Updates since call on 9/27

- Beta1 was released on October 4\textsuperscript{th}
- One comment about station duplicates
  - Ran analysis using Station Name, Lat/Lons, and Year data began
- Added new data up to deadline of 12/1
  - Swedish Data from GCOS Surface Network (GSN)
  - Uruguay Data (thanks to Madeline Renom)
  - Norwegian data from eKlima Portal
  - Environment Canada (replaces homogenized source from beta1)
  - HadISD (replaces GSOD source from beta1)
  - Updated CRUTEM4 dataset (thanks to Colin Morice)
  - Updated data from University of Giessen (thanks to Juerg Luterbacher)
Updates since call on 9/27 (con’t)

- Changes to merge program
  - Added a metadata comparison of when the data record began
  - Updated lookup table used to determine if candidate station is merged, unique, or withheld after a data comparison is made
1) Read in target dataset
2) Read in candidate dataset
3) Calculate Metadata comparisons
   a. Geographic distance
   b. Height distance
   c. Jaccard Index
   d. Year Data Record Began
4) Calculate Data Comparisons
   a. Overlapping Data: Index of Agreement
   b. Non-Overlapping Data: Rely on metadata
5) Fate of candidate station is one of the following:
   a. Merge with the target station
   b. Station is unique and added to the target dataset
   c. Station has not enough information and withheld
Validation of Merge Program

• Using an independent dataset, originally from ISD with known time of observation biases
• GHCN-Daily used as the target dataset
• Independent Dataset used as candidate dataset
• 1,251 stations that should match with GHCN-Daily

**Beta1**
- # merged: 1108 # total: 1251 %: 88.57
  - # matched correctly = 1026 total merged= 1108 % = 92.60
- # unique: 33 # total: 1251 %: 2.64
- # withheld: 110 # total: 1251 %: 8.79

**Beta2**
- # merged: 1139 # total: 1251 %: 91.05
  - # matched correctly = 1053 total merged= 1139 % = 92.45
- # unique: 10 # total: 1251 %: 0.80
- # withheld: 102 # total: 1251 %: 8.15
RESULTS OF RECOMMENDED MERGE (beta2)
Number of Stations

BLACK=GHCN-M V3 | BLUE=Recommended (beta1) | RED=Recommended (beta2)
Number of Gridboxes (Globe)

BLACK=GHCN-M V3 | BLUE=Recommended (beta1) | RED=Recommended (beta2)

% Land Area Sampled vs. Years

Years:
- 1860
- 1890
- 1920
- 1950
- 1980
- 2010
Number of Gridboxes (NH)

BLACK=GHCN-M V3 | BLUE=Recommended (beta1) | RED=Recommended (beta2)
Number of Gridboxes (SH)

BLACK=GHCN-M V3 | BLUE=Recommended (beta1) | RED=Recommended (beta2)
Updates since call on 9/27 (con’t)

• Changes to variants
  ▪ Code changes from recommended merge were applied to variants
  ▪ Variants name changed
<table>
<thead>
<tr>
<th>Variant</th>
<th>Source Deck Change?</th>
<th>Threshold Change?</th>
<th>Code Change?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (colin)</td>
<td>Prioritize sources from NMA’s.</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>2 (david)</td>
<td>NMA’s with TMAX and TMIN given highest priority</td>
<td>overlap_threshold changed from 60 months to 24 months</td>
<td>no</td>
</tr>
<tr>
<td>3 (peter)</td>
<td>No TAVG sources used, rest ranked by order of longest station record present</td>
<td>Thresholds to merge and unique station are lowered to merge more stations</td>
<td>metadata_probability weighted to favor distance_probability over all others</td>
</tr>
<tr>
<td>4 (jay)</td>
<td>no</td>
<td>no</td>
<td>During data comparisons, candidate station only merged or unique</td>
</tr>
<tr>
<td>5 (matt)</td>
<td>All homogenized sources removed</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>6 (more-unique)</td>
<td>no</td>
<td>All thresholds adjusted to make more candidate stations unique</td>
<td>no</td>
</tr>
<tr>
<td>7 (more_merged)</td>
<td>no</td>
<td>All thresholds adjusted to make more candidate stations merge with target stations</td>
<td>no</td>
</tr>
</tbody>
</table>
## Variants of Merge Program

- **Number of Stations**

<table>
<thead>
<tr>
<th>Variant</th>
<th>beta1</th>
<th>beta2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommended</td>
<td>37,322</td>
<td>34,630</td>
</tr>
<tr>
<td>Variant 1 (colin)</td>
<td>27,412</td>
<td>19,765</td>
</tr>
<tr>
<td>Variant 2 (david)</td>
<td>29,110</td>
<td>20,515</td>
</tr>
<tr>
<td>Variant 3 (peter)</td>
<td>34,492</td>
<td>31,523</td>
</tr>
<tr>
<td>Variant 4 (jay)</td>
<td>42,374</td>
<td>35,289</td>
</tr>
<tr>
<td>Variant 5 (matt)</td>
<td>39,040</td>
<td>34,345</td>
</tr>
<tr>
<td>Variant 6 (more-unique)</td>
<td>47,331</td>
<td>43,292</td>
</tr>
<tr>
<td>Variant 7 (more-merged)</td>
<td>35,714</td>
<td>31,487</td>
</tr>
</tbody>
</table>
Number of Stations

BLACK=GHCN-M V3 | BLUE=Recommended (beta2) | RED=Variants (beta2)
Number of Gridboxes (Globe)

BLACK=GHCN-M V3 | BLUE=Recommended (beta2) | RED=Variants (beta2)

% Land Area Sampled

Years

1860 1890 1920 1950 1980 2010
Variants of Merge Program

- Trends (1850-2010)
  - GHCN-M v3 QCU: 0.543

<table>
<thead>
<tr>
<th>Variant</th>
<th>beta1</th>
<th>beta2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommended</td>
<td>0.653</td>
<td>0.638</td>
</tr>
<tr>
<td>Variant 1 (colin)</td>
<td>0.659</td>
<td>0.664</td>
</tr>
<tr>
<td>Variant 2 (david)</td>
<td>0.654</td>
<td>0.663</td>
</tr>
<tr>
<td>Variant 3 (peter)</td>
<td>0.822</td>
<td>0.798</td>
</tr>
<tr>
<td>Variant 4 (jay)</td>
<td>0.637</td>
<td>0.637</td>
</tr>
<tr>
<td>Variant 5 (matt)</td>
<td>0.652</td>
<td>0.632</td>
</tr>
<tr>
<td>Variant 6 (more-unique)</td>
<td>0.698</td>
<td>0.667</td>
</tr>
<tr>
<td>Variant 7 (more-merged)</td>
<td>0.620</td>
<td>0.616</td>
</tr>
</tbody>
</table>
To Do List prior to v1.0.0 release

• Continue to assess any public feedback (if any)

• Changes to Stage2 / Stage3 data (nothing outstanding)
  - porting the stage 2 provenance flags into stage 3
  - adding a stage 2 provenance flag describing the number of days used to create a monthly average
  - creating multiple formats for the stage 3 datasets, including CF compliant NetCDF
  - create a blacklist file of stations that are known to be incorrect, so they automatically end up in the withheld bin

• Send draft of paper to Geoscience Data Journal
  - Draft was sent to working group last week