Climate Data Record (CDR) Program

Technical Report

International Surface Temperature Initiative

Global Land Surface Databank

Version 1.1.0

Written by Jared Rennie, CICS-NC

1. Introduction

The purpose of this document is to provide an update to the International Surface Temperature Initiative's (ISTI) global land surface temperature databank. This dataset contains global monthly mean temperature (maximum, minimum, and mean) on multiple time scales. Data are collected from in situ networks as well as other national and international providers. The ISTI Steering Committee was formed in 2010 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 ensuring users can fully understand the provenance of the data in the merged holding to the extent that it is known.

In June 2014, the first version of the global databank was released (Rennie *et al*., 2014), which included data from nearly 50 different sources and an algorithm to resolve duplicate stations and piece together complete temperature time series. Since then, there have been monthly updates, appending new data to existing stations. Thanks to user feedback, along with additional analysis, minor changes were introduced and implemented to the merge program to ensure the most accurate data were incorporated in the final product. This, along with updates to current sources required a small change to the versioning system. The remainder of this document will highlight the changes implemented in the global land surface databank, version 1.1.0. More information about the structure of the databank, including sources, formats, and merge algorithm, can be found on the databank website (www.surfacetemperatures.org/databank)

2. Updates to Stage 1 and Stage 2 Data

The databank design includes six data Stages, starting from the original observation to the final quality controlled and bias corrected products. For the purposes of this update, only three stages were modified: digitized data (Stage One), data converted to a common format (Stage Two), and the merged dataset (Stage Three).

The highest priority source comes from the Global Historical Climatology Network – Daily (GHCN-D) dataset (Menne *et al.* 2012). In June 2015, GHCN-D underwent a large update, which included a new average temperature element (TAVG), along with the addition of 1,400 stations that are a part of the World Meteorological Organization's (WMO) Regional Basic Climatology Network (RBCN). Because these stations are important for real time updates, it was necessary to include this new version in the latest merge.

Further assessment was also done on one of our sources known as "russsource." This source contained over 36,000 stations reporting maximum and minimum temperature. While the original format was consistent across all stations, it was discovered that this source included 27 individual sources. It was decided to split these sources up and place them individually in the merge following the source hierarchy defined by the databank working group. Because of some duplication with sources used in GHCN-D, only 20 of the 27 sources were included. In addition, station ID's were brought into the Stage Two data, so that the merge's ID test could be implemented. The same was done for the source known as "ghcnsource."

Other than the above, no additional sources were added to the source hierarchy (Table 1). One source however was removed (crutem4), because it was determined that the use of these stations as a last resort was causing stations to be unique because of the data changes through bias corrections. Candidate stations from crutem4 were matched with their respective target stations through metadata tests, but were chosen as unique from the data tests, because of these corrections. In order to avoid excessive station duplication, this source was removed.

3. Changes to Merge Algorithm

The merge algorithm, as described by Rennie *et al.* 2014, underwent no code changes. However, a couple of thresholds were modified in order to maximize the amount of data the final recommended product would have (Table 2). The thresholds are defined in a configuration file that is required for the program to run successfully.

The first step of the merge algorithm takes into account the metadata between a target and candidate station, including the stations latitude, longitude, elevation and name. A quasi-probabilistic comparison is made and the result is a metadata metric between 0 and 1. In version 1.0.0, this metric needed to pass a threshold of 0.50 in order to be considered for merging. Analysis showed that too many stations were being pulled through and forcing merges between stations that shouldn't have. As a result, a stricter threshold of 0.75 was applied, in order to avoid this issue.

In addition, once a candidate station is chosen to merge with a candidate station, it needs to fill in a gap of at least 60 months (5 years) in order to be added to the target station. It was determined that this gap was too large, and target stations with short gaps in its data were not being filled in by qualifying candidate stations. This gap threshold has been reduced to 12 months as a result.

Similar to version 1.0.0, all decisions made were tested against an independent dataset generated from hourly data for US stations available in the Integrated Surface Dataset (Smith *et al.* 2011). Results, shown in Table 3, show a small change between the results of version 1.0.0 and version 1.1.0.

