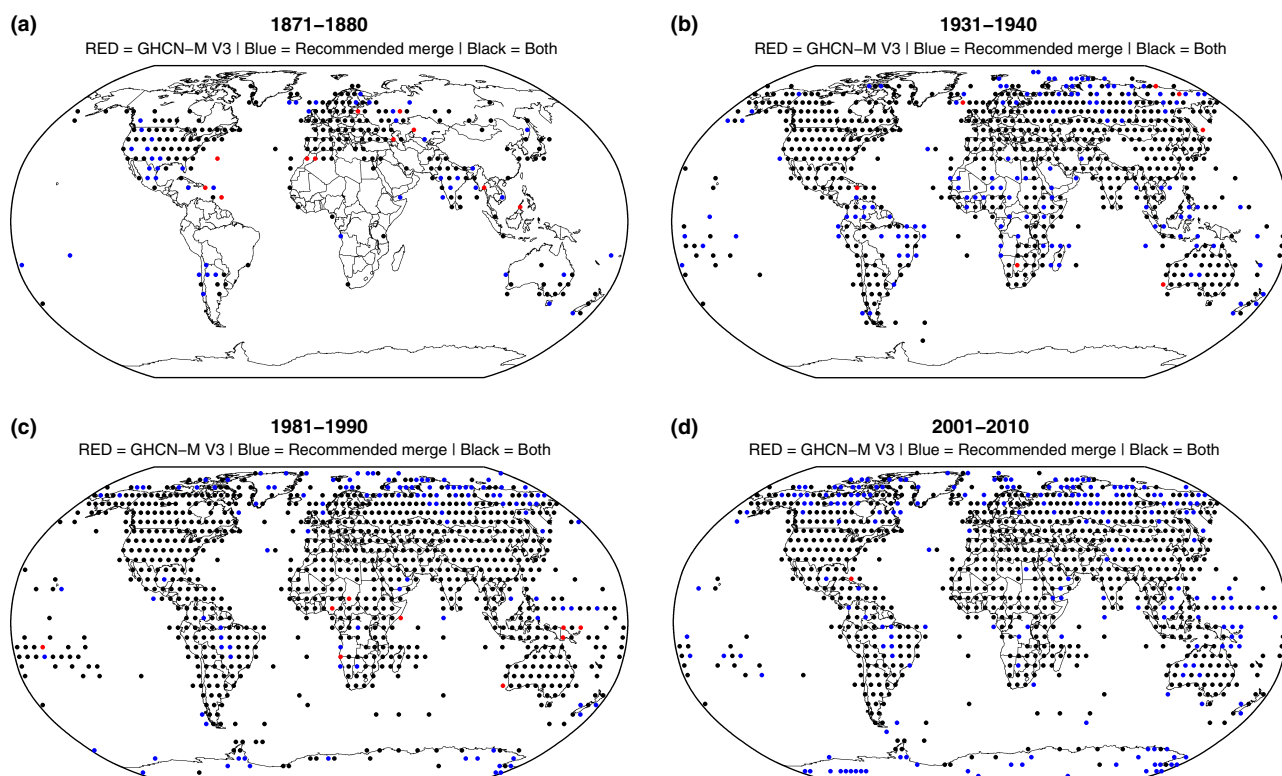
areas that are lacking on both spatial and temporal scales, or lacking maximum and minimum temperature data.

#### 4. Data access and version control

Data are provided from a primary ftp site hosted by the Global Observing Systems Information Center

(GOSIC; <http://gosic.org>) and World Data Center A at NOAA/NCDC. In addition, World Data Center B at Obninsk, Russia established an ftp site that is routinely updated to mirror the data on the primary site. <ftp://ftp.ncdc.noaa.gov/pub/data/globaldatabank/>, <ftp://ftp.meteo.ru/pub/data/globaldatabank/>. Other mirrors may be established in the future. Data are provided in ASCII to facilitate quick access and use by anyone in





**Figure 12.** Spatial coverage of both databank and GHCN-M version 3 data during the periods (a) 1871–1880, (b) 1931–1940, (c) 1981–1990, (d) 2001–2010. Blue indicates a  $5^\circ \times 5^\circ$  box containing at least one station that exists only in the merge product, red for only GHCN-M version 3, and black for both.

**Table 14.** Metadata for station CAPE NATURALISTE for GHCN-M version 3 and GHCN-D.

	ID	NAME	LATITUDE	LONGITUDE	ELEVATION (m)
GHCN-D	ASN00009519	CAPE NATURALISTE	−33.5372	115.0189	109
GHCN-M	50194600000	CAPE NATURALI	−33.5300	115.0000	98

Because the GHCN-M longitude resides on the 115 degree line, it is placed in the grid box centred over the ocean (−32.5, 112.5), whereas the GHCN-D station is placed in the grid box centred over land (−32.5, 117.5).

