Databank Merge Update for Wednesday, June 6th

Some key highlights about the merge program ***during our last meeting*** included the following:

- Outlined the decisions made to determine metadata criteria, including the geographic distance, height distance, and station name similarity using the Jaccard Index
- Overviewed methods for data comparisons between a master station and a candidate station, including the use of two different equations: Normalized Root Mean Square Deviation (NRMSD) and the Index of Agreement (IA)
- Showed some preliminary results using TMAX data only

Since then we have made numerous updates, including:

- Adding 7 new sources, totaling 43 (see table on next page)
    - Antarctic Data provided by AWS/AMC
    - Greenland / WMSSC data provided by NCAR
    - Argentina data provided by the University of Buenos Aires
- Overlap Comparisons: Moving forward with the Index of Agreement
- Non-Overlap Comparisons: Decided to make decision based solely upon metadata
- Adding TMIN and TAVG into the merge process
- Validation: Using a pseudo source of stations with a known time of observation bias

A quick run through of the how the program runs is below:

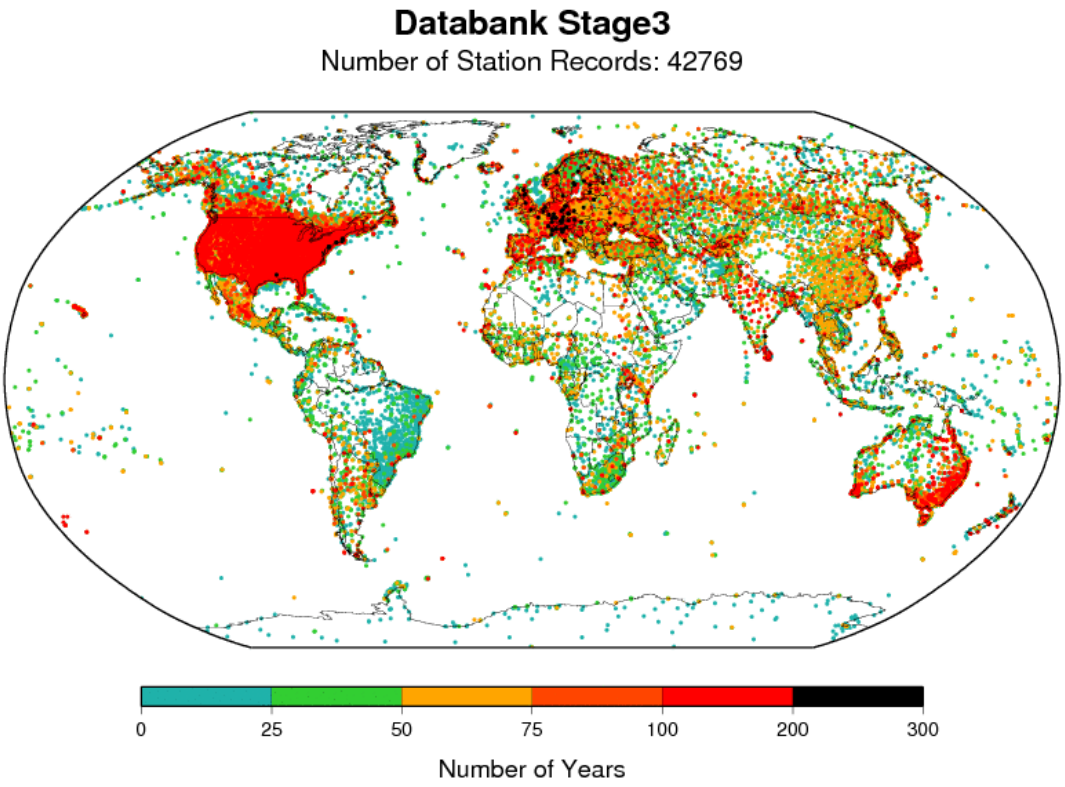
1) Read in master dataset
2) Read in candidate dataset
3) Calculate Metadata comparisons
    a. Geographic distance
    b. Height distance
    c. Jaccard Index
4) Calculate Data Comparisons
    a. Overlapping Data: Index of Agreement
    b. Non-Overlapping Data: Rely on metadata
5) Fate of candidate station is one of the following:
    a. Merge with the master station
    b. Station is unique and added to the master dataset
    c. Station has not enough information and withheld