4. Results

Version 1.1.0 of the recommended merge contains 35,932 stations (Figure 1), nearly 4,000 stations more than v1.0.0 (32,142). Figure 2 depicts that the addition of stations reflect the most recent period, as there is relatively a 10% increase in the number of stations since 1950. It should be noted that there is a drop in coverage prior to 1950 with the new version. However it is the author's opinion that this was reflected by removing crutem4 as one of the sources. Including this source had made candidate stations unique, due to differences in its data as a result of the data providers bias corrections. While the number of stations is lower during this time period for v1.1.0, it should be noted that the number of gridboxes used in analysis (Figure 3) was either equal, or slightly higher than v1.0.0.

Stage Three normally includes a merge recommended and endorsed by ISTI, along with variants showing the structural uncertainty of the algorithm. Due to time constraints, these variants are not available, however will be provided at a later date.

5. References

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.

Rennie, J. J., Lawrimore, J. H., Gleason, B. E., Thorne, P. W., Morice, C. P., Menne, M. J., Williams, C. N., de Almeida, W. G., Christy, J.R., Flannery, M., Ishihara, M., Kamiguchi, K., Klein-Tank, A. M. G., Mhanda, A., Lister, D. H., Razuvaev, V., Renom, M., Rusticucci, M., Tandy, J., Worley, S. J., Venema, V., Angel, W., Brunet, M., Dattore, B., Diamond, H., Lazzara, M. A., Le Blancq, F., Luterbacher, J., Mächel, 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*, **1**: 75–102. doi: 10.1002/gdj3.8

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

6. Tables

Table 1. Summary of Stage Two sources, in prioritized form, used for the recommended version of the merge program, version 1.1.0.

#	Name	Tx	Tn	Tg	#	Name	Tx	Tn	Tg
1	ghcnd	Y	Y	Y	35	ukmet-hist	Y	Y	N
2	mexico	Y	Y	N	36	knmi	Y	Y	Y
3	vietnam	Y	Y	N	37	eklima	Y	Y	Y
4	usforts	Y	Y	N	38	russsource-antarctica	Y	Y	N
5	channel-islands	Y	Y	N	39	russsource-argentina	Y	Y	N
6	ecuador	Y	Y	N	40	russsource-brazil	Y	Y	N
7	pitcairnisland	Y	Y	N	41	russsource-chile	Y	Y	N
8	giessen	Y	Y	N	42	russsource-cuba	Y	Y	N
9	brazil-inmet	Y	Y	N	43	russsource-greece	Y	Y	N
10	brazil	Y	Y	N	44	russsource-indonesia	Y	Y	N
11	argentina	Y	Y	N	45	russsource-iran	Y	Y	N
12	greenland	Y	Y	N	46	russsource-new_zealand	Y	Y	N
13	india	Y	Y	N	47	russsource-south_africa	Y	Y	N
14	gsn-sweden	Y	Y	Y	48	russsource-mexico	Y	Y	N
15	canada-raw	Y	Y	Y	49	russsource-fao	Y	Y	N
16	wwr	Y	Y	Y	50	russsource-fwa	Y	Y	N
17	colonialera	Y	Y	N	51	russsource-australia	Y	Y	N
18	east-africa	Y	Y	Y	52	russsource-australia_de	Y	Y	N
19	uganda	Y	Y	Y	53	russsource-australia_wwr	Y	Y	N
20	antarctica-aws	Y	Y	N	54	russsource-ghcn	Y	Y	N
21	antarctica-palmer	Y	Y	Y	55	russsource-climat	Y	Y	N
22	antarctica-southpole	Y	Y	Y	56	russsource-conus_climat	Y	Y	N
23	ispd-swiss	N	N	Y	57	russsource-ak_hi_climat	Y	Y	N
24	ispd-ipy	N	N	Y	58	germany	N	N	Y
25	ispd-sydney	N	N	Y	59	ghcnsource	N	N	Y
26	antarctica-scar-reader	N	N	Y	60	wmssc	N	N	Y
27	mcdw	N	N	Y	61	central-asia	Y	Y	Y
28	spain	Y	Y	Y	62	arctic	N	N	Y
29	uruguay-inia	Y	Y	Y	63	histalp	N	N	Y
30	uruguay	Y	Y	N	64	hadisd	Y	Y	N
31	swiss-digihom	Y	Y	Y	65	climat-uk	Y	Y	Y
32	ispd-tunisia-morocco	Y	Y	Y	66	climat-prelim	Y	Y	Y
33	sacad_non-blended	Y	Y	Y	67	mcdw-unpublished	N	N	Y
34	japan	Y	Y	Y					

Table 2. List of user defined thresholds in the merge program (version 1.0.0 and version 1.1.0). Changes are noted in red. These thresholds can be altered in the configuration file. Note that the first metadata threshold must be less than the second.