**Table 15.** The 1850–2010 temperature trends, in degrees Celsius, calculated using the Climate Anomaly Method compared to the unadjusted GHCN-M version 3, recommended merge, and merge variants.

	Trend (°C)
GHCN-M v3 (unadjusted)	0.544
Recommended	0.668
Variant one	0.661
Variant two	0.665
Variant three	0.819
Variant four	0.668
Variant five	0.653
Variant six	0.645
Variant seven	0.603

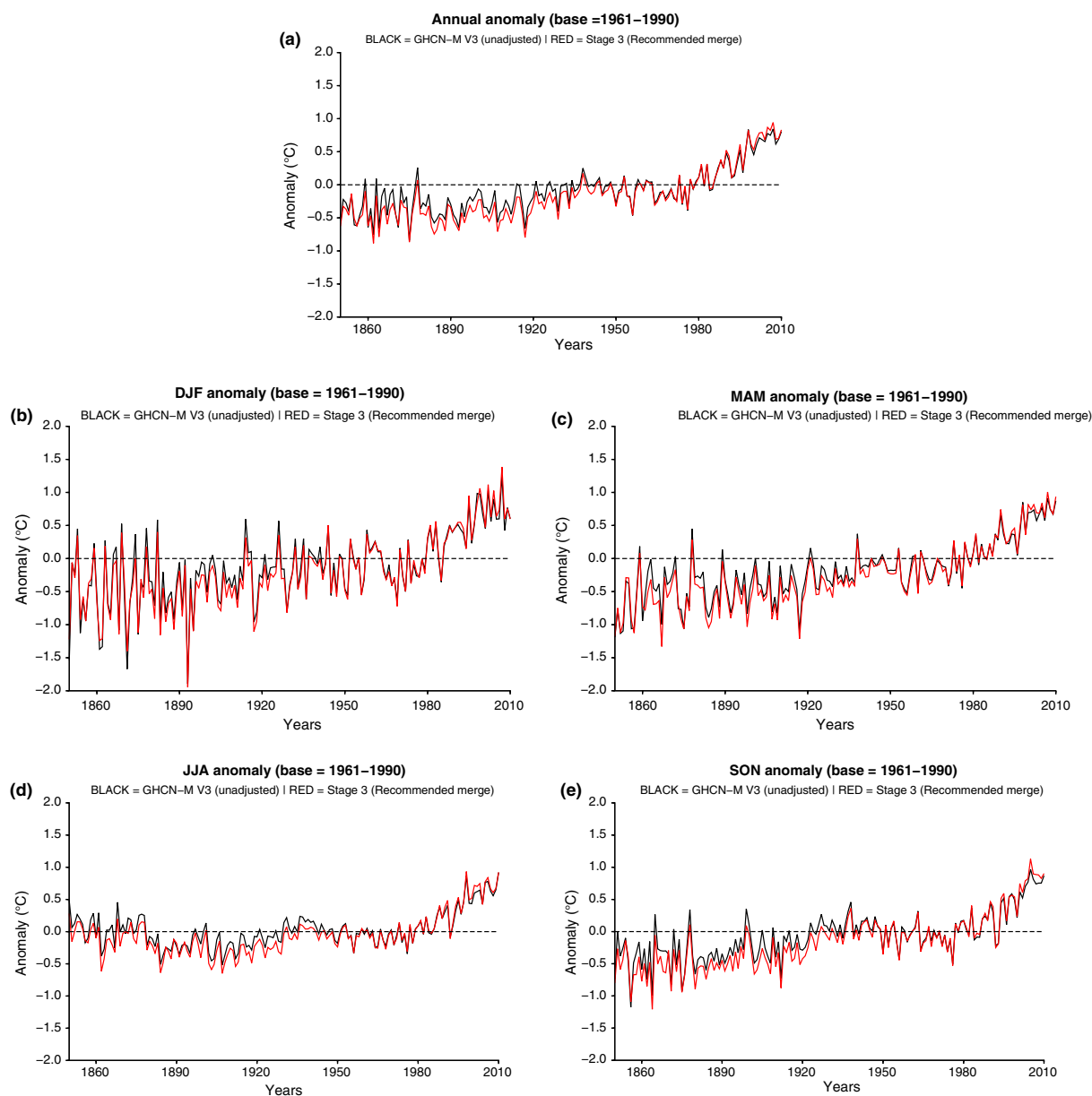
the international community. The stage 3 dataset comes in three different formats. The first is the format similar to stage 2 data, and the second is a

format similar to GHCN-M version 3. Conversion of data files to the NetCDF Climate and Forecast convention are also provided.

