GHCN-M Technical Report
No. GHCNM-15-01

Modifications to GHCN-Monthly (version 3.3.0) and USHCN (version 2.5.5) processing systems

Byron Gleason
Claude Williams
Matt Menne
Jay Lawrimore
NOAA's National Centers for Environmental Information
Asheville, NC

June 9, 2015

This report, GHCNM-15-01, provides descriptions of software modifications made and implemented in GHCN-Monthly version 3.3.0 and USHCN version 2.5.5. *Note: An earlier version of this document dated April 9, 2015 incorrectly annotated the USHCN version upgrade as 2.6.0.*

U.S. DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
National Centers for Environmental Information
Asheville, NC 28801-5001

Table of Contents

List of Acronyms

GHCN-M: Global Historical Climatology Network-Monthly
NCEI: National Centers for Environmental Information
NOAA: National Oceanic and Atmospheric Administration
PHA: Pairwise Homogeneity Adjustment
USHCN: U.S. Historical Climatology Network

Abstract

The software used to perform operational updates and reprocessing of GHCN-M version 3 and USHCN version 2 was modified to incorporate updated data sources and to correct software bugs associated with the quality control and bias correction processes. The steps taken to correct each problem area are described in this technical report. A comparison of global analyses performed using the software before and after the corrections and modifications is included. The impact on annual means resulted in little change in global annual land surface temperature rankings. The change in the century-scale global land surface air temperature trend (1880-2014) is 0.03°C/century. The software modifications are incorporated into a new release of GHCN-M, version 3.3.0 and USHCN version 2.5.5.

1. Introduction

The Global Historical Climatology Network-Monthly (GHCN-M) version 3.0.0 dataset was released to the public in May 2011. Subsequent modifications were implemented to improve processing speed through improvements to processing data within sparse matrices. This change was implemented as version 3.1.0 and released in November 2011 (Technical Report GHCNM-12-01R). Other minor updates were performed in August 2012 to correct software coding errors (bugs) in the Pairwise Homogenization Algorithm (PHA; Menne and Williams 2009) that had reduced the algorithm's efficiency in identifying inhomogeneities in the GHCN monthly temperature series. These changes are described in Technical Report GHCNM-12-02. In this technical report additional improvements to the GHCN-M processing system are described as part of the release of version 3.3.0. In addition because the USHCN data set is an important part of GHCN-M changes associated with the upgrade from USHCN v2.5.0 to v2.5.5 are also described.

The GHCN-M processing system consists of two major phases; Phase 1) data collection, integration and quality control; Phase 2) bias correction and data output. Changes were incorporated into both phases in the updated version described herein (v3.3.0). Comparisons against the previous version (v3.2.2) are described throughout this document.

Phase 1 changes include data source updates and a correction to two small software bugs in the quality control process. The data source update involved the addition of Monthly Climatic Data of the World (MCDW) data for 2012 and 2013 and the addition of previously missing

CLIMAT data provided by the UK Hadley Centre for a three month period in 2011. These are described in section 2a. Changes made to the GHCN-M quality control process to correct software bugs are described in section 2b. Software coding errors recently found in the Phase 2 bias correction process and corrections to the algorithms are described in Section 2c.

2. Data Source Updates and Software Modifications

a. Data source updates

The GHCN-M version 3 data set is comprised of ten sources of monthly mean temperature data as shown in Table 1. Two of these sources were recently updated, and their data was refreshed in version 3.3.0. In the first instance summary of the month data collected and processed as part of the Monthly Climatic Data of the World (MCDW) data set were incorporated. This consisted of updating the MCDW source of data in GHCN-M to incorporate data that had undergone final MCDW processing. This resulted in no additional data or changes to data in GHCN-M. All values had been previously received via preliminary and intermediate sources. This update simply resulted in a change in the source flag for some monthly values from either ‘P’ or ‘C’ to ‘M’ (Table 1).

The second data source update involved a secondary source of CLIMAT data. Among the sources used in GHCN-M is a set of CLIMAT data provided by the UK Hadley Centre. Version 3 of this UK data set was included in GHCN-M version 3.2.2. The Hadley Centre recently released version 4 which included additional data from Australia that had been missing in the previous

version. In particular three months from September-November 2011 were filled with observations and are now included in GHCN-M version 3.3.0.

b. Quality Control software changes

The GHCN-M quality control process was designed to be objective, reproducible, traceable, and applied consistently throughout the data set. The process for version 3 mean temperature begins with three basic integrity checks followed by one outlier and one spatial consistency check. These checks are month-over-month duplicate, yearly duplicate, isolated value, climatological outlier, and spatial outlier (Lawrimore et al. 2011). Once an observation fails a quality control check, the value is excluded from subsequent checks during that processing cycle.