Name	Description	Version 1.0.0	Version 1.1.0
Metadata Threshold	The first metadata threshold that takes into account the distance, height, and jaccard metrics	0.50	**0.75**
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	0.90
Posterior Threshold Same-TXN	Threshold where TMAX/TMIN candidate station has to exceed in order to merge with the target station	0.50	0.50
Posterior Threshold Unique-TXN	Threshold where TMAX/TMIN candidate station has to exceed in order to be considered a unique station	1.30	1.30
Posterior Threshold Same-TVG	Threshold where TAVG candidate station has to exceed in order to merge with the target station	0.50	0.50
Posterior Threshold Unique-TVG	Threshold where TAVG candidate station has to exceed in order to be considered a unique station	0.90	0.90
Overlap Threshold	Overlap period that must exist between the target and candidate station in order to calculate a data comparison via the Index of Agreement	60	60
Gap Threshold	Gap period that must exist when merging a candidate station with the target station	60	**12**

Table 3. Results of validation scheme for versions v1.0.0 and v1.1.0, using an independent dataset.

Version	# Stns	# Merged	# Unique	# Withheld
v1.0.0	1952	1668 (85.45%)	5 (0.26%)	279 (14.29%)
v1.1.0	1952	1657 (84.89%)	7 (0.36%)	288 (14.75%)