The databank consists of subdirectories for daily and monthly data. In some cases, the data provider has agreed to contribute regular data updates. As updates arrive, the previous version is moved to the archive directory and permanently stored in a directory specific to the source. Within the archive directory, each version is maintained in a separate subdirectory designated by the year, month, and day the data were first added to the databank.

It is preferable that the entire source dataset be transferred as updates are made rather than collection of only the most recent observations. By acquiring the full source dataset each time an update is made, the databank can better ensure the most up-to-date data.

A version number is assigned to new sources or updates to sources as they are added to the data-



**Figure 13.** Annual (a) and seasonal (b–e) global temperature anomalies (with respect to the 1961–1990 climatology) using both the recommended version of the merge program and the unadjusted version of GHCN-M version 3.

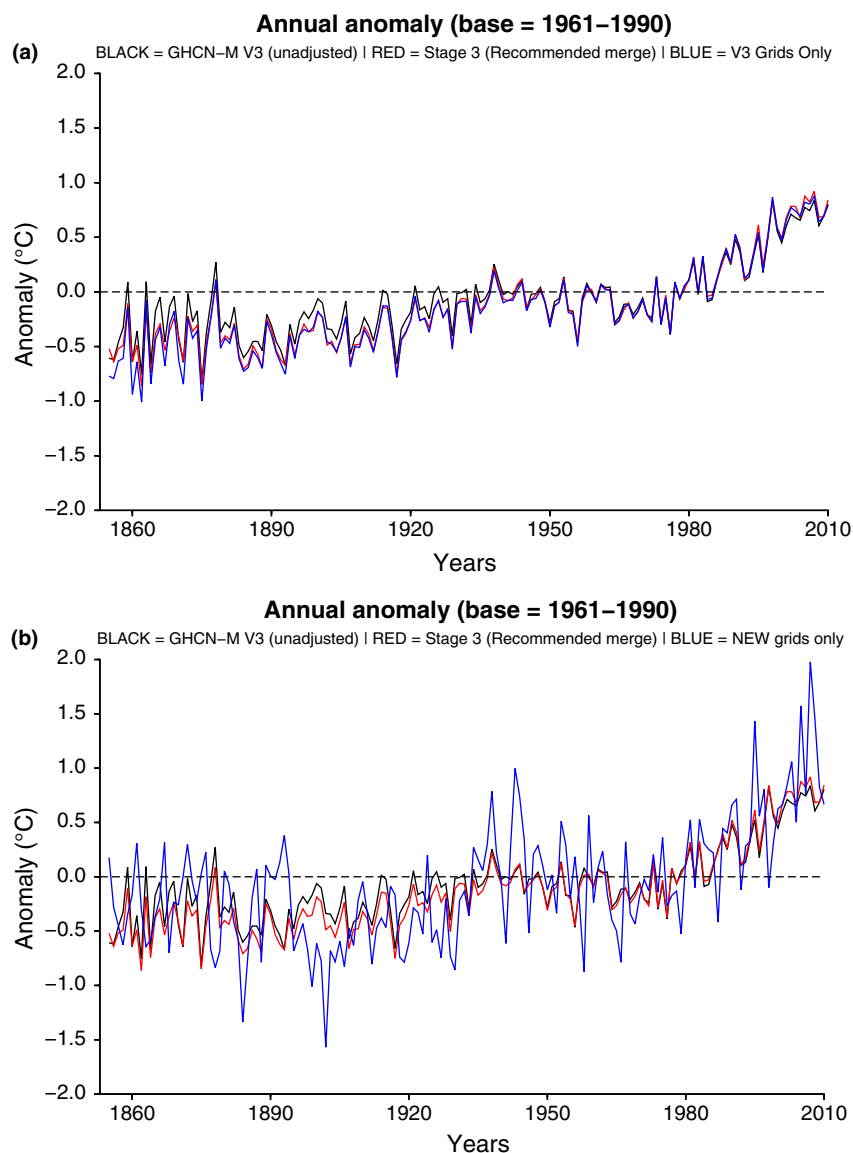
bank. All files from a single source are combined into a single tar file, which is compressed using gzip. The version control protocol follows those applied to GHCN-M version 3 (Lawrimore *et al.*, 2011) and is described in Appendix B. It is stressed that users should always clearly identify the variant of the merge they are utilizing. As all variants will be archived this will facilitate independent replication and transparency of any analyses built off of the databank holdings.

