NCDC Technical Report
NCDC No. GHCNM-12-01R

Modifications to Pairwise Homogeneity Adjustment software to improve run-time efficiency

Claude Williams
Matt Menne
Jay Lawrimore
Physical Scientist
NOAA/National Climatic Data Center
Asheville, NC

This report, GHCNM-12-01R, provides descriptions of software modifications made and implemented in GHCN-Monthly version 3.1.0. It is a revision to GHCNM-12-01 made to remove discussion of the Rothenberger fixes which are included in GHCNM-12-02.

August 1, 2012

U.S. DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
National Climatic Data Center
Asheville, NC 28801-5001

Table of Contents

Figures:

Abstract

The software used to perform operational updates of GHCN-M version 3 was modified to improve its run-time efficiency. To reduce the time required to ingest, quality control, and bias correct the version 3 data, traditional array processing was replaced with a technique based on sparse matrices. A comparison of U.S. and global analyses performed using the software before and after the corrections and modifications are included. The impact on annual means resulted in a change in century-scale global and US trends of less than 0.002°C/Decade. The software modifications are incorporated into a new release of GHCN-M, version 3.1.0.

1. Introduction

The Global Historical Climatology Network-Monthly (GHCN-M) version 3.0.0 dataset was released to the public in May 2011. Since GHCN-M was first developed in the 1990s it has been an internationally recognized source of data for the study of observed variability and change in land surface temperature. It provides monthly mean temperature data for 7280 stations from 226 countries and territories, ongoing monthly updates of more than 2000 stations to support monitoring of current and evolving climate conditions, and homogeneity adjustments to remove non-climatic influences that can bias the observed temperature record (Lawrimore et al. 2011). Version 3 introduced a number of improvements and changes to the dataset since the time version 2 was released in 1998 (Peterson and Vose, 1997). Among the changes that were incorporated into version 3 is the use of new approaches to homogenization.

The software used to perform homogenization and operational updates of GHCN-M version 3 was modified to improve its run-time efficiency. During the first ten days of each month the version 3 dataset is updated nightly to incorporate monthly mean temperature observations from the previous month which are transmitted over the Global Telecommunications System. This process provides the data necessary for global analyses that are conducted as part of NCDC's State of the Climate reporting while also supporting other climate monitoring activities such as those performed at NASA-GISS. To reduce the time required to perform the ingestion, quality control, and bias correction of the version 3 data, traditional array processing was replaced with a technique based on sparse matrices.

These changes, which are discussed in the following sections, are incorporated into a new release of GHCN-M, version 3.1.0.

2. Implementation of Sparse Matrices

The GHCNMv3.0.0 Pairwise Homogeneity Algorithm (v52g) contains massive work arrays to collect and analyze the full network changepoints over the three phases of the process. The size of the arrays is so large that the full dataset could not be processed as a single entity on the existing servers (6Gb memory). To work around the hardware limitations, the station data were split into nine regions for separate processing. The regional data were patched back together after the processing was complete. With the incorporation of new and more powerful servers (74Gb memory), the processing was put back together and run as one region.

But although the code could run as a single entity, too many resources were required for efficient run-time processing. More than 23Gb of memory was required and the global processing took more than four hours to complete for each of three datasets (mean, maximum, and minimum temperature). Although this was manageable at the time, it could not support planned additions of thousands of stations to GHCN-M in upcoming versions.

To address this problem, the software was modified to replace the existing array structure with procedures using sparse matrices, using a method similar to that of Duff, Erisman, and Reid (1986). Conceptually, the procedure involves a rearrangement of a full matrix (which begins and ends at the period of record limits for the dataset) to a "reduced" matrix that contains storage space for each station's individual period of record only. After implementation, the sparse matrix version is 15% and 25% of the size of the previous full-matrix versions for GHCN-M and USHCN processing, respectively, and the execution time is 50% less than the legacy process.

3. Changes to global anomalies and trends

The changes to array handling led to minor changes in the final series of many station series. Figure 1 is a map of the differences between the GHCNMv3.0.0 and GHCNMv3.1.0. The figure enhances the stations with differences (colored red and blue) since they plot on top of those with very small changes (colored gray). Many of these stations have short period of records so that the trend change does not reflect a change over the entire 116 years which is the period of record for data selection from the network. Several of the station series are displayed in Figure 2 showing the changes between the GHCNMv3.0.0 (red) and the GHCNMv3.1.0 (green) PHA output. Figure 3 shows the distribution of the differences in all individual station trends between the operational as applied in GHCNMv3.0.0 (52g) and the new version as implemented in GHCNMv3.1.0 (RBFix). The impact of these changes on the average trends for the USHCN and the GHCN is approximately 0.0015°C/Decade. There are no biases inherent to the changes, and therefore the average monthly global temperature series (Fig. 4) contain insignificant differences. Similarly, the USHCN station subsets of maximum and minimum temperatures show no biases in the distribution of the station trend differences (Fig. 5) and almost indiscernible differences in average series (Fig. 6 & 7).

