Benchmarking and Assessment Working Group

2014 Progress Report
October 2014

Current Members:

John Christy	- University of Alabama, Huntsville, USA
Meaghan Flannery	- Australia Bureau of Meteorology
Waldenio Gambi de Almeida	- CPTEC/INPE, Brazil
Byron Gleason	- NOAA NCDC, USA
Kenji Kamiguchi	- Japan Meteorological Agency
Albert Klein-Tank	- KNMI, Netherlands
Jay Lawrimore (Chair)	- NOAA NCDC, USA
David Lister	- Climatic Research Unit, East Anglia, UK
Matthew Menne	- NOAA NCDC, USA
Albert Mhanda	- MSD, Meteorological Services of Zimbabwe
Colin Morice	- UK Met Office, Exeter, UK
Vyacheslav Razuvaev	- Roshydromet, Russia
Jared Rennie	- CICS-NC/NOAA NCDC, USA
Madeleine Renom	- IFFC, Univ of the Republic, Uruguay
Matilde Rusticucci	- Univ of Buenos Aires, Argentina
Jeremy Tandy	- UK Met Office, Exeter, UK
Peter Thorne (ex-officio)	- NERSC, Bergen, Norway
Steve Worley	- National Center for Atmospheric Research, USA

October 2013 to October 2014 Objectives:

1) Complete software enhancements associated with code review
2) Complete operational readiness requirements for version 1 release
3) Version 1 release of Stage-3 data
4) Begin collection of parallel measurements
5) Enhance metadata collections
6) Add to collections in data sparse areas

Objectives Met:

1) Complete software enhancements associated with code review

To meet internal NCDC requirements associated with release of version 1 of the Databank, a code review of the Databank software was conducted from November 20-22, 2013. The review was conducted by two Fortran and scripting language experts and moderated by the lead Databank software developer. The review was separated into the four types of code in the databank process:

- • Convert source data in its native form (Stage One) to a common format (Stage Two)
- • Convert Stage Two DAILY data to Stage Two MONTHLY Data
- • Format Checker for ALL Stage Two Data

- Merge Program combining all MONTHLY Stage Two Data into a consolidated global data set (MONTHLY Stage Three).

The first three parts of the program were written in the Perl Scripting Language, while the Merge Program was coded in FORTRAN 95. The review team made 19 recommendations that primarily focused on adherence to Climate Data Record Program (CDRP) standards and other best programming practices. The lead Databank programmer responded to each of the recommendations which resulted in improvements to the overall software design. Changes included consolidation or removal of eight modules and the addition of five new modules. In addition the overall complexity of the software system was reduced through refactoring. As an example a static analysis of the main merge program before and after refactoring showed a 45% reduction in its complexity, which will greatly simplify long-term maintenance of the program.

2) Complete operational readiness requirements for version 1 release

An NCDC Operational Readiness Review of version 1 of the Monthly Temperature Databank was completed on May 30, 2014 and approval given for release of the Databank Stage 3 version 1.

3) Version 1 release of Stage-3 data

The official release occurred on June 30 accompanied by announcements on the NCDC "News" online and by the Nansen Environmental and Remote Sensing Center (NERSC). An announcement also was posted on the Research Data Archive at the US National Center for Atmospheric Research (NCAR; http://rda.ucar.edu/). It was subsequently highlighted on the UK Met Office research news webpage (http://www.metoffice.gov.uk/research/news/2014/isti-databank-release) and by the National Physical Laboratory in the UK and noted in a Guardian article.

The version 1 release contains data from 49 sources and totals slightly more than 32,000 stations. The GHCN-Daily dataset provides approximately 85% of the stations. The Databank received a Digital Object Identifier (DOI) of 10.7289/V5PK0D3G.

A peer review journal article, *The international surface temperature initiative global land surface databank: monthly temperature data release description and methods* (Jared Rennie et al. 2014) was published in the open source Geoscience Data Journal and is available in early view at http://onlinelibrary.wiley.com/doi/10.1002/gdj3.8/abstract .