## 5. Concluding remarks and outlook

Construction of a land surface databank is a major undertaking requiring time and international coordination. It has been preceded by many groundbreaking

efforts that established the foundation that makes this possible, and it comes at a time when the need for high quality, traceable, and complete data are clearer than ever.

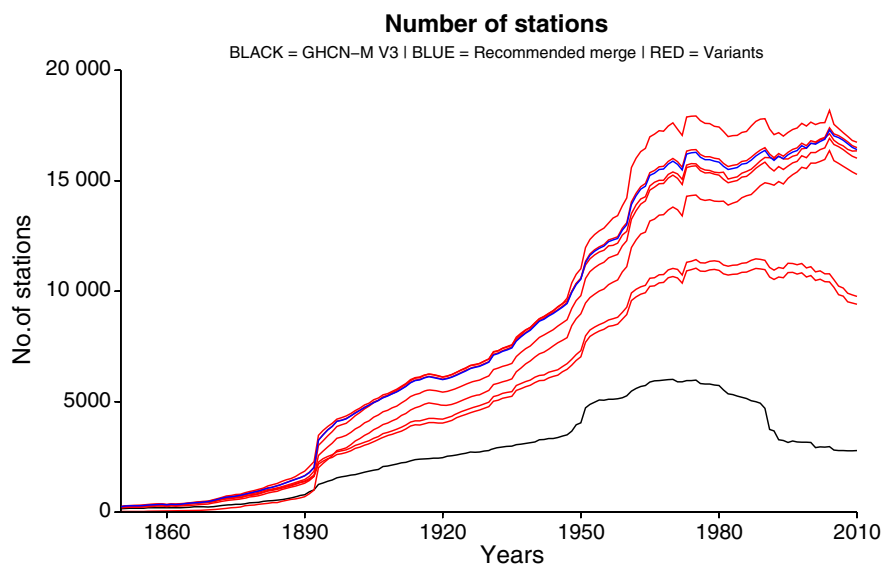
As an integral part of the ISTI, the databank contains many stages, reaching as far back as scans of the original paper copy in those cases where it can. All digitized data have been converted into a common format and then merged together to resolve duplicate stations. This algorithm takes on a quasi-probabilistic approach, and consists of metadata matching and data equivalence criteria. The recommended merge, along with variants are provided to characterize uncertainty. For all stages, data provenance and version control have been applied, and all data and software code used to analyse the data are made



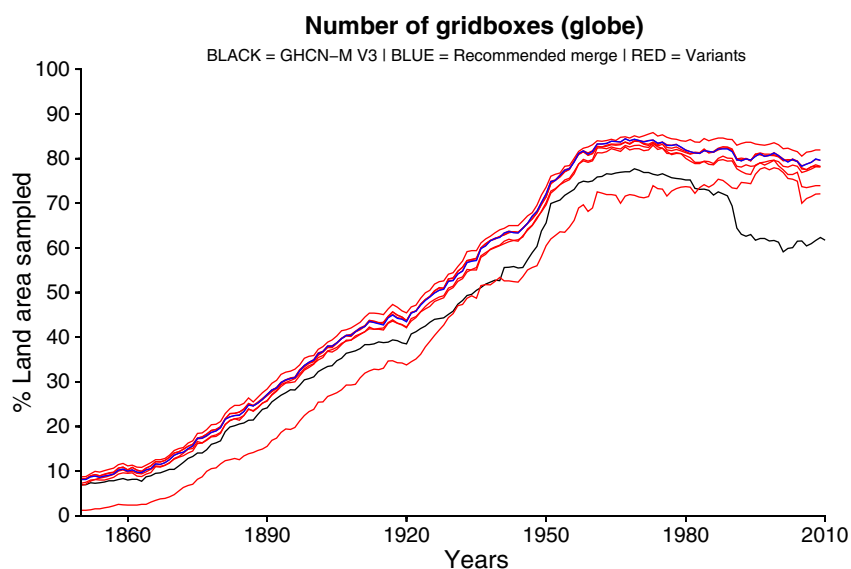
**Figure 14.** Annual global temperature anomalies (with respect to the 1961–1990 climatology) for the recommended version of the merge program (red) and unadjusted GHCN-M version 3 (black). These are contrasted with the anomalies from the mutually inclusive coverage (both databank and GHCN-M version 3) (a), and the anomaly from the new unique boxes in the databank (b).

