Databank Working Group

2012 Progress Report

November 2012

Current Members:

Jay Lawrimore (Chair)	- NOAA NCDC, USA
John Christy	- University of Alabama, Huntsville, USA
Waldenio Gambi de Almeida	- CPTEC/INPE, Brazil
Kenji Kamiguchi	- Japan Meteorological Agency
Albert Klein-Tank	- KNMI, Netherlands
Matthew Menne	- NOAA/NCDC, USA
Matilde Rusticucci	- Univ of Buenos Aires, Argentina
Vyacheslav Razuvaev	- Russian Research Institute of Hydrometeorological Information
Madeleine Renom	- IFFC, Univ of the Republic, Montevideo, Uruguay
Jeremy Tandy	- UK Met Office, Exeter, UK
Peter Thorne (ex-officio)	- CICS-NCDC, USA
Steve Worley	- National Center for Atmospheric Research, USA

Ex-Members:

Rod Hutchinson	- Australian Bureau of Meteorology
Bryan Lawrence	- BADC, UK

New Members:

Meaghan Flannery	- Australia Bureau of Meteorology
David Lister	- Climatic Research Unit, East Anglia, UK
Albert Mhanda	- ACMAD, Niger
Jared Rennie	- NOAA NCDC, USA

Objectives for October 2011 to October 2012:

1) Continue to add sources of Daily and Monthly timescale data to the Databank. Work with DWG members and others in identifying and collecting readily available sources of digital data.
2) Build upon Data Rescue activities and leverage crowd sourcing efforts to begin volunteer digitization of land surface records.
3) Develop an approved methodology for merging sources of data to create a monthly Stage 3 data product. Include a hierarchy of source data from which to build the merged dataset.
4) Launch version 1 of the Databank for monthly timescale data in April 2012, making all data, processes, and software freely available and accessible.
5) Complete and submit journal article describing version 1 of the Databank and its underlying principles.
6) Work with the Benchmarking and Assessment working group to expand opportunities for incorporating Stage 4 and 5 data into the databank.

Objectives Met:

1) Continue to add sources of Daily and Monthly timescale data to the Databank. Work with DWG members and others in identifying and collecting readily available sources of digital data.

As the year began the databank consisted of 21 sources of data. Because of the attention the working group members gave to identifying new sources of data, this number more than doubled in the following twelve months. By October 2012 the databank contained data from 45 sources with most of these at the daily and monthly timescales. The list of sources is available at http://www.surfacetemperatures.org/databank .

This greatly increases the temporal and spatial coverage throughout the world when compared to existing global datasets such as the Global Historical Climatology Network-Monthly. As shown in Figure 1 the additional sources provides a more than three-fold increase in the number of stations once merged into a single Stage-3 dataset (Figure 1). This results in 5 to 10% increase in spatial coverage (Figure 2). In addition to improving coverage of global land area, the greater station density is especially beneficial in neighbour comparisons used for quality control and bias correction.

