Databank Working Group Update: May 6th, 2015

Monthly Databank Update

Jared Rennie, Matt Menne and Jay Lawrimore

The databank has gone through a number of routine updates since the release of version 1. We have set up a near real time update system, which incorporates the latest data (including GHCN-Daily and CLIMAT streams). These updates occur on a monthly basis, and the newest data is posted on the FTP site no later than the 11th of each month. These updates have been the basis for testing and evaluation of the next version of GHCN-Monthly (version 4), which includes updates to its quality control and bias correction processes. We have an end to end process on our developmental server, and we are currently calling it version 4 alpha 1 (v4.a.1).

During the evaluation of version 4 alpha, we have taken a closer look at some of the specifics of the databank version 1.0.1 stage 3 merge and have identified a number of cases of undesirable merge results that had a variety of causes. These causes are listed below.

1. Including homogenized data sources with joined (threaded) station records when these records are also included separately as higher ranking sources. In these circumstances, the merge algorithm cannot be expected to reconcile the two sources and the best decision is to withhold the threaded station in the lower ranking source; however, this does not always happen automatically with the current merge algorithm.
2. Grid box comparisons between GHCN-Monthly v4 alpha and the operational version 3 revealed a suspicious drop in currently station coverage in the v4 alpha (Databank) product, particularly with the most recent months. Analysis showed that stations from CLIMAT streams were not appending new data correctly because of metadata/data match issues between the databank stations and the CLIMAT sources.
3. In addition, in some cases, there was lack of data in a station's base period (here we define as 1961-1990) that prevented the calculation of a gridbox anomaly. In diagnosing these situations, we noted additional cases of station records merging with the wrong records from higher ranking sources. These merge problems appeared to result from the use a low metadata metric threshold for allowing data comparisons, which led to errors that favored data merges over adding unique stations.
4. Additional merge issues were also caused by not exploiting id matching for two major sources (ghcnsource and russsource) and requiring a source to contribute at least 5 years of new data to an existing record.

These issues were remedied as follows:

1. Removing homogenized sources that caused issues with threaded data
   Central-asia, arctic, histalp, crutem4
2. Ensuring that all of the RCBN stations that submit CLIMAT matches are represented in GHCN-Daily. This had led to the addition of approximately 1200 new stations in GHCN-Daily from the Global Summary of the Day (GSOD). These additions allow a direct id match between the CLIMAT WMO number and GHCN-Daily. It also permits the calculation of monthly mean temperature directly from daily data for all CLIMAT sites in real-time, which will help overcome CLIMAT transmission problems when they occur.

3. A series of merges were run on the validation dataset and GHCN-Daily with differing metadata thresholds to see if there was a more optimal threshold to use that would yield a better balance between merging sources and assigning a station as unique. Increasing the first metadata threshold from 0.50 to 0.75, requiring matches on metadata to be stricter. Running through our validation scheme proved 0.75 to be the most efficient (Table 1).
4. Updating two stage2 sources (ghcnsource and russsource) to incorporate the ID's that were incorporated in its respective Stage1 data. Including the ID's would help match stations through the merge's ID test module. In addition the gap threshold from 60 months to 12 months in order to help piece together stations with small gaps, especially during the defined base periods (Figure 1a and 1b).

A recent run of the merge, with the above changes, shows that there is substantial improvement in gridbox coverage. Figure 2 shows our analysis of gridboxes for the year 2014. The purple squares include data in both V3 and V4, the red squares include data in V4 not in V3, and the blue dots are data in V3 not in V4. Overall, there are 149 new gridboxes in v4.a.1, and only 6 gridboxes with v3 only data. Earlier assessments showed that we had 50+ gridboxes with v3 only data, so this is a huge improvement. We have not had enough time to analyze these gridboxes before the call, but we suspect some of them are due to updated metadata putting a station in a different gridbox.

Next steps include updating data that has been in the queue for quite some time. The cutoff date for new sources has passed (2/28/2015). Efforts to convert to Stage 2 have lagged, due to this gridbox analysis and other priorities. However we will move forward to putting these sources (Table 2) into the databank and incorporate in the latest merge, and we will soon have a Databank version 1.1 release. The Working Group needs to review the list of stations to confirm whether all should be included in the next merge. If some are already in ECA&D for example, it may be best to omit from the next merge since they can enter via GHCN-Daily. This raises the larger issue of managing a multi-element land surface databank rather than a temperature only databank.