**Table 16.** Description of variants computed and used to compare with the recommended merge.

Variant	Source deck change?	Threshold change?	Code change?
1	Prioritize sources from NMA's.	No	No
2	NMA's with $T_{MAX}$ and $T_{MIN}$ given highest priority	<i>overlap_threshold</i> changed from 60 months to 24 months	No
3	No $T_{AVG}$ sources used, rest ranked by order of the longest station record present	Thresholds to merge and unique station are lowered to merge more stations	<i>metadata_metric</i> weighted to favour <i>distance_metric</i> over all others
4	No	No	During data comparisons, candidate station only merged or unique
5	All homogenized sources removed	No	No
6	No	All thresholds adjusted to make more candidate stations unique	No
7	No	All thresholds adjusted to make more candidate stations merge with target stations	No



**Figure 15.** Station count by year from 1850 to 2010, compared to the unadjusted GHCN-M version 3 (black), recommended merge (blue), and merge variants (red).



**Figure 16.** Percentage of global coverage in 5° grid boxes compared to the unadjusted GHCN-M version 3 (black), recommended merge (blue), and merge variants (red).

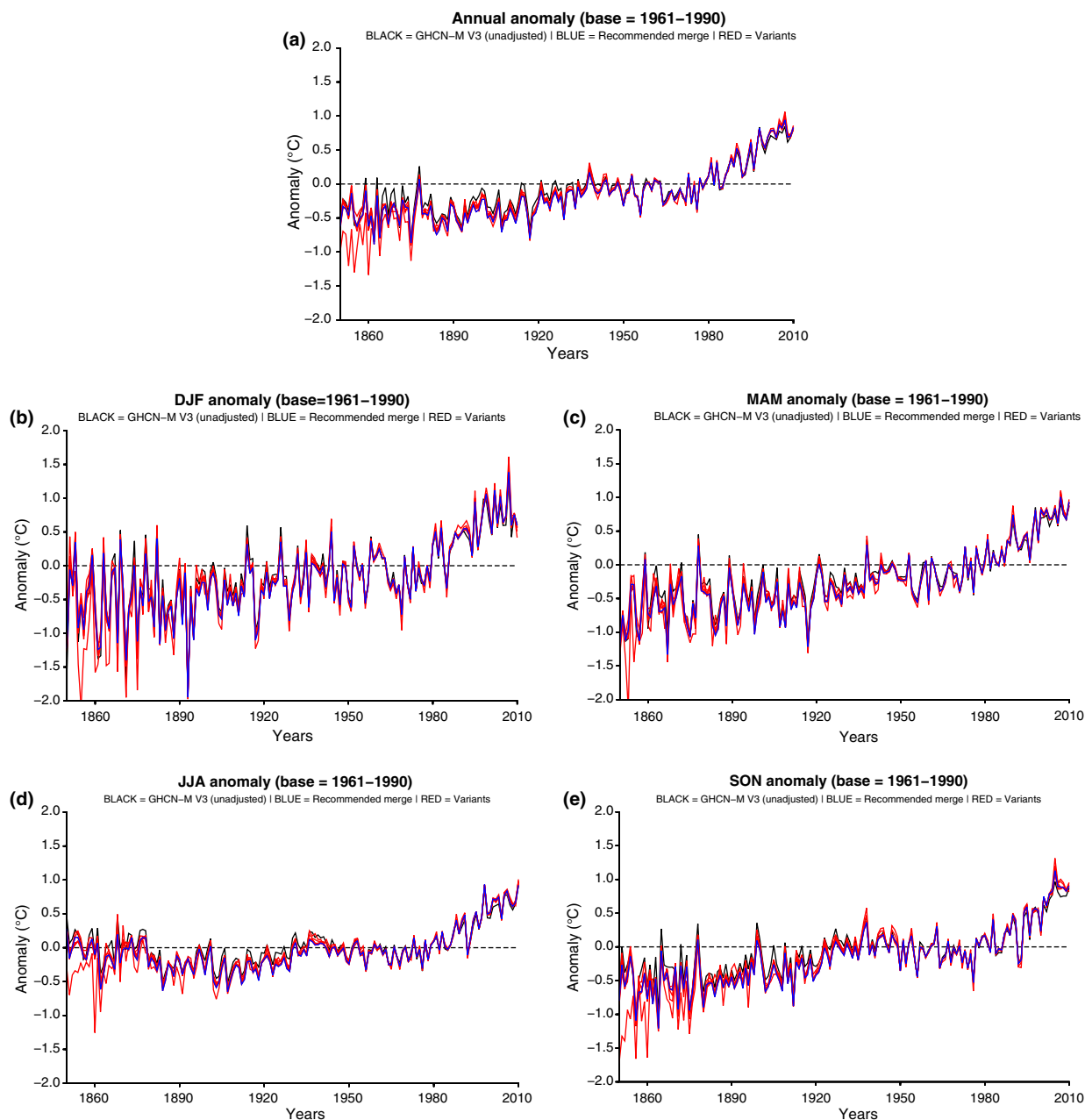
available to the public, to maintain openness and transparency.