Monthly updates of the most recent month's data are now being made to the Databank from GHCN-Daily source and global CLIMAT data. These updates are available by the 11th of each month. The previous month's data are placed in an archive directory on the databank ftp site and are also permanently archived at NCDC (ftp://ftp.ncdc.noaa.gov/pub/data/globaldatabank/archive/).

New sources will be incorporated over time and a remerge conducted on an annual basis.

4) Begin collection of parallel measurements

Parallel measurements consist of the side-by-side collection of climate observations using different instrumentation taken over a sustained period of time. Over the past several decades a number of independent researchers, typically associated with universities or federal governments, collected parallel measurements in an effort to quantify measurement biases caused solely by differences in instrumentation. The Databank's focus is on the creation of a central storehouse of such observations for the long-term preservation and access to such observations.

This activity leverages the leadership efforts of Victor Venema at the University of Bonn. A description of the effort and the database structure are provided at http://variable-variability.blogspot.com/2014/08/database-with-parallel-climate-measurements.html . The proposal for a file format (https://drive.google.com/file/d/0B7sJmg1UW9uGdXZ0M3BuRU54ckk/edit?usp=sharing) and a list of potential data sources (https://ourproject.org/moin/projects/parallel) are also available. Data contributions are being encouraged and many additions to the parallel database are expected in the coming year.

In the coming year a new task team will be established within the Databank Working Group. Its principal aim will be collection of parallel measurements and development of a living and accessible archive. Terms of Reference for this task team will be established.

5) Enhance metadata collections

The Databank Working Group made little progress in the collection of new metadata holdings for existing and new data. Metadata primarily consists of station name, location, and elevation. Additional metadata such as current and historical information on instrumentation, station environment, and maintenance practices remain practically non-existent for the majority of databank stations. Exceptions include new station metadata provided by the US Environmental Protection Agency for a small network of stations in the state of Oregon (Oregon Crest to Coast network), and metadata for the US Climate Reference and the US Cooperative Observers Network. The USCRN and COOP network metadata are available in NCDC's Historical Observing Metadata Repository (HOMR; http://www.ncdc.noaa.gov/homr/).

6) Add to collections in data sparse areas

Strides were made in the collection of new data sources aimed at enhancing coverage in data sparse areas, principally Africa and S. America. The following ten additional sources were collected and will be integrated into the version 1.1 release which will take place during the first half of 2015.

- UK Stations from the Met Office (300+ stations)
- German data released by DWD (1000+ stations)
- EPA's Oregon Crest to Coast Dataset (24 stations)
- LCA&D: Latin American Climate Assessment and Dataset (148 stations)

- Daily Chinese Data (380 stations)
- NCAR Surface Libraries (unknown number of stations)
- Stations from Meteomet project (240 stations)
- Libya Stations sent by their NMS (9 stations)
- C3/EURO4M Stations (80 stations)
- Additional Digitized Stations from the University of Giessen (10 stations)
- Homogenized Iranian Data (50 stations)
- Long-term Swiss data (7 stations)

New data sources in the most data sparse areas include the Latin American Climate Assessment and Dataset consisting of stations within Suriname. In addition the set of Libyan stations provided by their National Meteorological Service consists of nine stations with period of record data from 1944 to 2010.

There are other sources which are expected to be available in the coming year. Although some of these will have been homogenized, the databank will attempt to acquire the unadjusted data. Sources include the following.

- HISTALP: expected to be completed by early 2015 (Ingeborg Auer).
- NORDHOM: data available from Nordic countries by the end of this year (Erik Engstrom, SMHI)
- Monthly data for the Pyrenees from about 1950 (Marc Prohom, Servei Meteorològic de Catalunya)

Objectives Not Met:

Objectives 4, 5 and 6 above remain open and will continue to receive attention from the Databank Working Group in 2015.