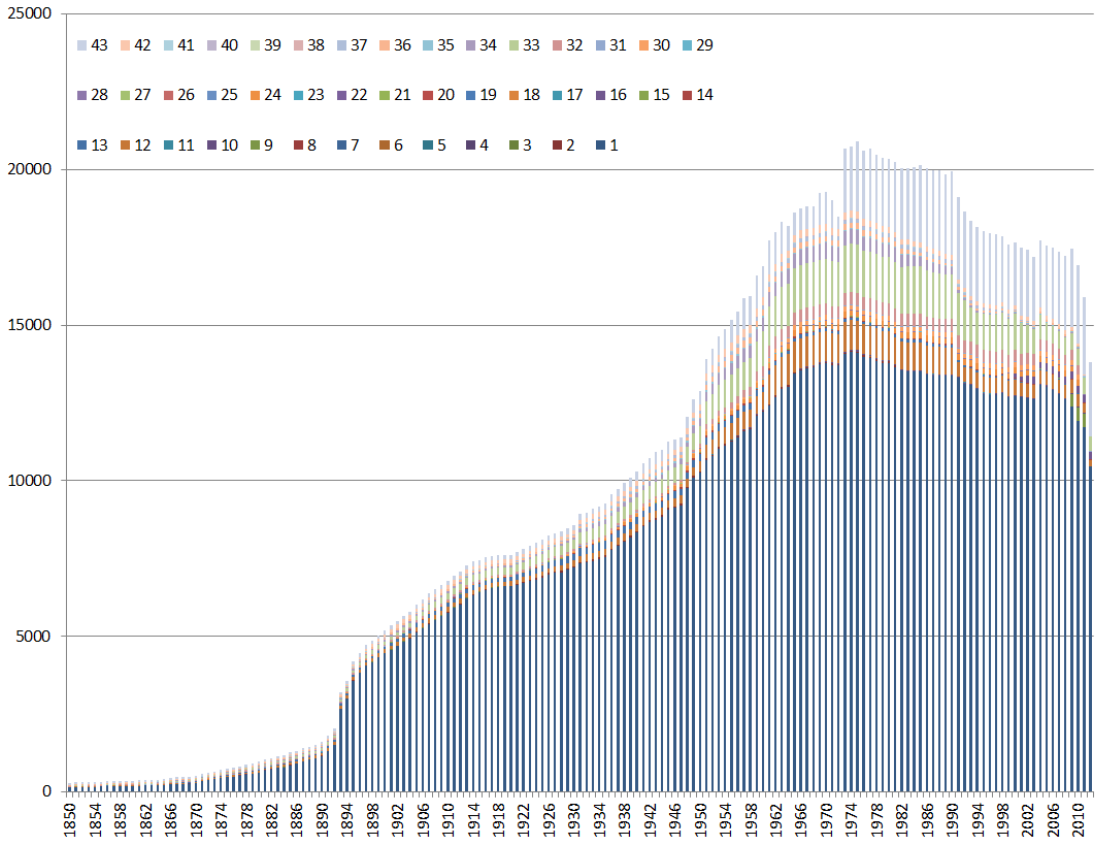
Steps 2 through 5 are run iteratively through all the sources which have TMAX/TMIN first. Afterwards, the master dataset creates TAVG from TMAX and TMIN (added together divided by 2). The updated master dataset then runs through all the sources that have TAVG. The final result is the Stage 3, merged dataset

There are 8 different thresholds throughout the program that will determine whether a station is merged, unique, or withheld. Slight changes in these thresholds may change the overall results. (More on this later)

Priority	Name	Source	TMAX?	TMIN?	TAVG?
1	GHCN-Daily Raw	NCDC	Y	Y	N
2	Mexico	CDMP	Y	Y	N
3	Vietnam	CDMP	Y	Y	N
4	US Forts	CDMP	Y	Y	N
5	Channel Islands	States of Jersey Met	Y	Y	N
6	Ecuador	Inst. Nacional De Met E Hidrologia	Y	Y	N
7	Pitcairn Island	Met Service of New Zealand	Y	Y	N
8	Beirut	University of Giessen	Y	Y	Y
9	Brazil	INPE, Nat. Institute for Space Research	Y	Y	N
10	Argentina	University of Buenos Aires	Y	Y	N
11	Greenland	NCAR	Y	Y	N
12	World Weather Records	WMO	Y	Y	Y
13	Colonial Era Archives	Griffith	Y	Y	N
14	East Africa	Univ. of Alabama Huntsville	Y	Y	Y
15	Uganda	Univ. of Alabama Huntsville	Y	Y	Y
16	UK Climat	UKMO	Y	Y	Y
17	Antarctica South Pole	Univ. of Wisc.-Madison	Y	Y	Y
18	Switzerland	ISPD	N	N	Y
19	Polar	ISPD	N	N	Y
20	Sydney	ISPD	N	N	Y
21	Antarctica	SCAR Reader Project	N	N	Y
22	Mon. Clim Data of World (MCDW)	NCDC	N	N	Y
23	Spain	Univ. Rovira I Virgili	Y	Y	Y
24	Russia	Roshydromet	Y	Y	Y
25	Uruguay	Inst. Nacional de Invest Agropecuaria	Y	Y	Y
26	Switzerland	Digihom/MetoSwiss/IAC-ETH	Y	Y	Y
27	Tunisia/Morocco	ISPD	Y	Y	Y
28	Europe / N. Africa	ECA Daily / KNMI	Y	Y	Y
29	Southeast Asia	SACA / KNMI	Y	Y	Y
30	Japan	JMA	Y	Y	Y
31	Uk Met Office Historical	UKMO	Y	Y	N
32	Europe / N. Africa	ECA Monthly / KNMI	Y	Y	Y
33	Max/Min Stations from R. Vose	NCDC	Y	Y	N
34	GHCN-M v2 Source	NCDC	N	N	Y
35	Mon. Surf. Station Clim. (WMSSC)	NCAR	N	N	Y
36	GHCN-M v2	NCDC	Y	Y	Y
37	Central Asia	NSIDC	Y	Y	Y
38	Canada	Environment Canada	Y	Y	Y
39	Australia	BOM	Y	Y	Y
40	Arctic	IARC/Univ of Alaska Fairbanks	N	N	Y
41	Greater Alpine Region	Histalp / ZAMG	N	N	Y
42	CRUTEM4	UKMO	N	N	Y
43	Global Summary of the Day	NCDC	Y	Y	N