The databank provides the foundation from which new methods of analysis, consistent benchmarking of performance and data serving to end-users will be established. Comments are encouraged and can be provided at <http://surfacetemperatures.blogspot.com/>. Open dialogue will be helpful in ensuring the best science is applied and that user needs are met in the construction, analysis, and access of the databank and associated products.

Data submissions are always welcome. Databank submission procedures are ongoing and designed to make the process easy while also ensuring that the data which are submitted are of high quality and traceable. The highest priority is on collection of

temperature data on daily and monthly timescales, but other elements and data collected on sub-daily timescales (e.g. hourly) are collected as they become available and the longer term aspiration is to create a multi-elemental set of integrated synoptic, daily, and monthly land holdings. This will take substantive time, effort, and international coordination.

Policies require the submission of a minimum amount of information about the contributed data including file formats and metadata such as station location and name. The most basic requirement is that the data be provided in the original native format. Examples of possible formats include ASCII text, Microsoft Excel, XML, NetCDF, and any other used by the provider to originally digitize or store the raw observations. This makes the submission easier for the data provider, is in



**Figure 17.** Annual (a) and seasonal (b–e) global temperature anomalies from 1850–2010, with a base period of 1961–1990, compared to the unadjusted GHCN-M version 3 (black), recommended merge (blue), and merge variants (red).

keeping with data provenance procedures that strive to collect the data in a form closest to the original observation, and reduces the possibility of data conversion errors that cannot be identified. A complete guide to data submission procedures is available online (<http://www.surfacetemperatures.org/databank>).

In addition, the user is encouraged to experiment with the techniques used in the merge algorithm. The program was designed to be modular, so that individuals have the option to develop and to implement other methods that may be more robust than described in this article. We will remain open to releases of new versions should such novel techniques be constructed and verified.

## Acknowledgements

This study was supported by NOAA through the Cooperative Institute for Climate and Satellites – North Carolina under Cooperative Agreement NA09NES4400006. We thank the many contributors of data, which made establishment of the databank possible. This includes many National Meteorological Agencies, who have provided data either directly towards the databank effort, or have provided data for input into products such as GHCN-M and GHCN-D. Special thanks go out to CICS-NC interns Jennifer Meyer and Andrew Rogers, who provided essential analysis towards the databank. We also thank members of the public who provided



feedback during the beta release of the databank, including Nick Stokes (<http://moyhu.blogspot.com/>). We would like to thank NCDC's graphics team for providing Figures 2 and 5. Thanks to Richard Chandler, Scott Stevens, Blair Trewin, Jesse Bell, Scott Applequist and anonymous reviewers for providing comments on the paper. A portion of the Antarctic observations (Turner *et al.*, 2004; Lazzara *et al.*, 2012) is partially based on work support by the National Science Foundation, Office of Polar Programs under Grant ANT-0944018 and ANT-1141908. The work of Colin Morice is supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101).