4. Conclusions

The application of changes to the Pairwise Homogeneity software as discussed in Sections 2 replaced traditional array management with the use of sparse matrices. The modified code was tested to ensure reliability and the results of the revised version were compared with the operational version to identify the impact of these software changes (section 3). There was no overall bias in the global or U.S. trends, but the trends for some individual station series were affected. More than half of the stations in each dataset had trend differences less than +/-.025°C per decade following application of the software changes. Since there were no overall biases, the differences in the globally averaged annual series are barely perceptible.

5. References

Duff, I.S., A.M. Erisman, and J.K. Reid., Direct Methods for Sparse Matrices. Clarendon Press, Oxford, UK, 1986. Also see: http://web.eecs.utk.edu/~dongarra/etemplates/node378.html.

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.

Peterson, T.C., and R.S. Vose, 1997: An overview of the Global Historical Climatology Network temperature database. Bulletin of the American Meteorological Society, 78 (12), 2837-2849.

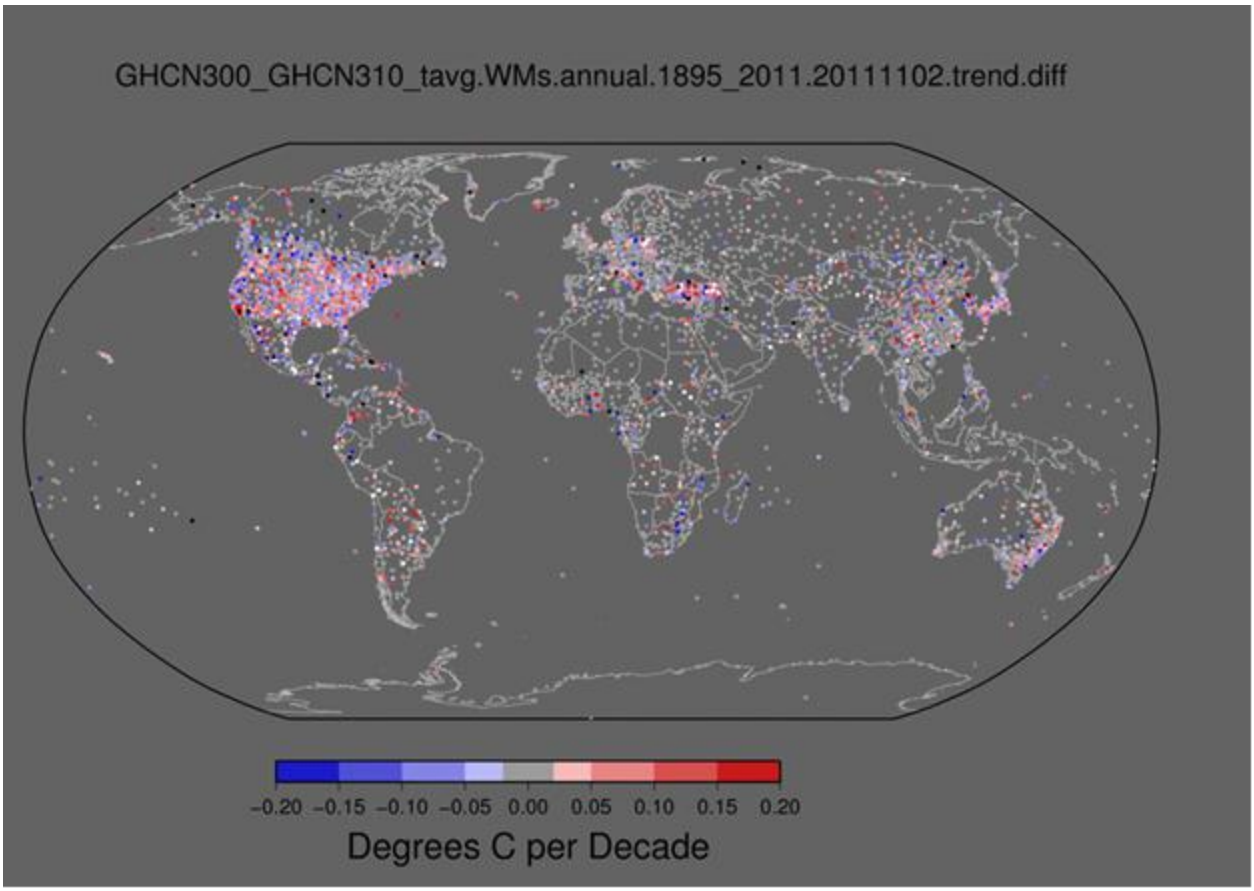

Figure 1. Map of individual Station trend difference between Operational GHCNMv3.0.0 and v3.1.0. Dots are plotted such that those with higher positive or negative differences overwrite those with differences closer to zero, thus highlighting the changes.