7. Figures

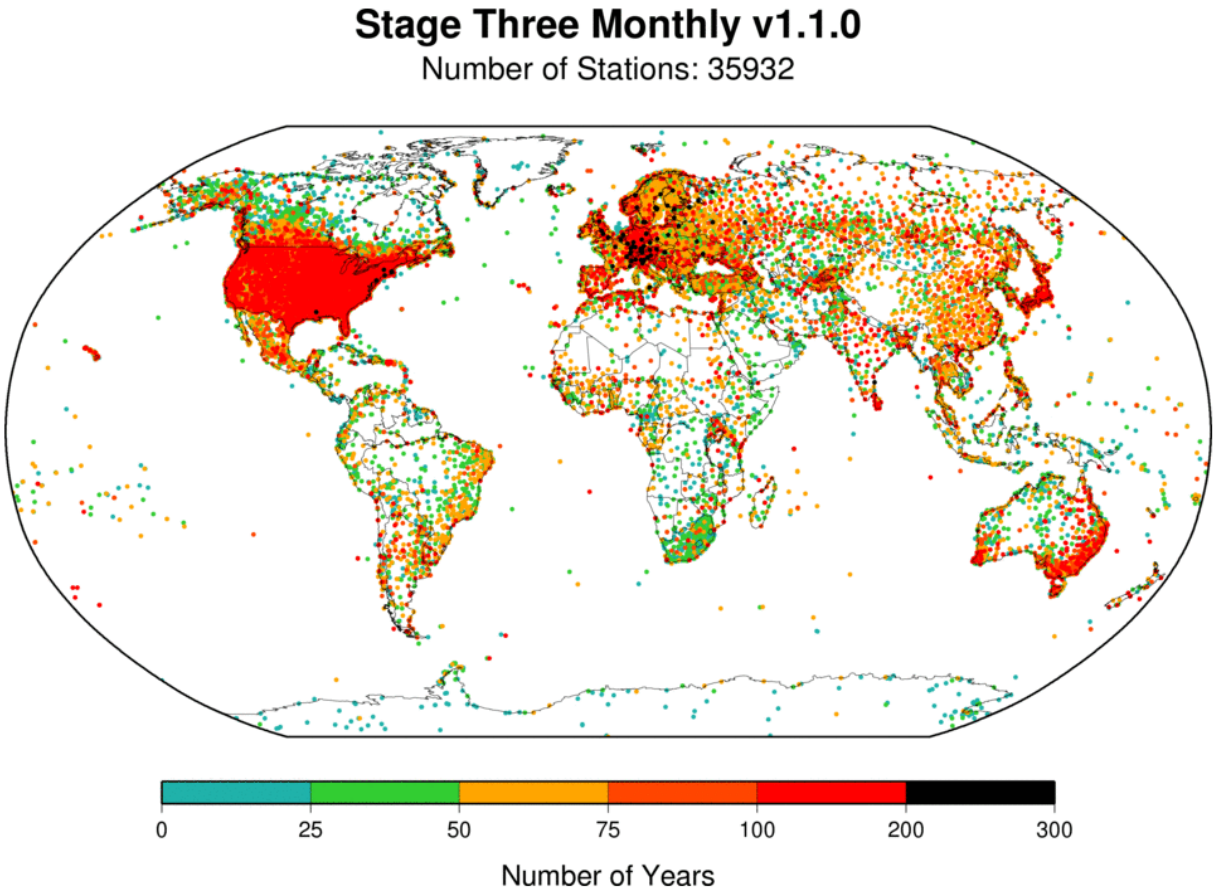

Figure 1. Location of all stations in the recommended Stage Three component of the databank. The color corresponds to the number of years of data available for each station. Stations with longer periods of record mask stations with shorter periods of record when they are in approximate identical locations.