## References

- Allan RJ, Brohan P, Compo GP, 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, doi: 10.1175/2011BAMS3218.1.
- Blunden J, Arndt DS. 2013. *State of the Climate in 2012*. *Bulletin of the American Meteorological Society* **94**: S1–S238, doi: 10.1175/2013BAMSStateoftheClimate.1.
- Caesar J, Alexander L, Vose R. 2006. Large-scale changes in observed daily maximum and minimum temperatures: creation and analysis of a new gridded data set. *Journal of Geophysical Research* **111**: D05101, doi: 10.1029/2005JD006280.
- Camuffo D, Bertolin C. 2012. The earliest temperature observations in the world: the Medici Network (1654–1670). *Climatic Change* **111**: 335–363, doi: 10.1007/s10584-011-0142-5.
- Compo GP, Whitaker JS, Sardeshmukh PD, Matsui N, Allan RJ, Yin X, Gleason BE, Vose RS, Rutledge G, Bessemoulin P, Brönnimann S, Brunet M, Crouthamel RI, Grant AN, Groisman PY, Jones PD, Kruk MC, Kruger AC, Marshall GJ, Maugeri M, Mok HY, Nordli Ø, Ross TF, Trigo RM, Wang XL, Woodruff SD, Worley SJ. 2011. The Twentieth Century Reanalysis Project. *Quarterly Journal of the Royal Meteorological Society* **137**: 1–28, doi: 10.1002/qj.776.
- Hansen J, Ruedy R, Glascoe J, Sato M. 1999. GISS analysis of surface temperature change. *Journal of Geophysical Research* **104**: 30997–31022, doi: 10.1029/1999JD900835.
- Hansen J, Ruedy R, Sato M, Lo K. 2010. Global surface temperature change. *Reviews of Geophysics* **48**: RG4004, doi:10.1029/2010RG000345.
- International Surface Temperature Initiative (ISTI). 2014. Global Land Surface Meteorological Databank. Global Observing Systems Information Center (GOSIC). doi: 10.7289/V5PK0D3G.
- IPCC. 2013. Climate Change 2013: the physical science basis. In Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Stocker TF, Qin D, Plattner GK, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (eds). Cambridge University Press: Cambridge, UK, New York, NY; 187–194.
- Jaccard P. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles* **37**: 547–579.
- Jones PD, Moberg A. 2003. Hemispheric and large-scale surface air temperature variations: an extensive revision and an update to 2001. *Journal of Climate* **16**: 206–223, doi: 10.1175/1520-0442(2003)016<0206:HALSSA>2.0.CO;2.
- Jones PD, Lister DH, Osborn TJ, Harpham C, Salmon M, Morice CP. 2012. Hemispheric and large-scale land surface air temperature variations: an extensive revision and an update to 2010. *Journal of Geophysical Research* **117**: D05127, doi: 10.1029/2011JD017139.
- Karl TR, Williams CN, Young PJ, Wendland WM. 1986. A model to estimate the time of observation bias associated with monthly mean maximum, minimum, and mean temperature for the United States. *Journal of Applied Meteorology and Climatology* **25**: 145–160, doi: 10.1175/1520-0450(1986)025<0145:AMTETT>2.0.CO;2.
- Lawrimore JH, Menne MJ, Gleason BE, Williams CN, Wuertz DB, Vose RS, Rennie J. 2011. An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3. *Journal of Geophysical Research* **116**: D19121, doi: 10.1029/2011JD016187.
- Lazzara MA, Weidner GA, Keller LM, Thom JE, Cassano JJ. 2012. Antarctic Automatic Weather Station Program: 30 years of polar observations. *Bulletin of the American Meteorological Society* **93**: 1519–1537, doi: 10.1175/BAMS-D-11-00015.1.
- Legates DR, McCabe GJ. 1999. Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model evaluation. *Water Resources Research* **35**: 233–241, doi: 10.1029/1998WR900018.
- Menne MJ, Williams CN, Vose RS. 2009. The United States Historical Climatology Network Monthly Temperature Data – Version 2. *Bulletin of the American Meteorological Society* **90**: 993–1007, doi: 10.1175/2008BAMS2613.1.
- Menne MJ, Durre I, Vose RS, Gleason BE, Houston TG. 2012. An Overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology* **29**: 897–910, doi: 10.1175/JTECH-D-11-00103.1.
- Peterson TC, Vose RS. 1997. An overview of the Global Historical Climatology Network temperature database. *Bulletin of the American Meteorological Society* **78**: 2837–2849, doi: 10.1175/15200477(1997)078<2837:AOOTGH>2.0.CO;2.
- Ropelewski CF, Janowiak JE, Halpert MS. 1984. The Climate Anomaly Monitoring System (CAMS). Report, Climate Analysis Center, NWS, NOAA: Washington, DC; 39 pp.
- Skansi M, Brunet M, Sigró J, Aguilar E, Arevalo Groening J, Bentancur O, Castellon Geier Y, Correa Amaya R, Jacome H, Malheiros Ramos A, Oria Rojas C, Pasten AM, Sallons Mitro S, Villaroel Jiménez C, Martínez R, Alexander L, Jones PD. 2012. Warming and wetting signals emerging from analysis of changes in climate extreme indices over South America. *Global and Planetary Change* **100**: 295–307, doi: 10.1016/j.gloplacha.2012.11.004.