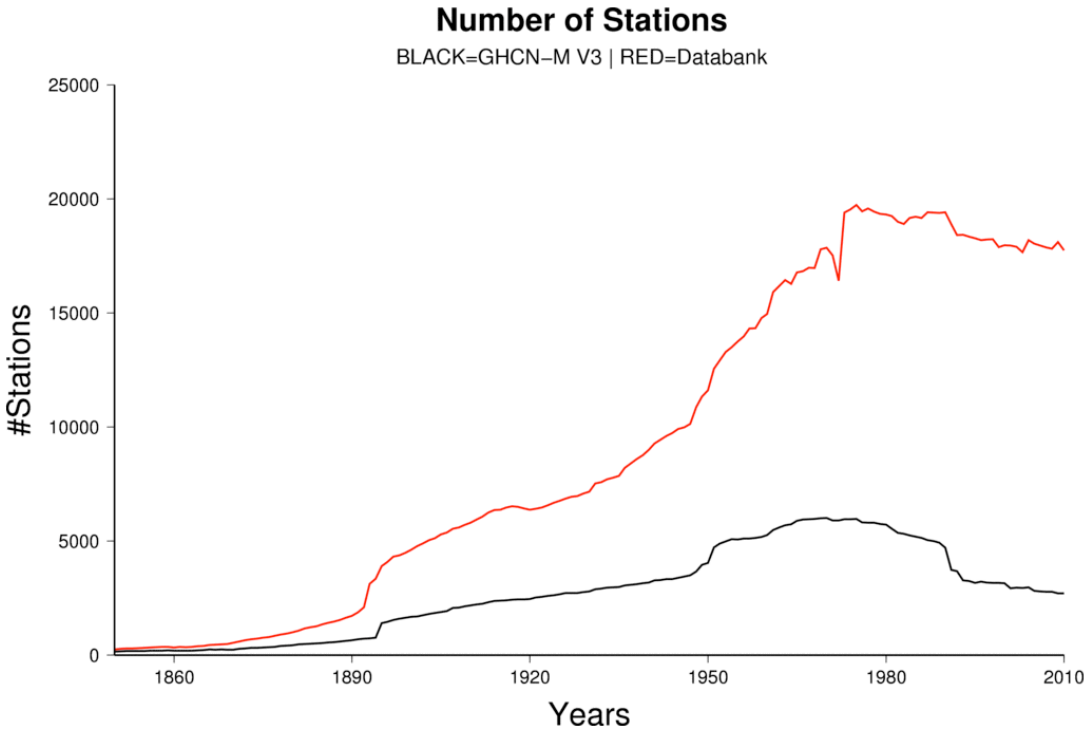

Figure 1. Number of stations in the Stage-3 dataset (red) and the GHCN-M version 3 dataset (black).

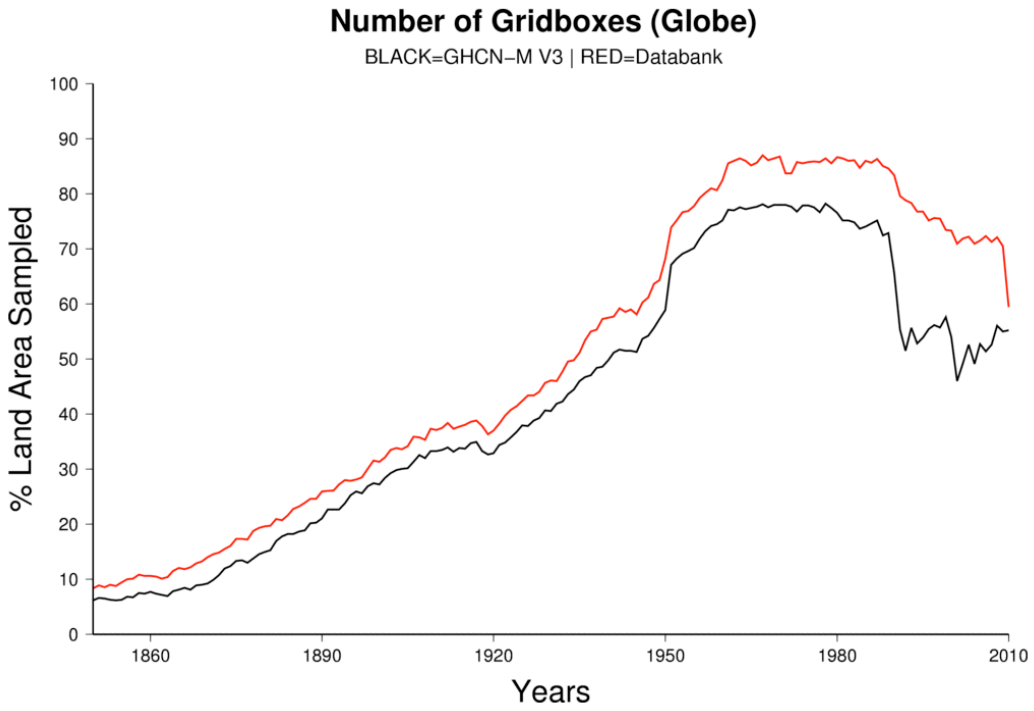

Figure 2. Percentage of land coverage based on averaging within 5X5 degree grid boxes. Databank (red) and GHCN-M version 3 (black).

