J14.1 U.S. AND GLOBAL IN SITU DATASETS FOR THE ANALYSIS OF CLIMATE VARIABILITY AND CHANGE

Jay H. Lawrimore*, Byron E. Gleason, Claude N. Williams, Matthew J. Menne, and William E. Angel
NOAA's National Climatic Data Center, Asheville, North Carolina

1. Introduction

The National Oceanic and Atmospheric Administration's (NOAA) National Climatic Data Center (NCDC) has a long tradition of dataset development and observational analysis for understanding the Earth's changing climate. Dataset development for the purpose of climate change study began in earnest at NCDC in the 1980s. The earliest efforts focused on data acquisition, quality control and analysis of data at a monthly temporal resolution. Within the past ten years greater attention has been placed on daily data, in part to provide insights into variability and change of extreme climate conditions.

NCDC continues to build on this tradition with new quality control methods for both daily and monthly data, improved algorithms for ensuring homogeneity of the climate record, and use of new technologies to improve data acquisition and processing. NCDC is positioned to provide higher quality monthly and daily climate data with better timeliness and easier access through enhancements to two widely used sources of U.S. and global data; the Global Historical Climatology Network-Monthly dataset (GHCN-M) and U.S. Cooperative Summary of the Day data.

The GHCN-M dataset is recognized worldwide as an authoritative source of instrumental land surface temperature and precipitation data. It is an essential starting point for studies of changes in land surface temperature and precipitation from the latter half of the 1800s through the present. The importance of this dataset to the climate community requires an ongoing commitment to ensuring the highest level of quality. As part of this commitment, NCDC will release an updated version of GHCN-M for monthly mean temperature, Version 2.5, in February 2010. This version will include better quality control and improved homogeneity adjustments for thousands of land surface stations worldwide.

A widely used source of daily data, NCDC's Cooperative Summary of the Day dataset for the U.S., often referred to as DSI-3200, is comprised primarily of data from the U.S. Cooperative Observer Program (COOP). With more than 8,000 stations currently operating, it meets an array of needs for weather and climate information. But for decades users have become accustomed to delays of approximately six months from the time of observation until quality controlled data are available for distribution. These delays largely result from the need to digitize paper forms and conduct quality control and processing. Within the past several years, efforts to establish automated data reporting technologies for the COOP network (Redmond et al. 2008) and separate efforts focused on developing new and more efficient quality control procedures have laid the groundwork for significant improvements to NCDC's Cooperative Summary of the Day data in 2010.

2. GHCN-Monthly Dataset

The Global Historical Climatology Network-Monthly version 2 dataset was released in 1997. It contains 7280 mean temperature stations, 4966 stations with monthly maximum and minimum temperature, and 20590 stations with monthly precipitation totals. Mean temperature and precipitation data begin as early as 1701 and 1697, respectively. The number of stations varies through time with fewer stations early in the record and a peak in number during the 1960s (Figure 1). Many sources of data were used in constructing GHCN (Peterson and Vose, 1997), but operational updates to the dataset are comprised primarily of data transmitted by WMO countries over the Global Telecommunication System.

2.a. Homogeneity Adjustments

Many temperature records in GHCN-M version 2 have been adjusted to remove inhomogeneities that are caused by non-climatic factors such as station moves, changes in instrument technologies and the station environment, and changes in observing practices. Such influences can introduce changes in the temperature record that are unrelated to true changes in the background climate.

* *Corresponding author address:* Jay H. Lawrimore, NOAA National Climatic Data Center, 151 Patton Avenue, Asheville, NC 28801; e-mail: Jay.Lawrimore@noaa.gov.