In version 3.3.0 minor modifications were made to the month-over-month duplicate check and the spatial outlier check to correct software bugs. The month-over-month duplicate check identifies errors resulting from a problem that can occur in the transmission of CLIMAT bulletins over the GTS; the retransmission and incorrect labeling of data that results in the mean temperature for the current data month being repeated from the prior month. In version 3.2.2 a coding error resulted in the monthly duplicate check not being applied to stations in Algeria. Fixing this error resulted in an additional 47 Algerian observations across 31 stations being flagged as invalid; about 0.00089% of the data in GHCN-M.

The second change involved a fix to a software bug in the spatial outlier check. For observations that are less than 5 sigma, but more than 2.5 sigma from the station's biweight mean

temperature, a comparison with neighbors is used to assess its validity. This check ensures extreme temperatures are flagged as invalid if they cannot be confirmed by similar observations at neighboring stations. The check is designed to confirm the validity of all observations that are -2.5 to -5.0 bi-weight standard deviations (unusually cold outliers) or +2.5 to +5.0 bi-weight standard deviations (unusually warm outliers) from their bi-weight mean. In version 3.2.2, this particular check was properly working only for values between +2.5 and +5.0 bi-weight standard deviations from the bi-weight mean (unusually warm outliers). This resulted in erroneously cold temperatures appearing as valid in v3.2.2. The fix to this software bug means that erroneous cool outliers will now be appropriately flagged. It affected 0.18% of the observations in GHCN-M.

c. Changes to bias correction algorithms

The nature of the homogeneity adjustments made to remove non-climatic influences that can bias the GHCN-M temperature record are described in Lawrimore et al. 2011 for GHCN-M version 3. In brief, adjustments are necessary because surface weather stations are frequently subject to minor relocations throughout their history of operation and may also undergo changes in instrumentation as measurement technology evolves. Furthermore, observing practices may vary through time, and the land use/land cover in the vicinity of an observing site can be altered by either natural or man-made causes. Any such modifications to the circumstances behind temperature measurements have the potential to alter a thermometer's microclimate exposure characteristics or otherwise change the bias of measurements relative to those taken under

previous circumstances. The manifestation of such changes is often an abrupt shift in the mean level of temperature readings that is unrelated to true climate variations and trends. Ultimately, these artifacts (also known as inhomogeneities) confound attempts to quantify climate variability and change because the magnitude of the artifact can be as large as or larger than the true background climate signal. The process of removing the impact of non-climatic changes in climate series is called homogenization.

In version 3 of the GHCN-M temperature data set, the apparent impacts of documented and undocumented inhomogeneities are detected and corrected through automated pairwise comparisons of mean monthly temperature series as detailed in *Menne and Williams* [2009]. This method improved upon earlier methods in part because it did not rely on the creation of a homogenous composite reference series from neighboring stations [*Menne and Williams,* 2005].

In version 3.3.0 five changes were made to the Phase 2 bias correction process. Two of these affected only data flags and three affected data values.

<u>Software changes that affect flags only</u>

1) The US Historical Climatology Network data set version 2.5.0 (Menne et al. 2009) is one important source of data for GHCN-M version 3. This data set consists of 1218 stations in the contiguous U.S. with data from 1895 to present. Approximately 5% of the data in USHCN version 2.5.0 originate from version 1 of USHCN (Karl et al. 1986). In USHCN v2.5.0 the days missing flags for USHCNv1 stations were presented in upper case. In

USHCN v2.5.5 and GHCN-M v3.3.0 the days missing flag for these stations is changed to lower case.