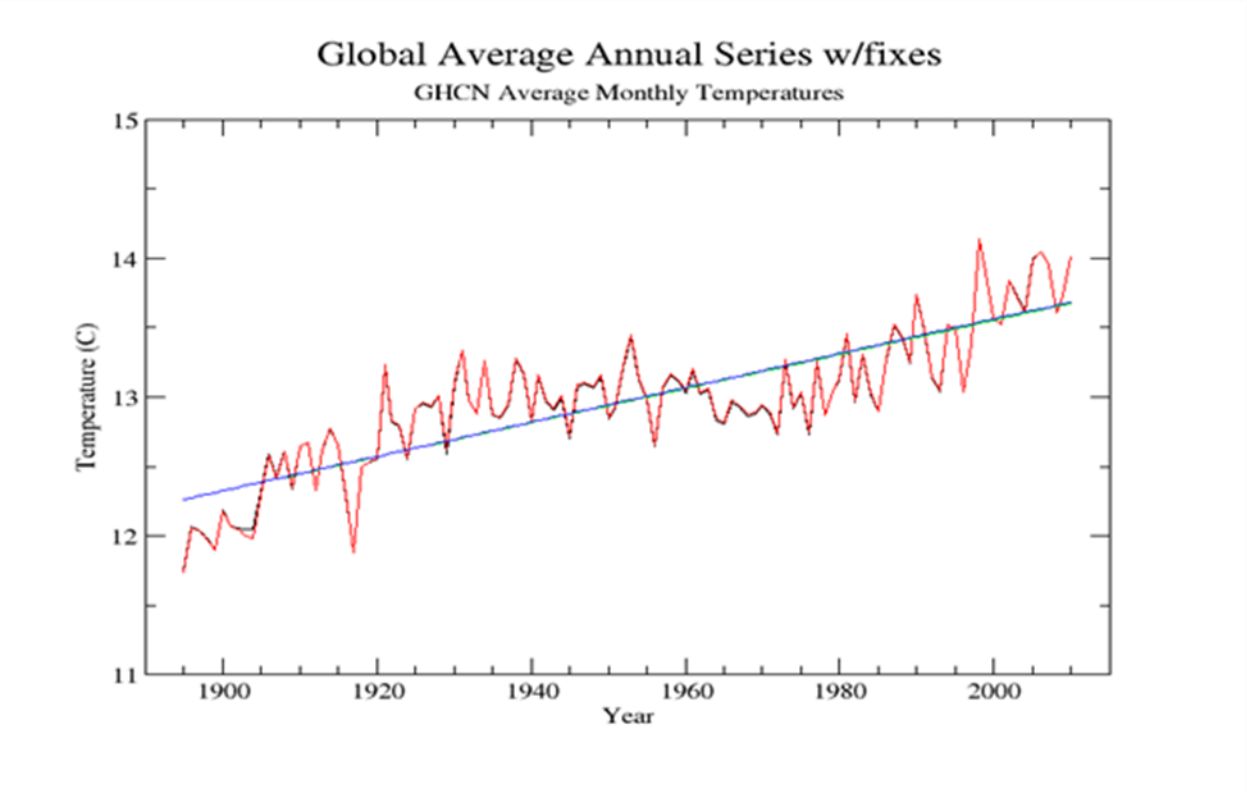

Figure 2. Global land surface annual mean temperatures from 1895 to 2010, before and after application of software changes. GHCNMv3 Operational version 3.0.0 (blue) and version 3.1.0 (red) after modifying the Operational version with Skyline Matrix recoding. Annual averages were calculated by arithmetically averaging all stations.