Here are the results from the latest version of the merge:



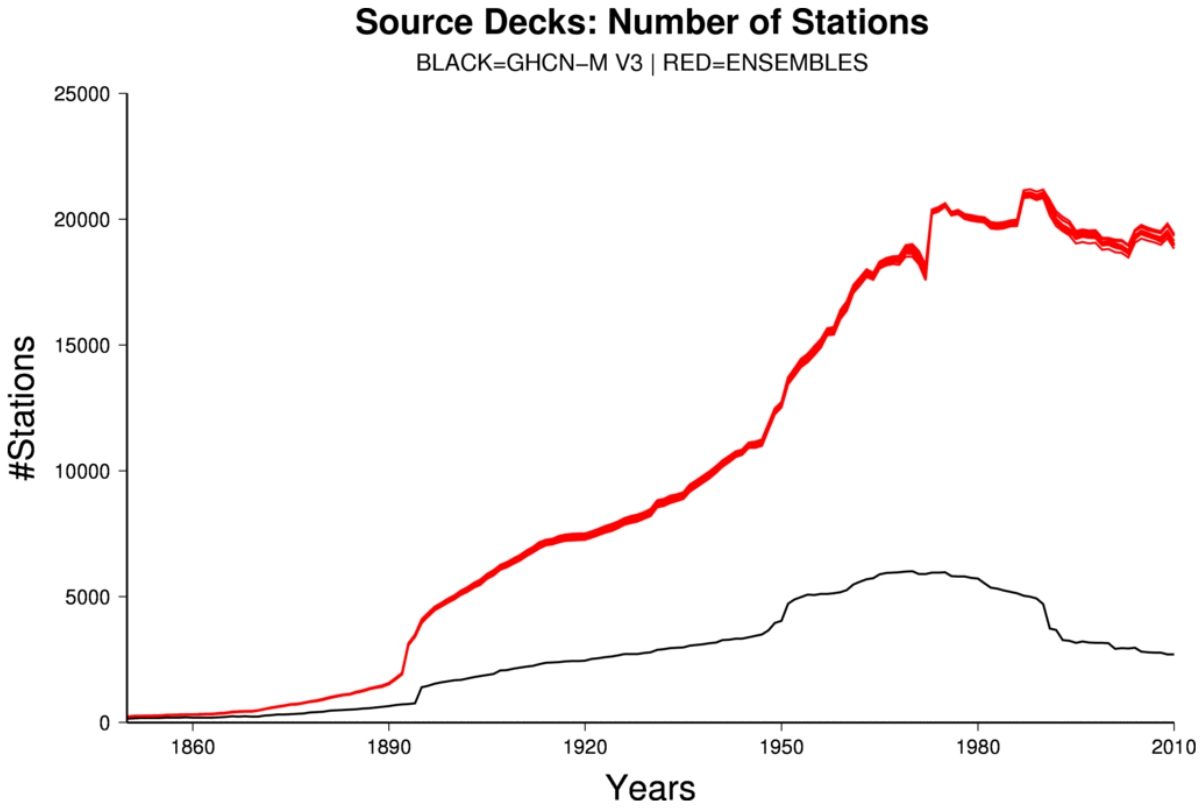


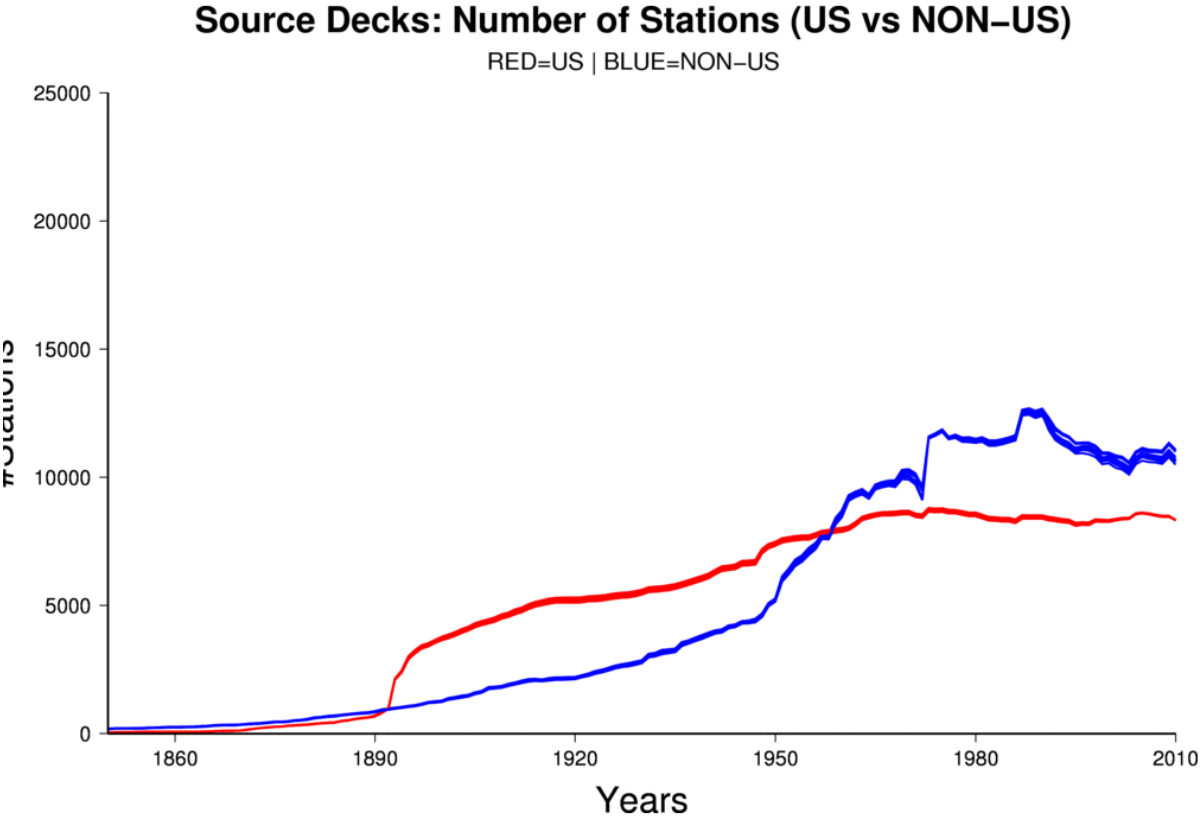
Currently, our efforts have focused on the sensitivity of the databank merge program. To do this we have created some ensembles of the merge program. We have done this one of two ways. First, making no changes to the code, but changing the source hierarchy of the 43 data sources. Some have been re-ordered randomly, others have been ordered subjectively. Results using this method are below:

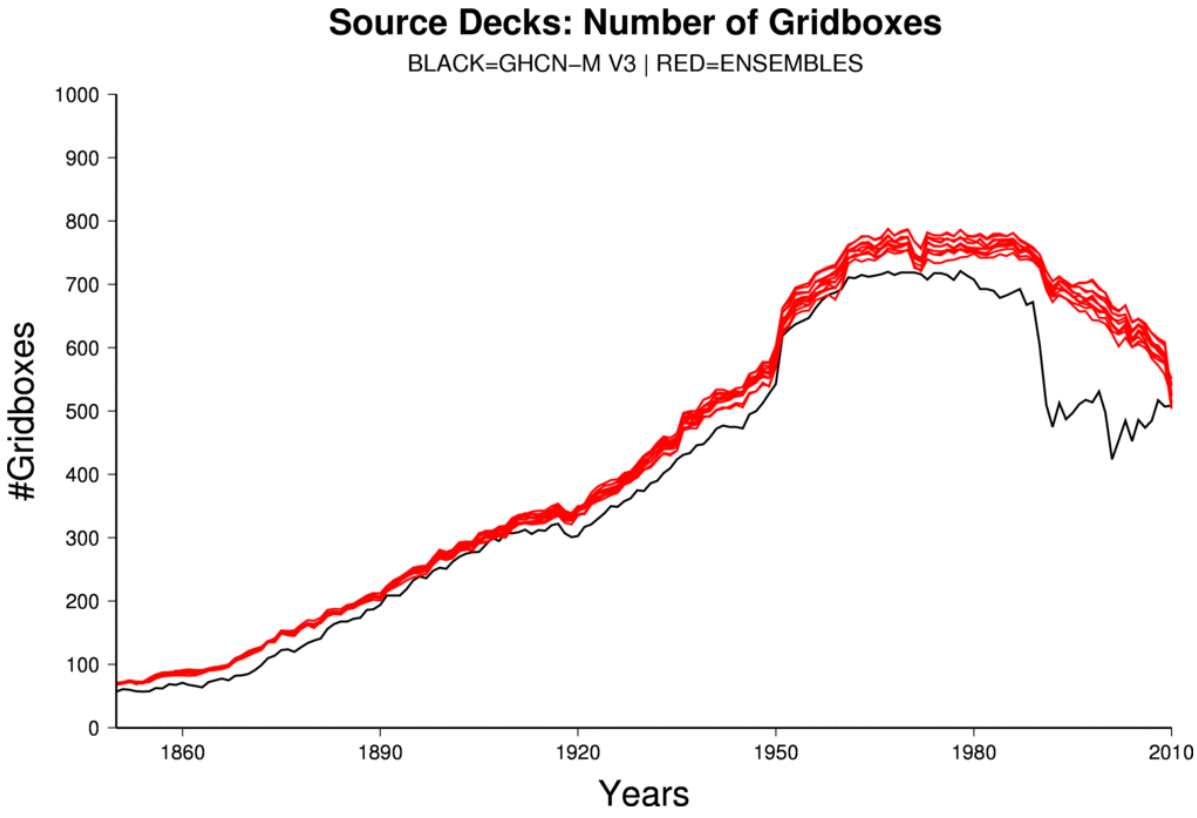
SOURCE DECK	# STNS	TREND °C/Century (1901-2010)	TREND °C/Century (1951-2010)
DEFAULT	**42769**	**0.844**	**1.829**
1	42623	0.902	1.876
2	42756	0.898	1.832
3	42525	0.863	1.865
4	42531	0.873	1.861
5	42522	0.855	1.809
6	42642	0.881	1.847
7	42606	0.913	1.888
8	42648	0.878	1.849
9	42649	0.884	1.875
10	42730	0.871	1.841
11	42749	0.871	1.814
AVERAGE	42646	0.878	1.849
STDEV	91	0.020	0.025
GHCNM-V3	7280	0.771	1.644