2) The second flag change is associated with output from the Pairwise Homogeneity Algorithm (PHA) process. When inhomogeneities are detected during the PHA process, either adjustments are made or the inhomogenous data are removed. In cases where there is not sufficient information to calculate an accurate adjustment (because of a lack of well correlated neighbors with homogeneous observations around the time of a target station change or because the time between inhomogeneities is too short to adjust), the inhomogeneous data are removed (i.e., set to missing). In such cases the months that are removed are flagged with an 'X' in the adjusted data set. The original observations remain available in the unadjusted data set. In the previous version the data considered to be unadjustable by PHA were set to missing and identified with the 'X' flag except in cases where the most recent data were considered unadjustable. In the USHCN v2.5.0, estimates were provided for the data removed by the PHA and appropriately identified as estimated. However these estimates were not distinguished from estimates provided for dates that were absent from the original observational record. In USHCN v2.5.5 appropriate flagging is now applied as follows: 'E' – absent from the original record and estimated, 'X' - removed by the PHA as unadjustable and estimated, 'Q' - removed after failing quality control checks and estimated.

Software bug fixes that affect data

3) The first bug fix that resulted in changes to GHCN-M adjusted data was associated with the calculation of correlation coefficients for nearest neighbors. As part of the bias correction process the 100 nearest neighbors are identified for each target station in GHCN-M and correlation coefficients calculated. In the previous version, arrays were not properly reinitialized to missing before neighbors for the next target station were processed. For most stations this was not problematic because there are at least 100 new neighbors available for the target station. But for some extremely remote stations there are not 100 neighbors in GHCN-M version 3. This resulted in neighbors from the previous target station remaining in the array and being taken as neighbors to the current target station. This problem affected 12 South Pacific and 1 Antarctic station (Fig. 1a). The solution fixed initialization of the arrays containing nearest neighbors for a given station before continuing to the next station.

4) Composite station changes: The source of this error was the unintentional reuse of stations in the process of identifying inhomogeneities in the PHA process. As mentioned above there are 1218 stations in GHCN-M that originate from the USHCN data set. Of the USHCN station network, 208 stations are comprised of threaded station$_s$; i.e., two or more nearby stations combined to form a single station (Menne et al. 2009). In the GHCN-M version 3.2.2 process, neighbors of the threaded USHCN stations used to identify inhomogeneities in the PHA process unintentionally included stations that had

been previously used to produce the threaded USHCN station record. In addition, the intent was to use only non-USHCN sites as neighbors in the PHA to homogenize USHCN stations. However, 46 USHCN stations were inadvertently used as neighbors of other USHCN stations. . In version 3.3.0 no stations used in producing the USHCN data records are used as neighbors of the USHCN (Fig. 1b).

5) Treatment of data that are not homogenized by the PHA because of long gaps or sparse neighbor network of stations (Long Gap Fixes): As stated previously the PHA algorithm relies on well correlated neighbors for identifying inhomogeneities in the target station. The well correlated neighbors are also used in determining the bias correction to apply to the data preceding the breakpoint. In some cases it is not possible to calculate the correct adjustment to apply because the network of neighboring stations is too sparse or because there are a very large number of years of missing data in the target station's period of record (i.e., a large data gap) and there are not a sufficient number of homogenous neighboring stations to span the long gap. In such cases the target station's data before the point where the inhomogeneity occurred are removed. Examples of this are shown in Figure 2. In version 3.2.2 the years that could not be adjusted were not being removed as intended. The software fix in version 3.3.0 now removes the segments that could not be adjusted because the network was too sparse or there were not enough homogeneous stations to span a long gap. The bug in version 3.2.2 affected 267 stations (177 of these from the USHCN network of stations; Fig. 1c&1d).

3. Changes to global anomalies and trends

The small additions of new source data and the software bug fixes resulted in minor changes to the GHCN-M global analysis. The global scale comparison is shown in Figure 3. All differences in annual average temperature are less than 0.1°C. The largest differences are in the first half of the 20th century. These are due principally to the software changes described in section 2c(5) associated with removal of data that could not be adjusted because of large data gaps or an insufficient number of neighbors. When viewed spatially (Fig. 4) for the 2014, differences in annual average anomalies are generally less than 0.5°C.

The warmest year on record for average global land surface temperature (2007) remains the same in the new version. There is little change in the rankings of the ten warmest years on record (Table 2), with the only change in ranking of the 10th warmest year. Similarly there is little change in the rankings of the monthly and seasonal rankings between the two versions (not shown). When combined with sea surface temperatures, the global land/ocean temperature rankings for the ten warmest years are unchanged; the year 2014 remains the warmest year on record.