2) *Develop an approved methodology for merging sources of data to create a monthly Stage 3 data product. Include a hierarchy of source data from which to build the merged dataset.*

In September 2012 the working group approved a methodology for merging the 45 sources of monthly mean temperature into a single Stage-3 dataset. The sources are merged based on a hierarchy with assigned priority for each. Sources with higher priority take precedence over 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, are closest to the original raw observation, have extensive metadata, come from a national or international holding, or have long and consistent periods of record are in the upper tier of the hierarchy. Once the hierarchy is established a series of metadata tests are applied for each target and candidate station. This is followed by comparisons of overlapping data when they exist.

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 broadly-speaking Bayesian in approach and based upon metadata matching and data equivalence criteria. A simple flow chart of the algorithm can be found in Figure 3.

The probability of a station match is first calculated based on a weighted sum of individual probabilities using station location, name, elevation, and record start dates

between the target and candidate station. Using a quasi-Bayesian approach, the probabilities are combined to form a posterior probability of possible station match, known simply as the metadata probability. Comparisons of station location are given the greatest weight while elevation is given the least because errors in elevation are known to occur frequently.

Metadata probability= ((9*dist)+(1*elev)+(5*name)+(5*start date))/20

For any of candidate stations that pass the metadata threshold (probability >0.50), a data comparison is made between that target station and candidate station when there is a minimum overlap between the two stations of 60 months. For target-candidate pairs without the minimum months of overlapping data the candidate can be merged with the target if the metadata probability exceeds 0.85.

The data comparison is performed for maximum and minimum temperature separately using the Index of Agreement (IA; Willmont 1981). Resulting values range between 0 and 1. Lengthier periods of overlap can lead to higher values of IA, so this was further refined to take into account the actual months of overlap between the target and candidate station. A lookup table was generated to provide a probability of station match (H1), as well as station uniqueness (H2). Bootstrapping was applied by changing the shifts in mean and variance of certain criteria, and calculating IA 1,000 times. A cumulative distribution function was fit 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 in order to be considered a station match.

The five resulting probabilities, one metadata and four data probabilities (tests for station match and uniqueness, for both TMAX and TMIN), are recombined to form two new posterior probabilities, one of station match, and one of station uniqueness.

$$posterior\ probability\ same_{TMAX/TMIN} = \frac{metadata\ probability * H1_{tmax} * H1_{tmin}}{3}$$

$$posterior\ probability\ unique_{TMAX/TMIN} = (1 - metadata\ probability) + H2_{tmax} + H2_{tmin}$$

If any of the values returned for *probability same* exceed the same threshold of 0.5, then the candidate station is merged with the target station with the highest *posterior probability same*. If none of the stations exceed that threshold, but one of the *posterior probability unique* values exceeds the unique threshold, then the candidate station becomes unique and is added to the target dataset. In some cases the probabilities are inconclusive and it is not possible to clearly determine if the station is unique or the same as the target. In such cases the station is withheld from the merge.

Although this is believed to be the best method for merging the source data, there are other plausible merging methods based on other source hierarchies and other metadata and data matching thresholds. To illustrate the impact of other decisions on the Stage-3 merge, several variants are included on the databank website along with the recommended merge. These are available at ftp://ftp0.ncdc.noaa.gov/pub/data/globaldatabank/monthly/stage3/ .