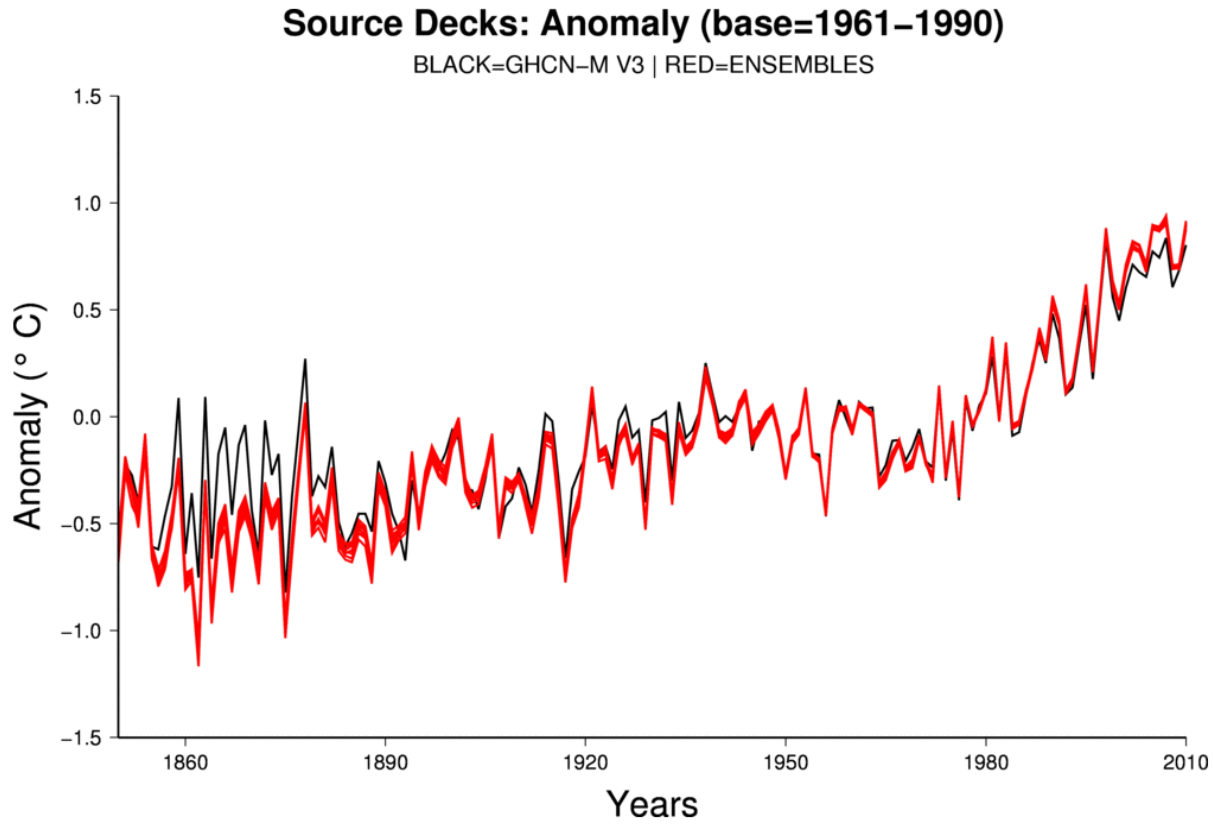
The following three plots depict the following with the ***source deck ensembles***. Data was quality controlled using the current algorithm used to create GHCN-M V3 before anomalies were generated

1. Number of stations over time (1850-2011)
2. Number of stations over time, stratified by US vs Non-US (1850-2011)
3. Number of 5°X5° Grids with available data over time (1850-2011)
4. Anomaly using a 1961-1990 base period (1850-2011)


Source Decks: Number of Stations
BLACK=GHCN-M V3 | RED=ENSEMBLES
#Stations
25000
20000
15000
10000
5000
0
1860
1890
1920
1950
1980
2010
Years


Source Decks: Number of Stations (US vs NON-US)
RED=US | BLUE=NON-US
25000
20000
15000
10000
5000
0
1860
1890
1920
1950
1980
2010
Years


Source Decks: Number of Gridboxes
BLACK=GHCN−M V3 | RED=ENSEMBLES
#Gridboxes
1000
900
800
700
600
500
400
300
200
100
0
1860
1890
1920
1950
1980
2010
Years


Source Decks: Anomaly (base=1961−1990)
BLACK=GHCN−M V3 | RED=ENSEMBLES
Anomaly (° C)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
1860
1890
1920
1950
1980
2010
Years

The other method is by changing the user-defined thresholds that go into the databank. There are currently 8 different thresholds that determine whether a candidate station is merged, unique, or withheld. A brief description is below:

metadata_threshold: the first metadata threshold that takes into account the distance, height, and jaccard probabilities (default is 0.50)
metadata_threshold2: the second metadata threshold used if there is no overlap period between the master and candidate station (higher than the first metadata threshold) (default is 0.85)
posterior_threshold_same_txn: threshold where TMAX/TMIN candidate station has to exceed in order to merge with the master station (default is 0.50)
posterior_threshold_unique_txn: threshold where TMAX/TMIN candidate station has to exceed in order to be considered a unique station (default is 1.30)
posterior_threshold_same_txn: threshold where TAVG candidate station has to exceed in order to merge with the master station (default is 0.50)
posterior_threshold_unique_txn: threshold where TAVG candidate station has to exceed in order to be considered a unique station (default is 0.90)
overlap_threshold: overlap period that must exist between the master and candidate station in order to calculate a data comparison via the Index of Agreement (default is 60 months)
gap_threshold: gap period that must exist when merging a candidate station with the master station (default is 60 months)