There are small changes in global annual average trends between the old and new version. The global land 1880-2014 annual average trend increased from 0.97°C/century to 1.00°C/Century. Spatially, the largest and most widespread changes in century-scale trends occurred in South America, West Africa, and southern Asia (Fig. 5). The global land surface

trend over the past 15 years (2000-2014) is little changed (+0.13°C/Decade versus +0.14°C/Decade).

4. **References**

Lawrimore, J., M. Menne, B. Gleason, C. Williams Jr., D. Wuertz, R. Vose, and J. Rennie (2011), An Overview of the Global Historical Climatology Network Monthly Mean Temperature Dataset, Version 3, J. Geophys. Res., doi:10.1029/2011JD016187, in press.

Menne, M.J., and C. N. Williams Jr. (2005), Detection of undocumented changepoints using multiple test statistics and composite reference series. *J. Climate*, 18, 4271–4286.

Menne, M.J., and C.N. Williams (2009), Homogenization of temperature series via pairwise comparisons. *J. Climate*, 22, 1700-1717.

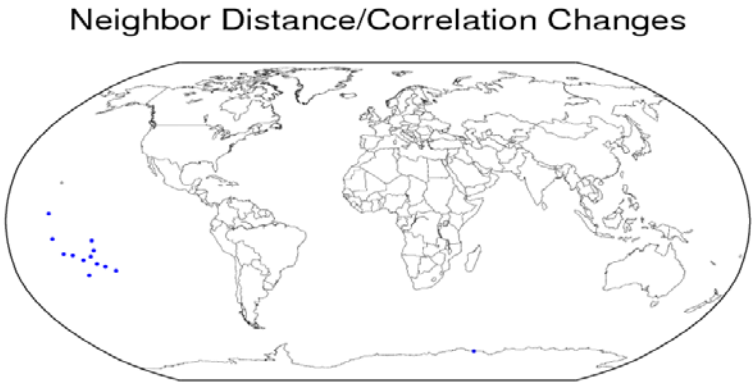

Figure 1a. Locations of the 12 South Pacific and 1 Antarctic station affected by the failure to properly initialize neighbor arrays.

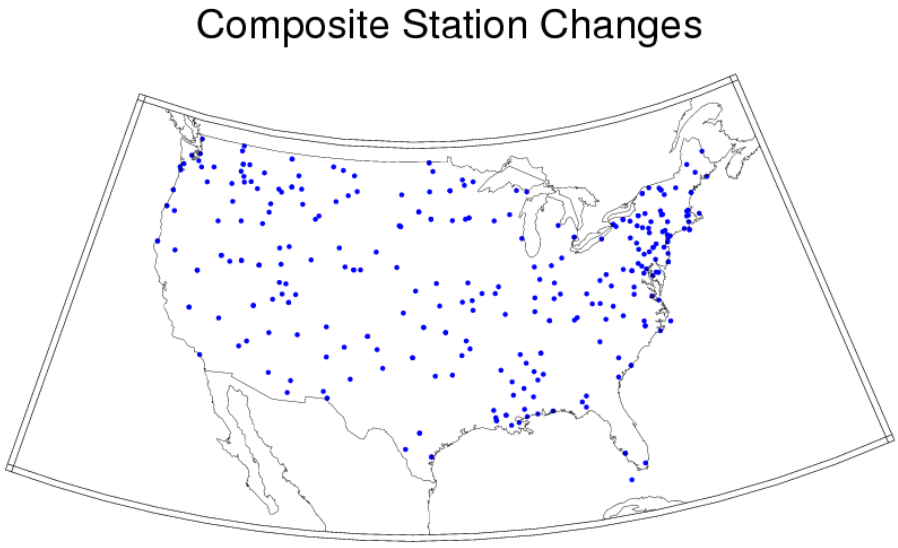

Figure 1b. The location of the 254 stations in version 3.2.2 for which neighbors of the threaded USHCN stations which were being used to identify inhomogeneities in the PHA process unintentionally included stations that had been previously used to produce the threaded USHCN station record.