Figures

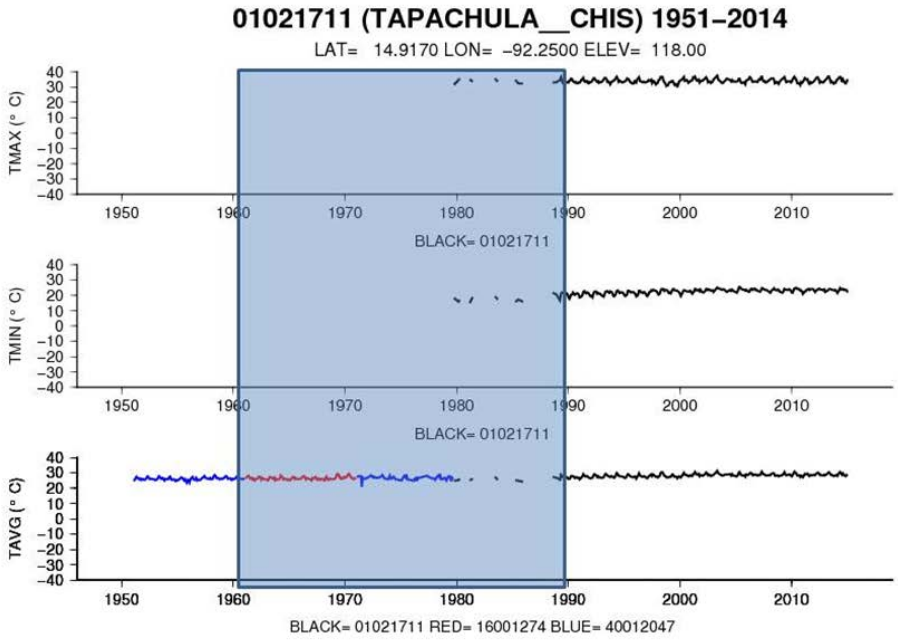

Figure 1a: Time series of TMAX/TMIN/TAVG data for Tapachula, Mexico. Blue box indicates data used for the 1961-1990 Base Period

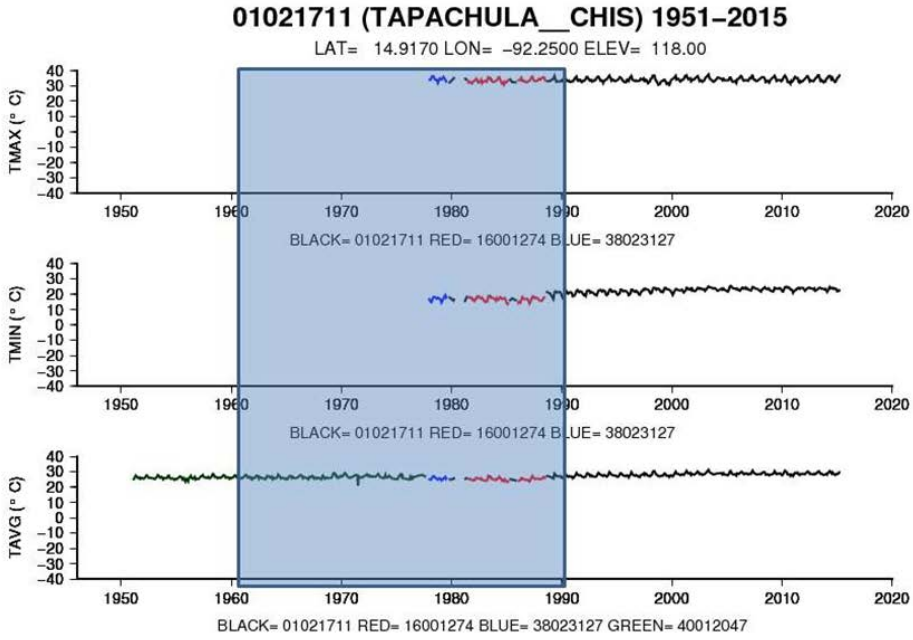

Figure 1b: Same as 1a, but with the latest version of the merge program, which includes lowering the gap threshold to 12 months.