For each threshold, different ensembles are generated to show the sensitivity of the merge just by changing one threshold and nothing else. Results of some of the thresholds are below (dark red is the default)

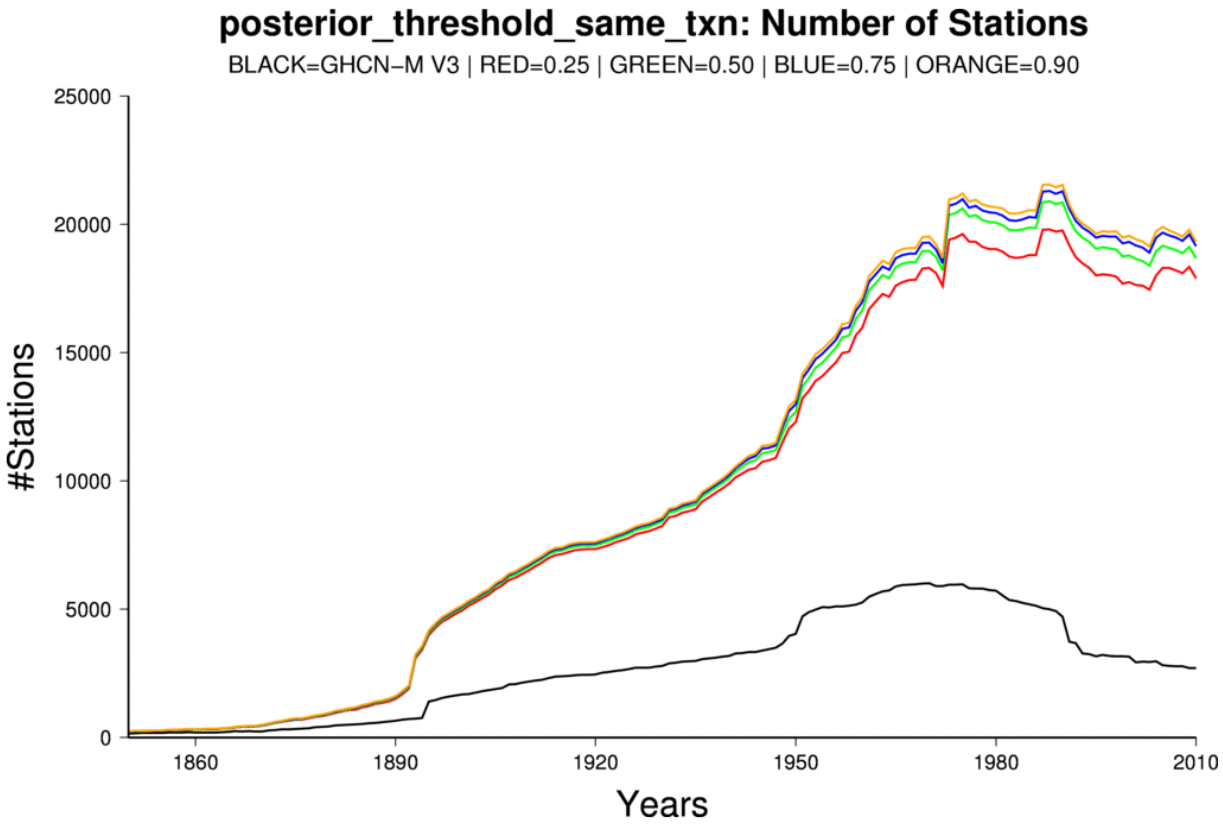
posterior	**THRESHOLD**	**# STNS**	**TREND °C/Century (1901-2010)**	**TREND °C/Century (1951-2010)**
threshold	0.25	40565	0.851	1.859
same	**0.50**	**42769**	**0.844**	**1.829**
txn	0.75	43425	0.867	1.837
	0.90	44042	0.863	1.833
	AVERAGE	42700	0.856	1.840
	STDEV	1515	0.011	0.013
	GHCNM-V3	7280	0.771	1.644