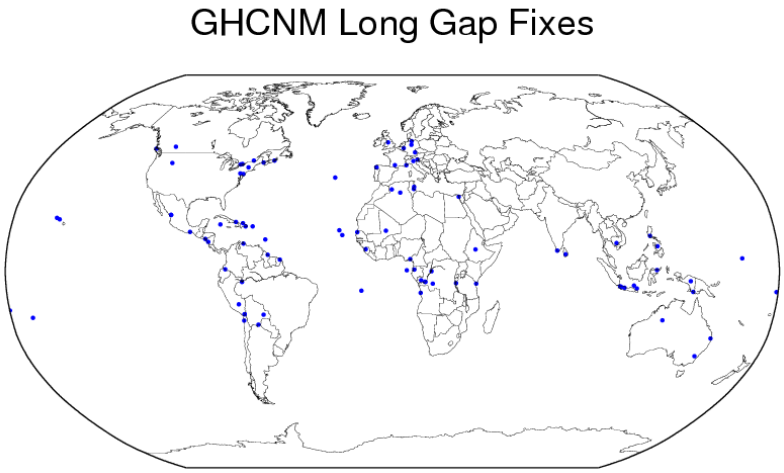

Figure 1c. Locations of the 90 GHCN-M version 3.2.2 stations for which data preceding the inhomogeneity should have been removed because there were insufficient neighbors to compute an accurate adjustment.

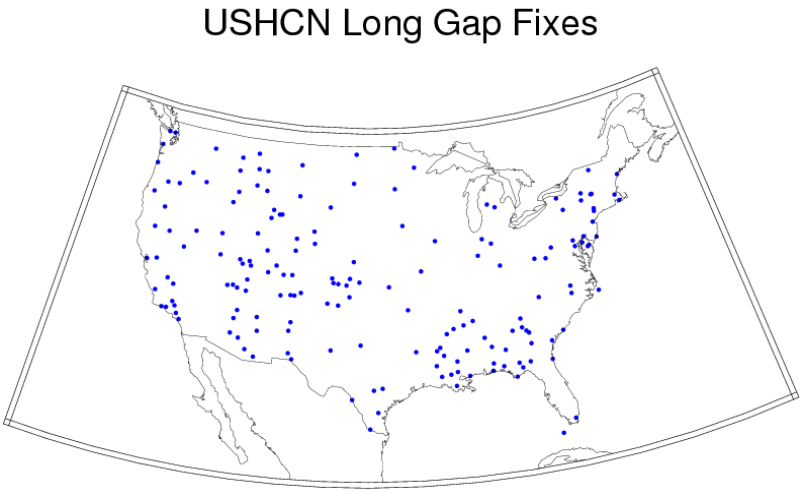

Figure 1d. Locations of the 177 USHCNv2.5 stations in GHCN-M version 3.2.2 for which data preceding the inhomogeneity should have been removed because there were insufficient neighbors to compute an accurate adjustment.