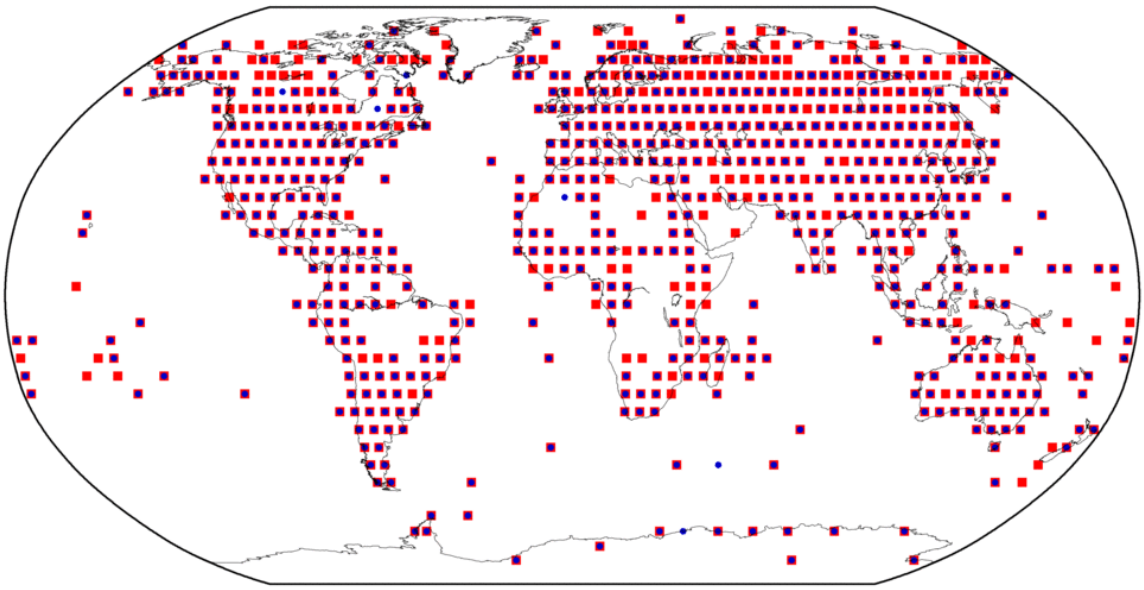

Figure 2: Results of gridbox analysis for the entire 2014 year. Purple is data in both v3.3.0 and v4.a.1. Red is data new in v4.a.1. Blue is data originally in v3.0.0 but not in v4.a.1

Tables

META THRESH	*All Stations (4,660 stations)*			*1 to 1 Matches with GHCN-D (1,952 stations)*		
	% MERGED	**% UNIQUE**	**% WITH**	**% MERGED**	**% UNIQUE**	**% WITH**
0.50	52.83	1.87	45.30	85.45	0.26	14.29
0.55	52.68	2.42	44.89	85.40	0.26	14.34
0.60	52.60	2.96	44.44	85.19	0.20	14.60
0.75	53.37	6.82	39.81	85.14	0.15	14.70
0.80	54.25	12.88	32.88	83.35	0.82	15.83
0.85	53.28	15.04	31.67	79.25	0.92	19.83
0.90	49.79	16.16	34.06	71.00	1.08	27.92
0.95	31.57	16.42	52.02	44.67	1.33	54.00
0.97	21.09	16.42	62.49	29.10	1.33	69.57
0.99	12.75	16.42	70.84	17.93	1.33	80.74

Table 1: Results of the validation scheme, using GHCN-Daily and observations from the Integrated Surface Dataset with obs taken at 12 UTC. Percentages of merges, uniques, and withhelds are shown for all stations in the validation dataset, along with stations that have a one-to-one match with GHCN-Daily.

Source Info	Number of Stations to be added
UK Stations provided by the Met Office (In ECA&D Monthly?)	300+
German data released by DWD (In ECA&D?)	1000+
EPA's Oregon Crest to Coast Dataset	24
LCA&D: Latin American Climate Assessment & Dataset	148
NCAR Surface Libraries	100+
Stations from meteomet project	240
Libya Stations sent by their NMS	9
C3/EURO4M Stations	80
Stations Digitized by Juerg Luterbacher	10
Homogenized Iranian Data	50
Long-term Swiss Data (In ECA&D?)	7

Table 2: Stations in the queue potentially to be added into Stage 2 of the databank.