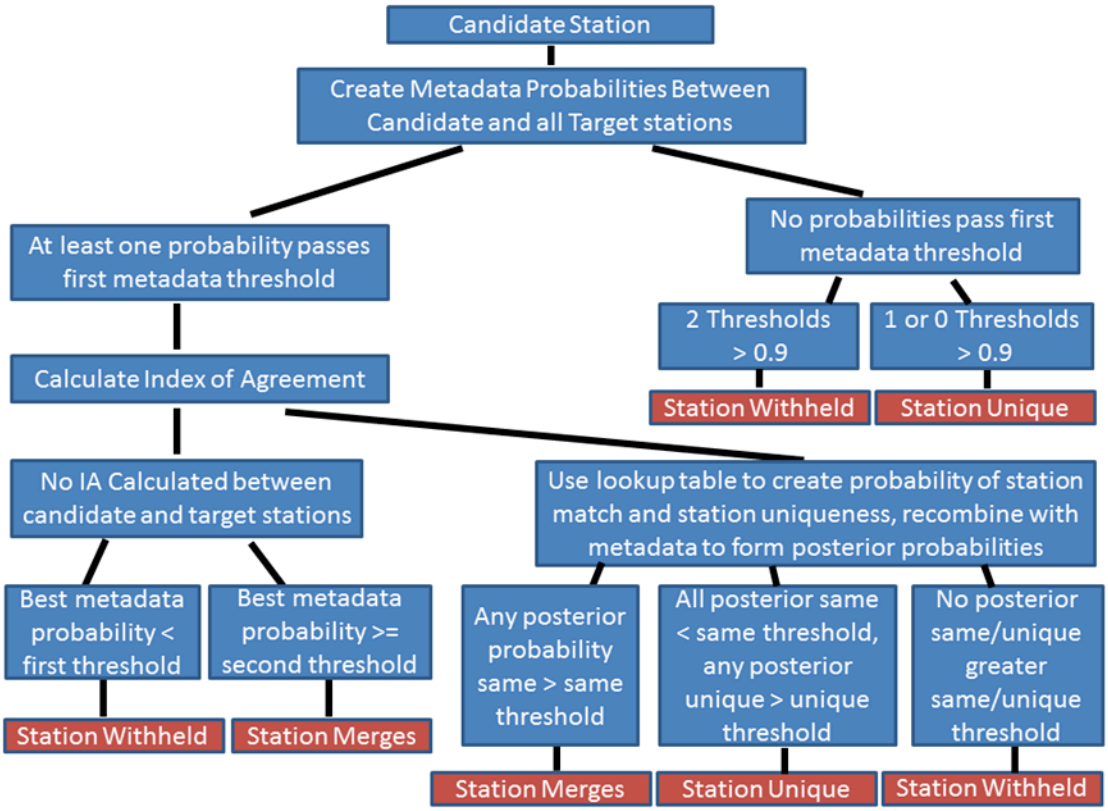

Figure 3. Flow chart of the merge algorithm.

3) *Launch version 1 of the Databank for monthly timescale data in April 2012, making all data, processes, and software freely available and accessible*

The monthly databank was released in beta form on 4 October 2012. The working group agreed to provide the databank in beta form for a period of three months. This gives the user community an opportunity to review and provide feedback on the databank and for modifications to be made in response to feedback. Notable changes to the merging algorithm made in response to user feedback include the addition of station record start date to the metadata comparison. Version 1 is expected to be released in January 2013.