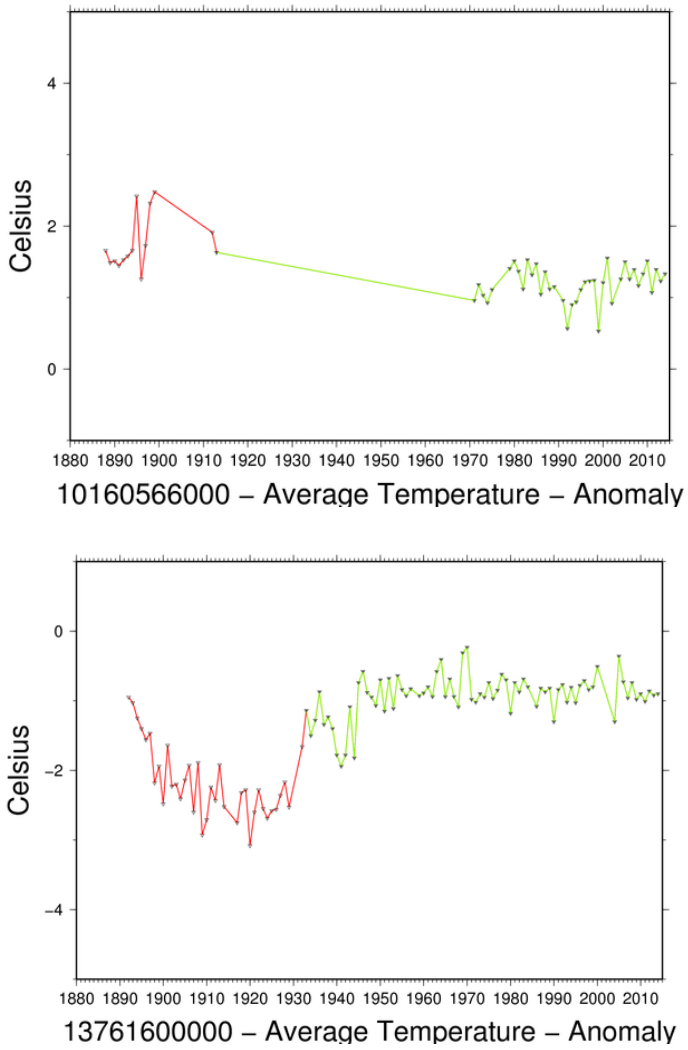

Figure 2. Two of the 267 stations for which data preceding the inhomgeneity (red line) should have been removed in GHCN-M version 3.2.2.

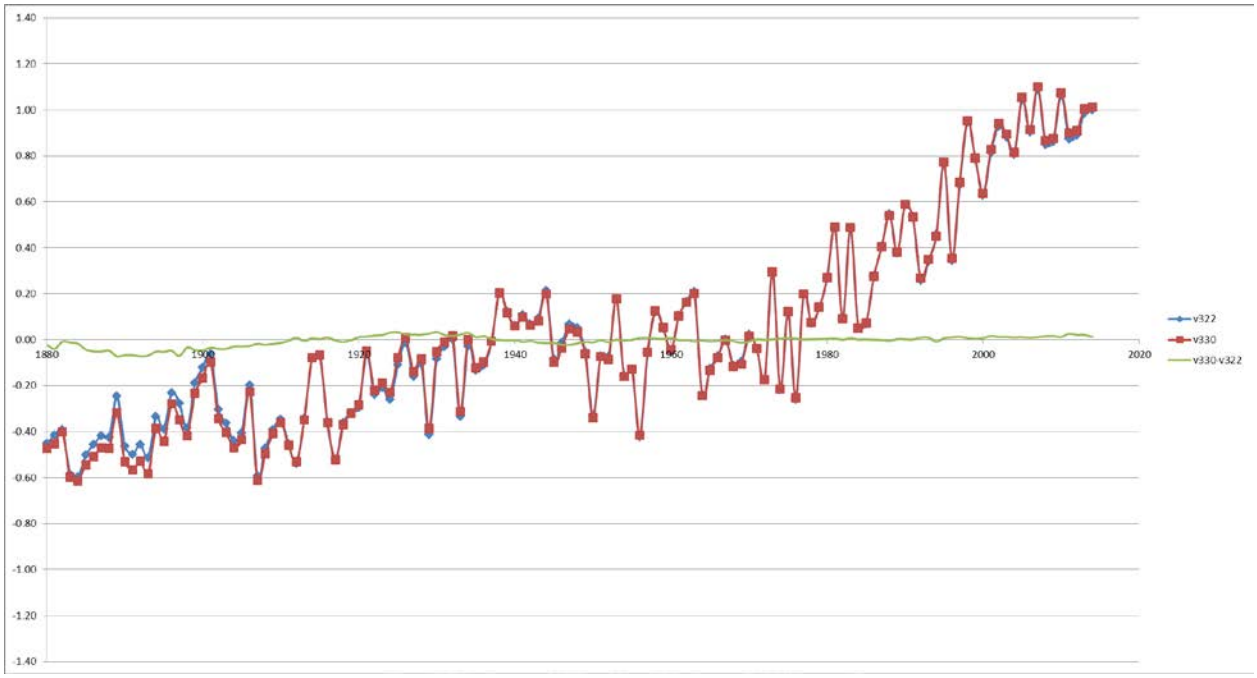

Figure 3. Time series of annual average global land surface air temperature anomalies in GHCN-M version 3.2.2 (blue), version 3.3.0 (red), and the difference (green) between v3.2.2 and v3.3.0 anomalies (v3.3.0 minus v3.2.2).