- Smith A, Lott N, Vose RS. 2011. The Integrated Surface Database: recent developments and partnerships. *Bulletin of the American Meteorological Society* **92**: 704–708, doi: 10.1175/2011BAMS3015.1
- Thorne PW, Willett KM, Allan RJ, Bojinski S, Christy JR, Fox N, Gilbert S, Jolliffe I, Kennedy JJ, Kent E, Klein Tank A, Lawrimore J, Parker DE, Rayner N, Simmons A, Song L, Stott PA, 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, doi: 10.1175/2011BAMS3124.1.
- Trewin BC. 2010. Exposure, instrumentation and observing practice effects on land temperature measurements. *Wiley Interdisciplinary Reviews: Climate Change*, **1**: 490–506, doi: 10.1002/wcc.46.
- Turner J, Colwell SR, Marshall GJ, Lachlan-Cope TA, Carleton AM, Jones PD, Lagun V, Reid PA, Iagovkina S. 2004. The SCAR READER Project: Toward a High-Quality Database of Mean Antarctic Meteorological Observations. *Journal of Climate* **17**: 2890–2898, doi: 10.1175/1520-0442(2004)017<2890:TSRPTA>2.0.CO;2.
- USGCRP. 2014. *Climate Change Impacts in the United States: The Third National Climate Assessment*, Melillo JM, Richmond TC, Yohe GW (eds). U.S. Government Printing Office: Washington, DC.
- Vose RS, Schmoyer RL, Steurer PM, Peterson TC, Heim R, Karl TR, Eischeid J. 1992. The Global Historical Climatology Network: Long-term monthly temperature, precipitation, sea level pressure, and station pressure data. Report ORNL/CDIAC-53, CDIAC: Oak Ridge, TN; 325 pp.
- Williams CN, Menne MJ, Thorne PW. 2012. Benchmarking the performance of pairwise homogenization of surface temperatures in the United States. *Journal of Geophysical Research* **117**: D05116, doi: 10.1029/2011JD016761.
- Willmott CJ. 1981. On the validation of models. *Physical Geography* **2**: 184–194.
- Willmott CJ, Ackleson SG, Davis RE, Feddema JJ, Klink KM, Legates DR, O'Donnell J, Rowe CM. 1985. Statistics for the evaluation and comparisons of models. *Journal of Geophysical Research* **90**: 8995–9005, doi: 10.1029/JC090iC05p08995.
- WMO. 2011. Guide to Climatological Practices. WMO-No. 100. WMO: Geneva, Switzerland.

## Appendix A

### Source prioritization associated with the recommended merge

The priority in Table 9 follows the nine overarching classes given below. The information necessary to assign each source to a given classification (1–9) should be readily available from the stage 2 data provenance flags:

1. Daily databank stage 3 (GHCN-D) – this provides a backbone of maximum and minimum values that is analyzed regularly and curated carefully on an ongoing basis with regular updates and stable resource support.

2. Data sources which contain maximum and minimum temperature data, have had no quality control or homogenization applied and have known provenance.
3. Data sources which contain maximum and minimum data, have had no quality control or homogenization applied with poorly known provenance.
4. Data sources that have no quality control or homogenization applied but only available as average and have known provenance.
5. Data sources that have no quality control or homogenization applied but only available as average and with poorly known provenance.
6. Data sources with quality control applied that have maximum and minimum data.
7. Data sources with quality control applied that have average data only.
8. Data sources with homogenization that have maximum and minimum data.
9. Data sources with homogenization that have average data only.

Within classes 2–9 the following set of criteria would be used to differentiate between the sources in the priority order with which they should be merged:

1. Whether the monthly data were calculated from dailies held in the databank.
2. Whether the data arise from World Weather Records or national holdings.
3. Average length of station record in the source.
4. Oldest station record start date or average station record start date with priority given to those with earlier start dates.
5. Number of stations in the source.

## Appendix B

### Versioning control procedure for the databank data sources

The version number for stage 2 data uses the following naming structure:

*source.timescale.stage2.X.Y.yyyymmdd.tar.gz*

where

1. *source* identifies the data provider.
2. *timescale* is monthly or daily.
3. *X* is incremented when there is a major change to the source dataset such as replacement or addition of a large percentage of data.
4. *Y* is incremented when there are small updates to the source dataset such as real-time updates to existing stations.
5. *yyymmdd* is the year, month, and day the data source was provided or updated.

Stage 3 uses a slightly different naming structure, namely:

*variant.timescale.stage3.vX.Y.Z.yyyymmdd.tar.gz*  
where

1. *variant* identifies the name of the merge variant.
2. *timescale* is monthly or daily.
3. *X* is incremented during major upgrades of unspecified nature and is always accompanied by a peer-reviewed manuscript.
4. *Y* is incremented during substantial modifications to the databank, including a new set

- of stations or substantive changes to the merging algorithm. This is accompanied by a technical note published to the FTP site.
5. *Z* is incremented during minor revisions to both data and processing software that are tracked via a change log file on the FTP site.
6. *yyymmdd* is the year, month, and day the data source was provided or updated.