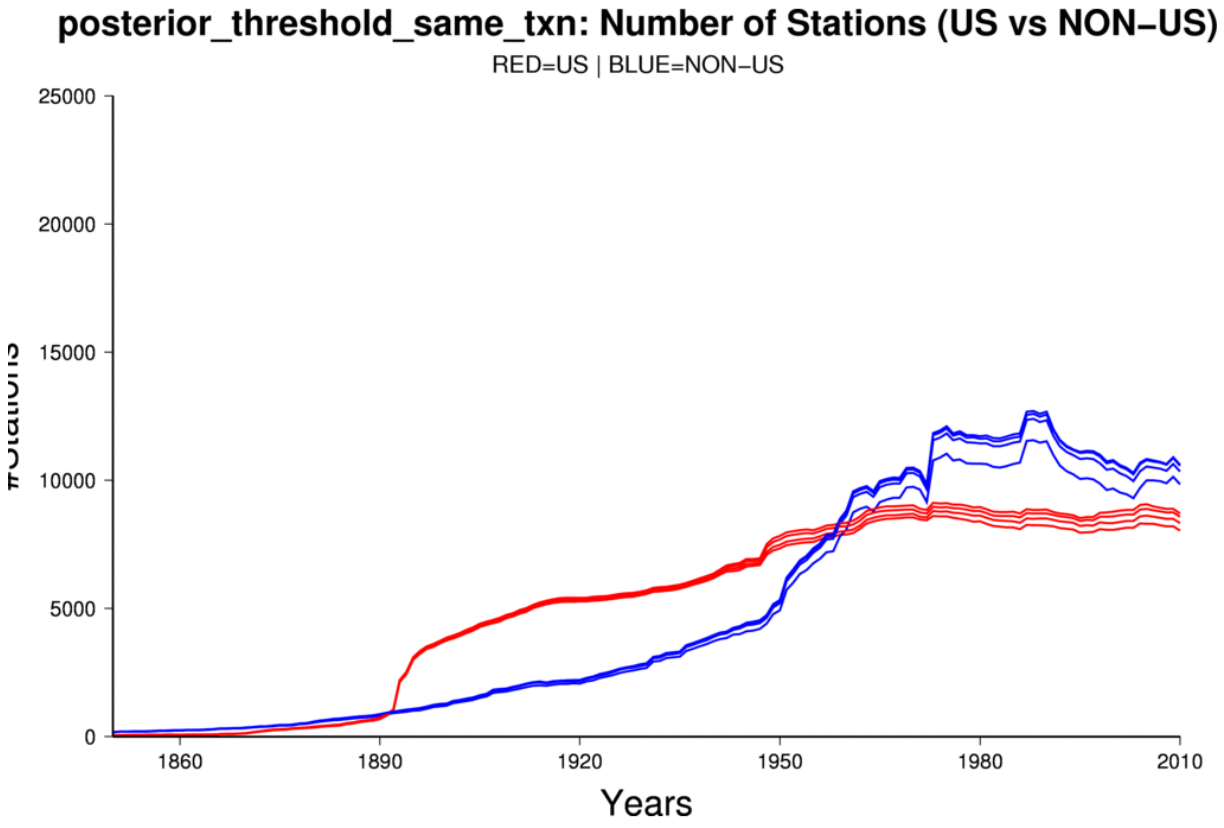
posterior	**THRESHOLD**	**# STNS**	**TREND °C/Century (1901-2010)**	**TREND °C/Century (1951-2010)**
threshold	0.75	43317	0.844	1.816
unique	**1.30**	**42769**	**0.844**	**1.829**
txn	1.75	42144	0.841	1.831
	2.30	41072	0.845	1.847
	AVERAGE	42326	0.844	1.831
	STDEV	963	0.002	0.013
	GHCNM-V3	7280	0.771	1.644

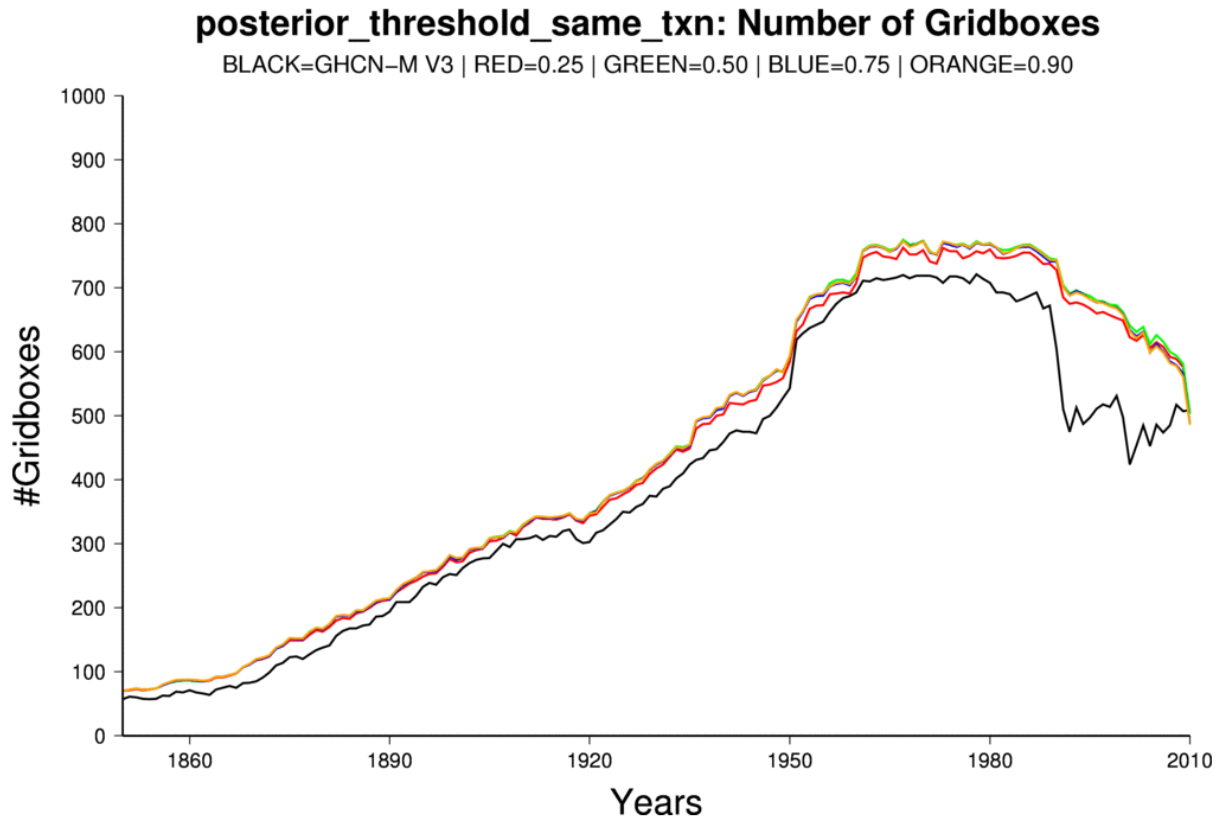
overlap threshold	THRESHOLD	# STNS	TREND °C/Century (1901-2010)	TREND °C/Century (1951-2010)
	12	44193	0.802	1.832
	60	**42769**	**0.844**	**1.829**
	120	41682	0.849	1.835
	180	41236	0.859	1.843
	AVERAGE	42470	0.839	1.835
	STDEV	1317	0.025	0.006
	GHCNM-V3	7280	0.771	1.644