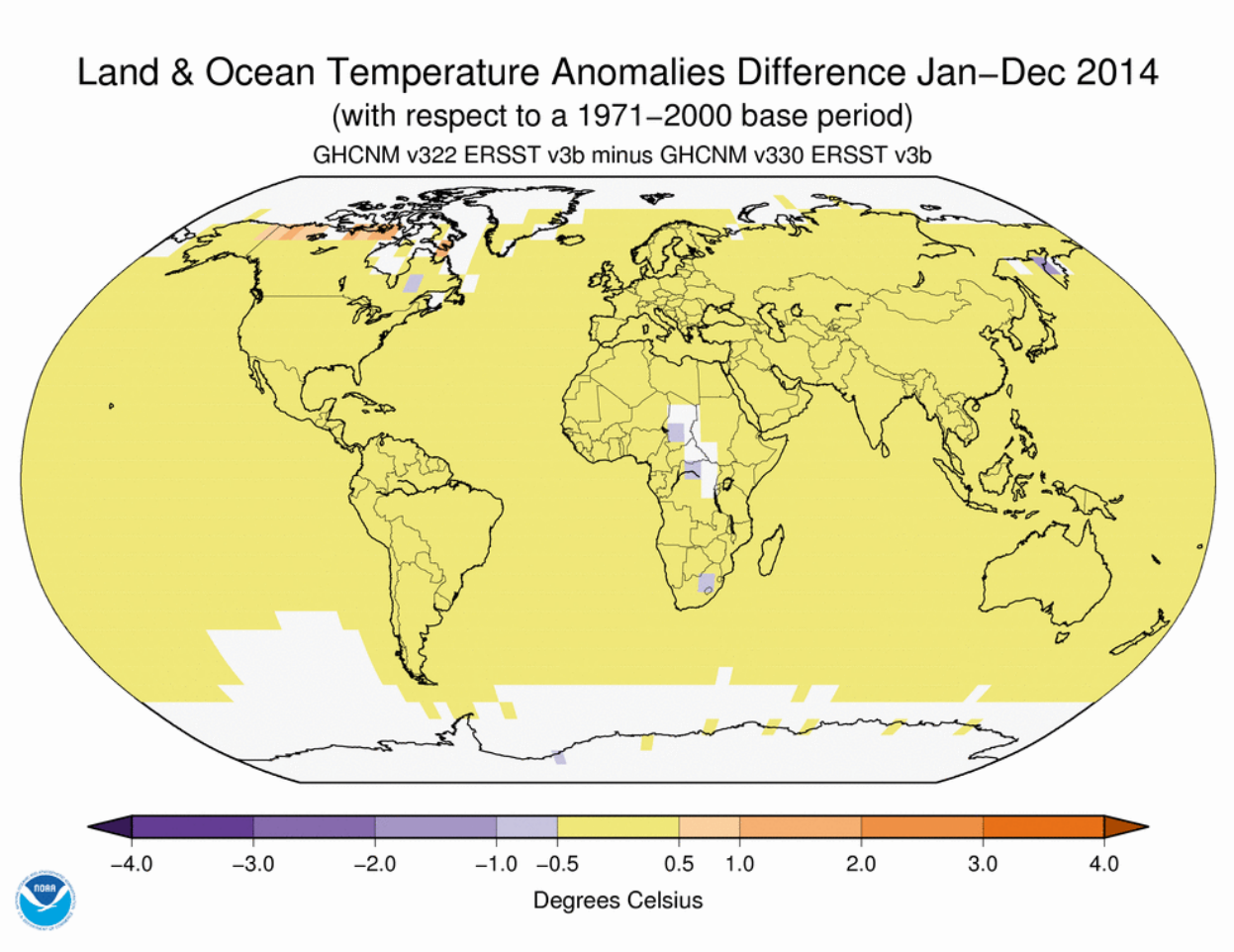

Figure 4. Differences between v3.2.2 and v3.3.0 Land&Ocean combined surface temperature anomalies; annual average for 2014. The sea surface temperatures are from the Extended Reconstructed Sea Surface Temperature (ERSST) data set version 3b and are unchanged in this analysis.