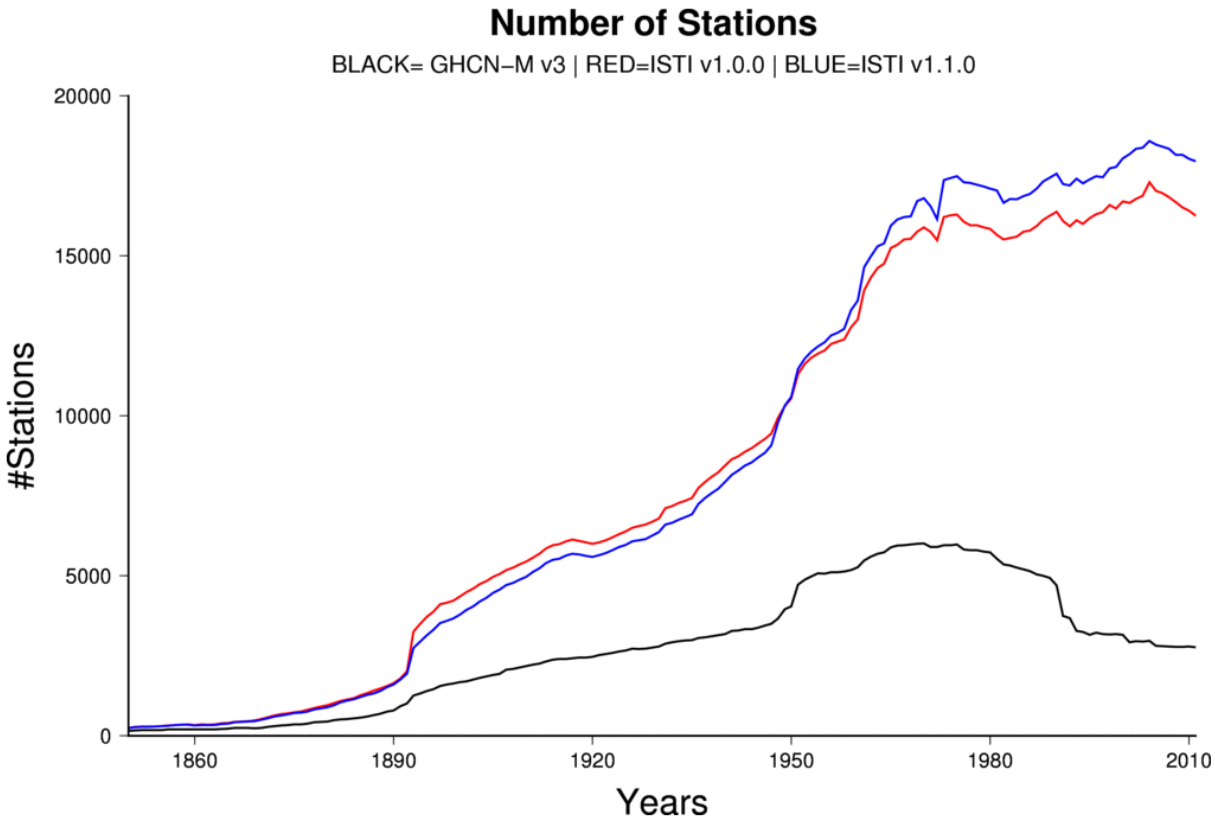

Figure 2. Station count of recommended merge v1.1.0 by year from 1850 to 2014, compared to version 1.0.0, along with GHCN-M version 3

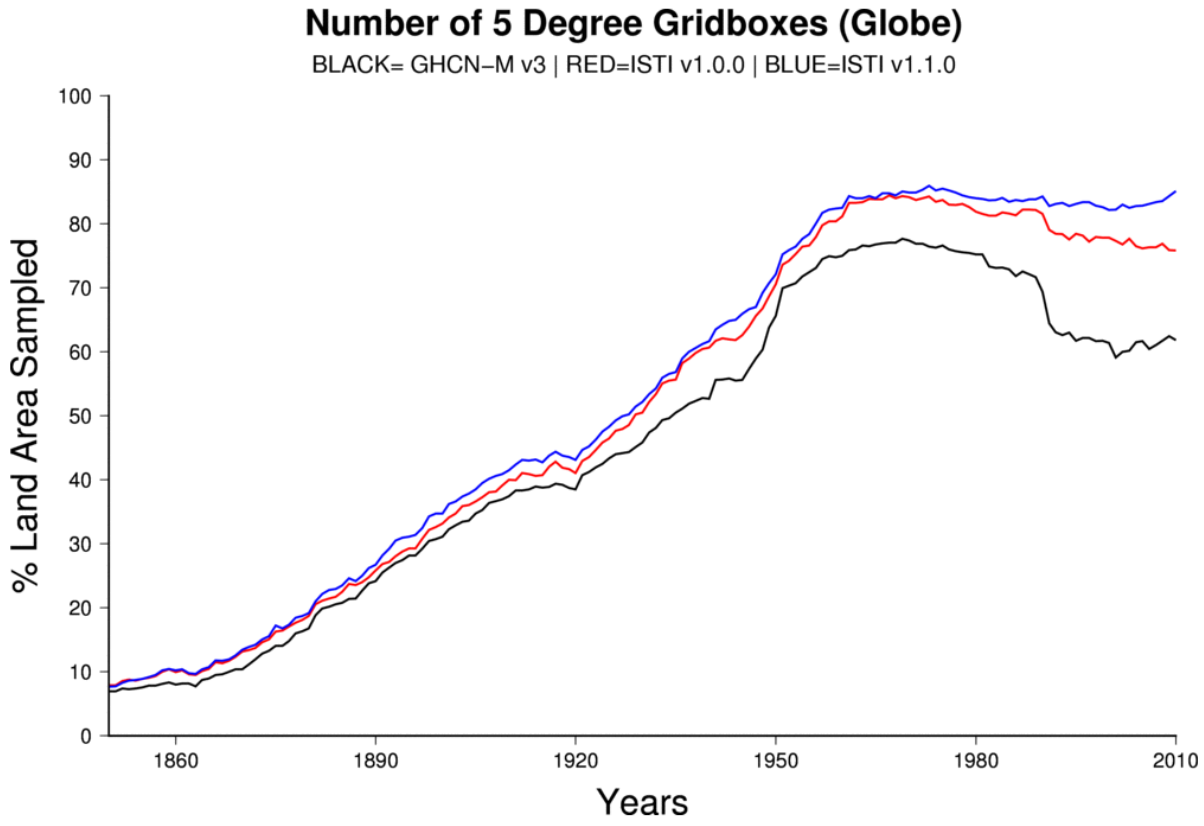

Figure 3. Percentage of global coverage with respect to 5 degree gridboxes for the recommended merge v1.1.0 by year from 1850 to 2014, compared to version 1.0.0, along with GHCN-M version 3