2014 Annual Overview:

The first major accomplishment of the Databank Working Group was completed in 2014 with the release of version 1 of the Stage 3 databank. This was made possible by contributions from every working group member and their combined efforts in developing the vision and practices that established the Stage 0 through Stage 3 data. Contributions from National Meteorological Services in every WMO region brought in 49 sources of data. New practices such as the establishment of data provenance tracking flags enabled the DWG to enhance the stewardship and transparency of the more than 32,000 stations included in the version 1 release.

As the working group brought the version 1 release to a close it continued to collect new sources of data to help fill temporal and spatial gaps in the temperature record. It also began establishing collections of parallel measurements and made initial strides toward enhancing metadata collections. These efforts will be further expanded upon in the coming year.

In 2015, the DWG will look toward the next phase of the databank; development of a daily timescale databank. Given the existing attributes of the NCDC's Global

Historical Climatology Network-Daily (GHCN-Daily) dataset, there is little need to reengineer a new daily merging algorithm. As such the DWG will have the GHCN-Daily dataset serve as the daily databank. The GHCN-Daily dataset contains more than 27,000 stations with daily maximum and minimum temperature data. Stage 1 and Stage 2 versions of GHCN-Daily sources for ISTI already occurs internal to NCDC as part of the GHCN-Daily processing system. Some of these are redundant to the current databank, but most are not. The existing NCDC daily merge algorithm will serve to create Stage 3 data. The GHCN-Daily merge algorithm is currently analogous to the monthly merge algorithm (i.e., uses station metadata matching/data matching).

In addition to providing the merged Stage 3 daily dataset, by providing the Stage 1 and Stage 2 data we will permit an alternative merge algorithm if another organization wants to commit resources to such an effort.

In the coming year the focus will be on incorporating new daily sources into GHCN-D as time permits. There may be some temperature only stations in GHCN-D. However, a major benefit of using GHCN-Daily as the Daily databank is that it is multi-element. There is a growing awareness of the need to maintain datasets that are integrated horizontally across all climate elements. GHCN-Daily establishes this principle, which provides the added benefit of using multi-elements to improve quality control and homogenization, and the source merging process through better data matching.

Objectives for October 2014 to October 2015:

1) Continue collection of parallel measurements and integrate into collection.
2) Enhance metadata collections
3) Add to collections of monthly and daily data in data sparse areas
4) Integrate at least 10 new sources into the Monthly databank and release version 1.1
5) Integrate new sources into GHCN-Daily

Suggested timeline and plan for achieving objectives:

Objective	Description	Responsible Members	Deadline
Establish Parallel Measurements task team	Task team will serve to establish the parallel measurements collection.	Victor Venema (Chair) Members TBD.	January 2015
Continue collection of parallel measurements and integrate parallel measurements into consolidated collection.	Integrate data into established format for parallel measurement collection.	All, lead by Victor Venema and Jared Rennie	September 2015
Add at least 10 new sources to Monthly databank and release version 1.1.	Conduct merge of new sources into monthly databank as part of version 1.1 release.	Jared Rennie and Merge Team	March 2015

Addition of new sources to GHCN-Daily	Work with NCDC Science Council and DWG to select and add candidate sources	Matt Menne	September 2015
Metadata collection	Add at least two new sources of metadata to Databank	Databank Working Group	September 2015
Conduct pilot experiment for extension of IMMA format to land meteorological data	Select one land source and translate into modified IMMA format	Lawrimore, Woodruff (Guest expert)	June 2015
Plan for advancing multi-element databank holdings	With the ISTI Steering Committee establish plan for multi-element holdings	Lawrimore, Menne, Thorne	June 2015
Ongoing activities			
Advocacy of the databank, efforts to augment holdings	Every effort should be made to engender data submissions	Steering committee, Databank working group	Ongoing
Data rescue	Continued pursuit of funding proposal for support of crowdsourcing of already imaged forms (such as NOAA foreign data library)	Data rescue task team / Databank Working Group	Ongoing until successful
Parallel measurements database data collection	Pursuit of parallel measurements data holdings	Databank Working Group / Victor Venema	Continuous