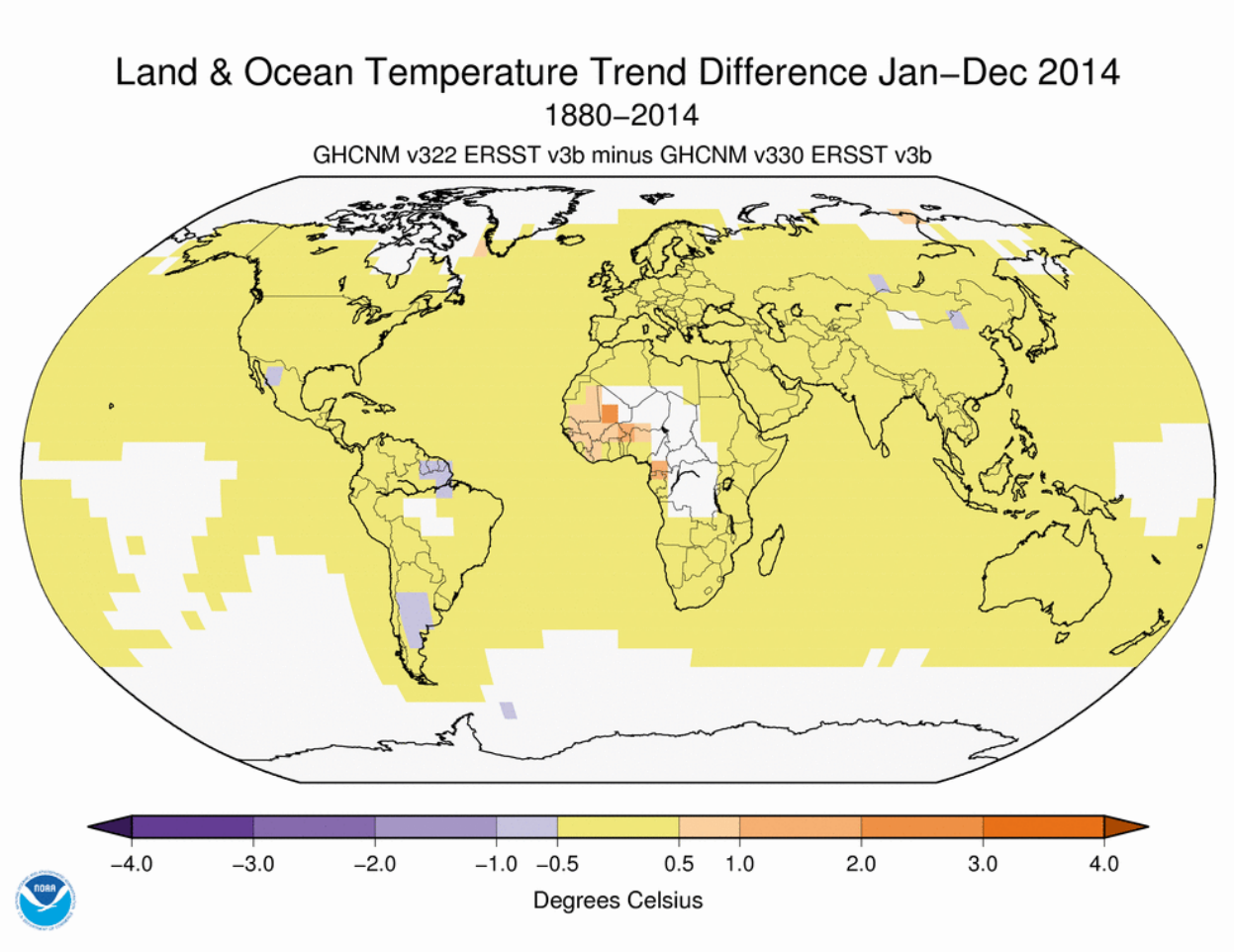

Figure 5. Differences between 1880-2014 trends (°C/Century) in annual average temperature; v3.3.3 minus v3.2.2. The sea surface temperatures are from the Extended Reconstructed Sea Surface Temperature (ERSST) data set version 3b and are unchanged in this analysis.

Table 1: Source datasets from which GHCN-M version 3.3.0 is constructed and maintained.

Priority	Source Dataset	Source Flag
1	Datzilla (Manual/Expert Assessment)	Z
2	USHCN-M Version 2	U
3	World Weather Records	W
4	KNMI Netherlands (DeBilt only)	N
5	Colonial Era Archive	J
6	MCDW final (DSI 3500)	M
7	MCDW monthly	C
8	UK Met Office CLIMAT	K
9	CLIMAT bulletin	P
10	GHCN-M Version 2	G

Table 2. The top 10 warmest years on record for global land annual average temperature for v3.2.2 and v3.3.0.

v3.2.2	Rank	v3.3.0	Rank
2007	1	2007	1
2010	2	2010	2
2005	3	2005	3
2014	4	2014	4
2013	5	2013	5
1998	6	1998	6
2002	7	2002	7
2006	8	2006	8
2012	9	2012	8
2003	10	2011	10