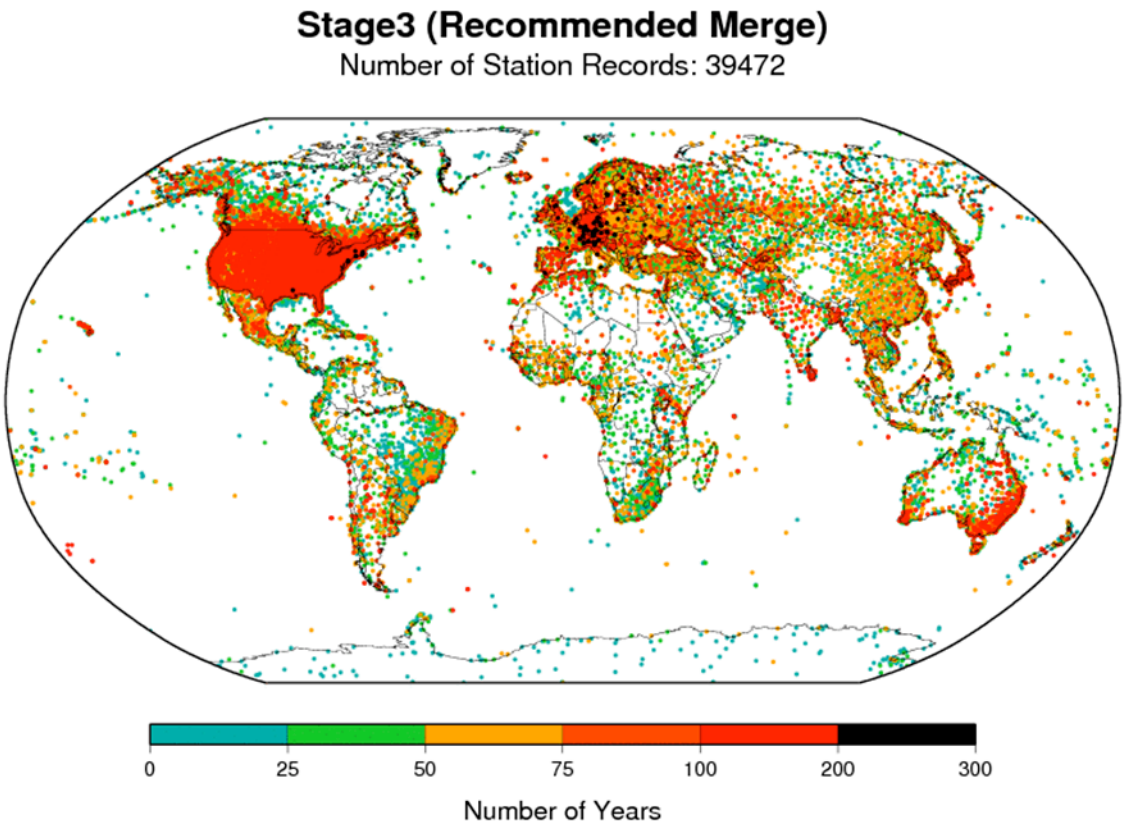

Figure 4. The location of the more than 39000 stations included in the Stage-3 version 1 beta release.

4) *Complete and submit journal article describing version 1 of the Databank and its underlying principles.*

An article entitled *An International Response to the Need for Better Global Temperature Data* was submitted to Eos as a Brief Report (1500 words) in September. This article provides a brief overview of the databank and its aims as part of the International Surface Temperature Initiative. The authors are awaiting reviewer comments.

A more in depth article that includes finer details on the databank and the merging methodology is nearing completion and will soon be submitted to the new open access Geosciences Data Journal.

Objectives Not Met:

1) Build upon Data Rescue activities and leverage crowd sourcing efforts to begin volunteer digitization of land surface records.

The working group continues to pursue crowd sourcing opportunities for digitizing the billions of historical land surface observations taken during the 19th and early 20th century that exist on imaged and hard copy forms. The success of crowd sourcing has been proven for marine records and a new effort initiated in 2012 for crowd source

digitization of tropical cyclone images further proves its viability. There are unique challenges associated with land surface records. The sheer range of forms and formats dictates a flexible system which will cost substantially more to build and maintain. In addition there is less historical information included with the land observations as there is in ship log records. Retaining the interest of volunteers will be a bigger challenge. To create an equally successful crowd sourcing project for land surface observations requires new funding sources. Through the leadership of the Crowdsourcing team sources of funding continued to be pursued. Proposals were submitted and expectations remain high that the databank effort will benefit from crowd sourcing in the future.

2) Work with the Benchmarking and Assessment working group to expand opportunities for incorporating Stage 4 and 5 data into the databank.

The focus remained on completing the Stage-3 merge in 2012. With the successful beta release in October 2012 and the expected release of version 1 in January 2013 collaboration with the Benchmarking and Assessment working group will be possible in 2013.

2012 Annual Overview:

2012 was a successful year for the global databank. Although the original goal of completing version 1 of the databank slipped from April to October, the completion of this milestone was a great success. The more than two-fold increase in the number of sources to 45 not only greatly benefited the Stage-3 merge of the monthly dataset, the large number of daily sources sets the stage for development of the daily dataset. The full version 1 release will be completed in January 2013, but this does not mean the end of development activities. The working group will continue collect and add new sources to the databank. These will be added to the monthly Stage-3 dataset in incremental updates. In 2013 work also is expected to begin on the daily version of the databank and collaboration with the Benchmarking and Assessment working group will focus on development of quality controlled and bias corrected stages of the monthly dataset.

Objectives for October 2012 to October 2013:

1) Incorporate user feedback, making further improvements to the monthly Stage-3 dataset and complete a full version 1 release in January 2013.
2) Continue to collect and add new sources of monthly, daily, and sub-daily data to the databank.
3) Complete incremental updates to the monthly Stage-3 dataset.
4) Begin development of the daily Stage-3 dataset.
5) Work with the Benchmarking and Assessment working group to expand opportunities for incorporating Stage 4 and 5 data into the databank.
6) Complete publication of articles describing the Databank and its goals as part of the International Surface Temperature Initiative.