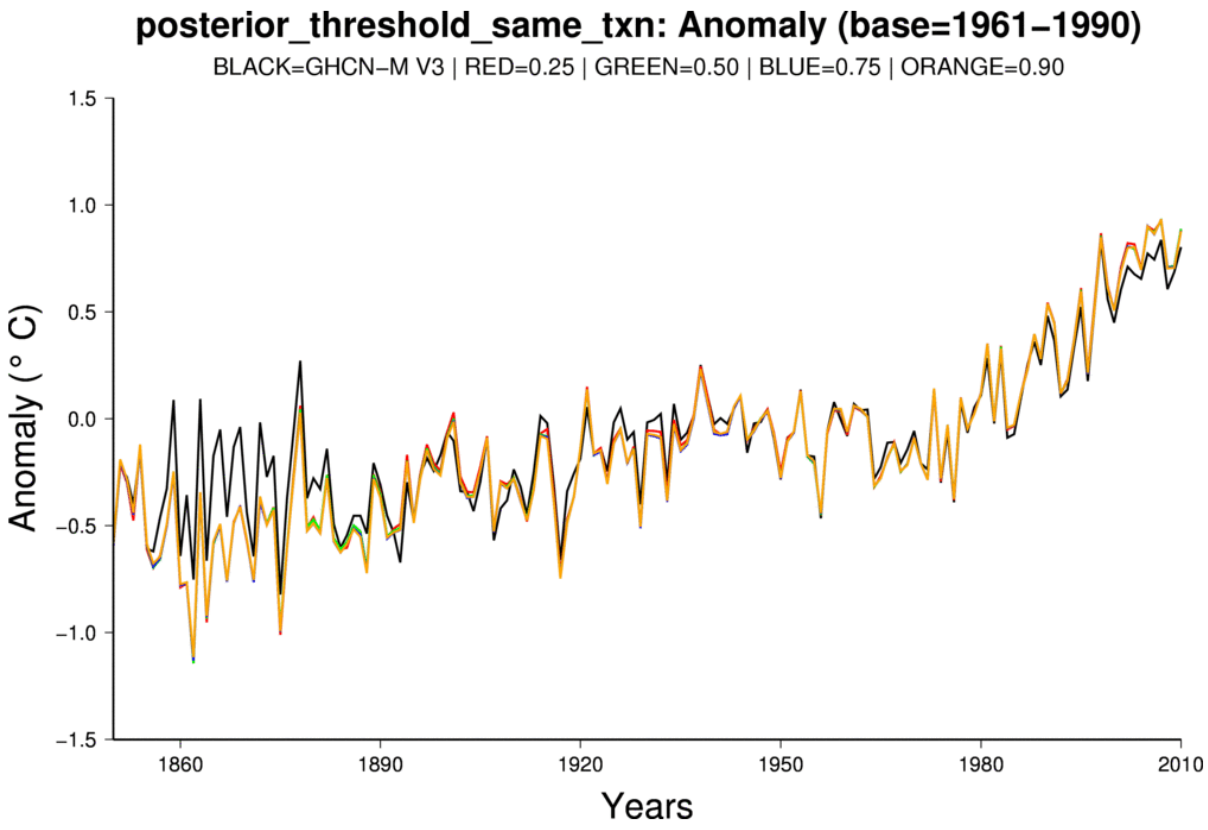
The following three plots depict the following with the ***posterior_threshold_same_txn ensembles***. Data was quality controlled using the current algorithm used to create GHCN-M V3 before anomalies were generated

1. Number of stations over time (1850-2011)
2. Number of stations over time, stratified by US vs Non-US (1850-2011)
3. Number of 5°X5° Grids with available data over time (1850-2011)
4. Anomaly using a 1961-1990 base period (1850-2011)




posterior_threshold_same_txn: Number of Stations (US vs NON−US)
RED=US | BLUE=NON−US
#Stations
25000
20000
15000
10000
5000
0
1860
1890
1920
1950
1980
2010
Years


posterior_threshold_same_txn: Number of Gridboxes
BLACK=GHCN−M V3 | RED=0.25 | GREEN=0.50 | BLUE=0.75 | ORANGE=0.90
#Gridboxes
1000
900
800
700
600
500
400
300
200
100
0
1860
1890
1920
1950
1980
2010
Years


posterior_threshold_same_txn: Anomaly (base=1961–1990)
BLACK=GHCN–M V3 | RED=0.25 | GREEN=0.50 | BLUE=0.75 | ORANGE=0.90
Anomaly (° C)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
1860
1890
1920
1950
1980
2010
Years