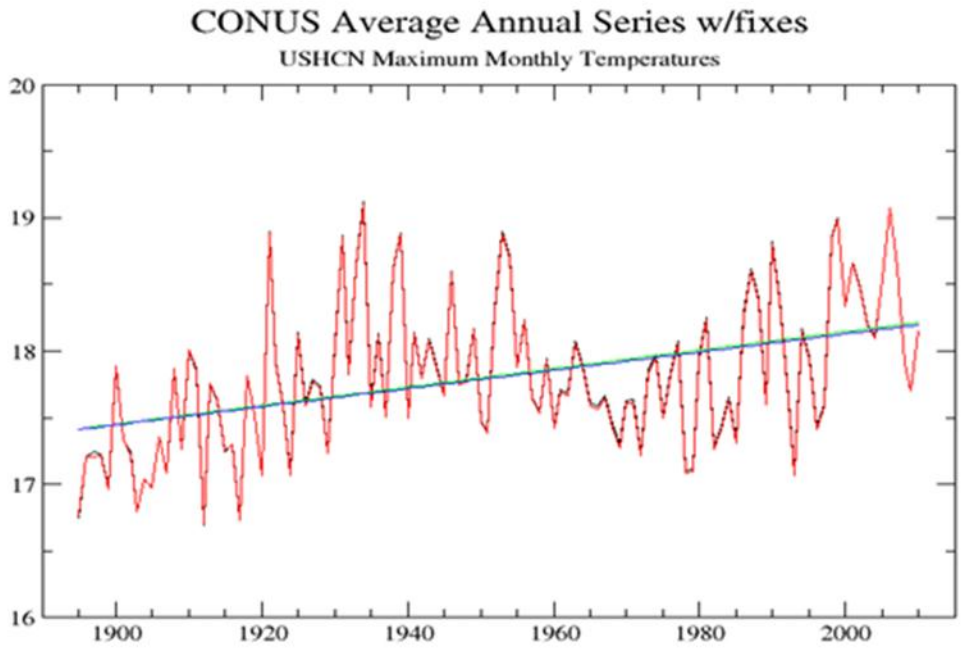

Figure 3. Contiguous U.S. mean annual maximum temperatures from 1895 to 2010, before and after application of software changes. Annual averages were calculated by arithmetically averaging all USHCN stations.

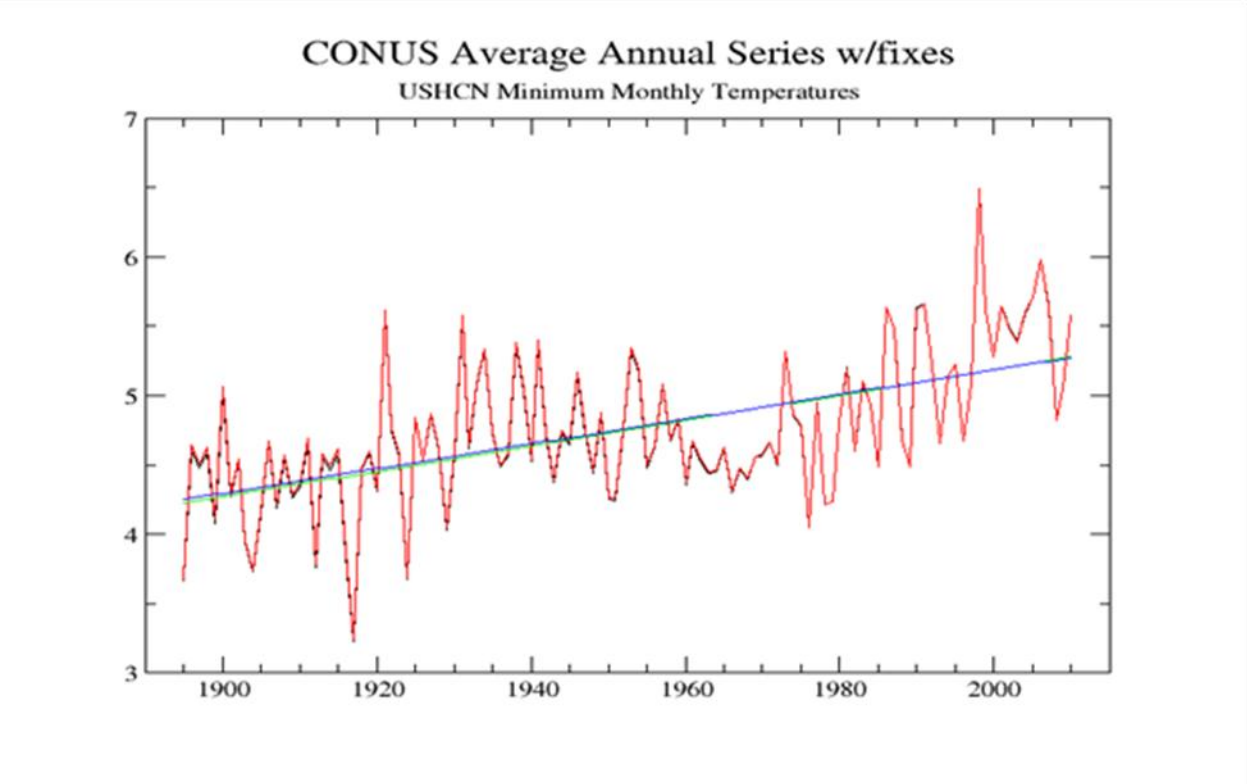

Figure 4. Contiguous U.S. mean annual minimum temperatures from 1895 to 2010, before and after application of software changes. Annual averages were calculated by arithmetically averaging all USHCN stations.