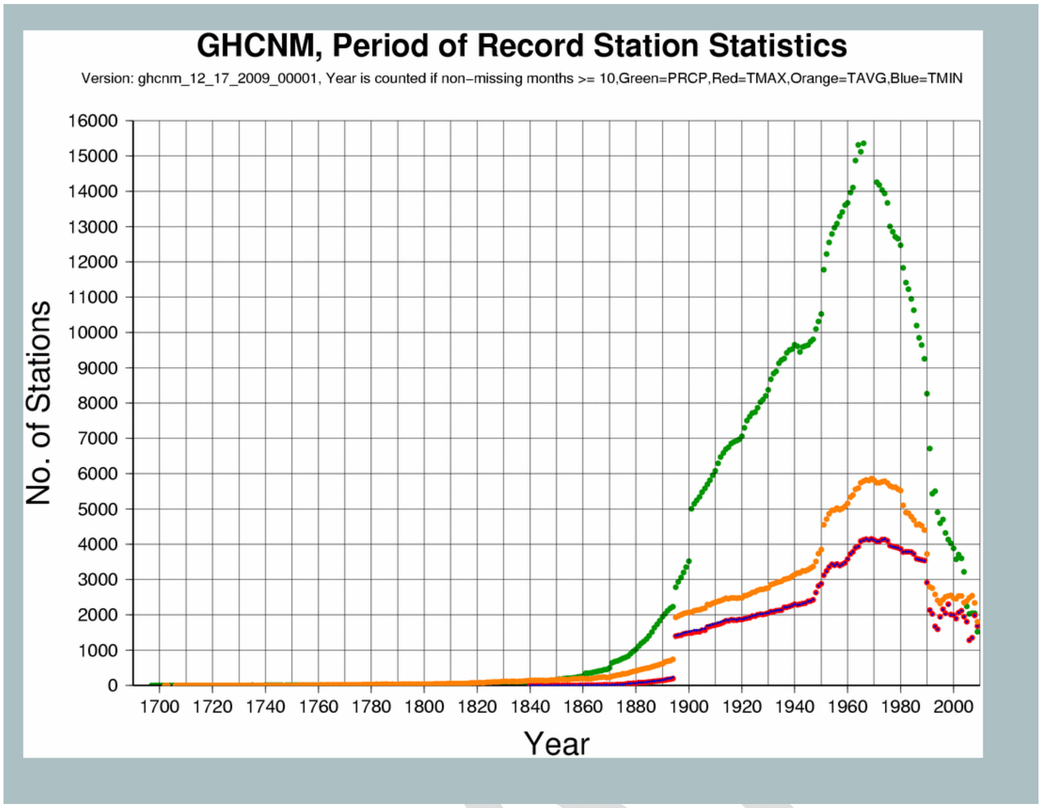

Figure 1. Number of stations with maximum, minimum, and mean temperature, and monthly total precipitation in GHCN Monthly Version 2 datasets.

The introduction of inhomogeneities in the observational record can occur gradually or abruptly. Gradual changes may include growing urban populations and infrastructure as well as changes in the natural environment such as vegetation growth and decline. Inhomogeneities that can occur abruptly include changes in the type of instrument used for measuring air temperature, a change in procedures used for making temperature measurements, and changes in the location of the observing station. Unless properly accounted for, inhomogeneities can alter conclusions regarding the rate of change or direction of trends in observed temperature.

Adjustments were applied to GHCN-M version 2 temperature data to remove inhomogeneities (Peterson and Easterling, 1994; Easterling and Peterson, 1995). While these homogeneity adjustment techniques improved the overall quality of the GHCN-M temperature data, concerns associated with the use of reference series for detecting inhomogeneities were identified in 2005 (Menne and Williams, 2005). This finding led to the eventual development of an adjustment methodology based on the use of station pair difference series between the target station and its surrounding neighbors (Menne and Williams, 2009).

The resulting pairwise algorithm for undocumented changepoint detection is superior to the reference series approach applied in GHCN-Monthly version 2 because it reduces the number of false alarms, and unlike the reference series approach there are no requirements for a group of series to have a common base period. As a result, the estimation of step-change magnitude is not confined to the shortest homogeneous interval within a group of neighboring series (Menne and Williams, 2009).

The pairwise algorithm was used in development of the U.S. Historical Climatology Network Version 2 dataset (Menne et al. 2009) and its improved characteristics make it well suited for inhomogeneity detection and adjustment in GHCN-M Version 2.5. The pairwise approach will initially be used in development of an adjusted monthly mean temperature dataset, followed by application to monthly maximum and minimum temperature in the ensuing months. An example of homogeneity adjustments for minimum temperature in Reno, Nevada is shown in Figure 2.

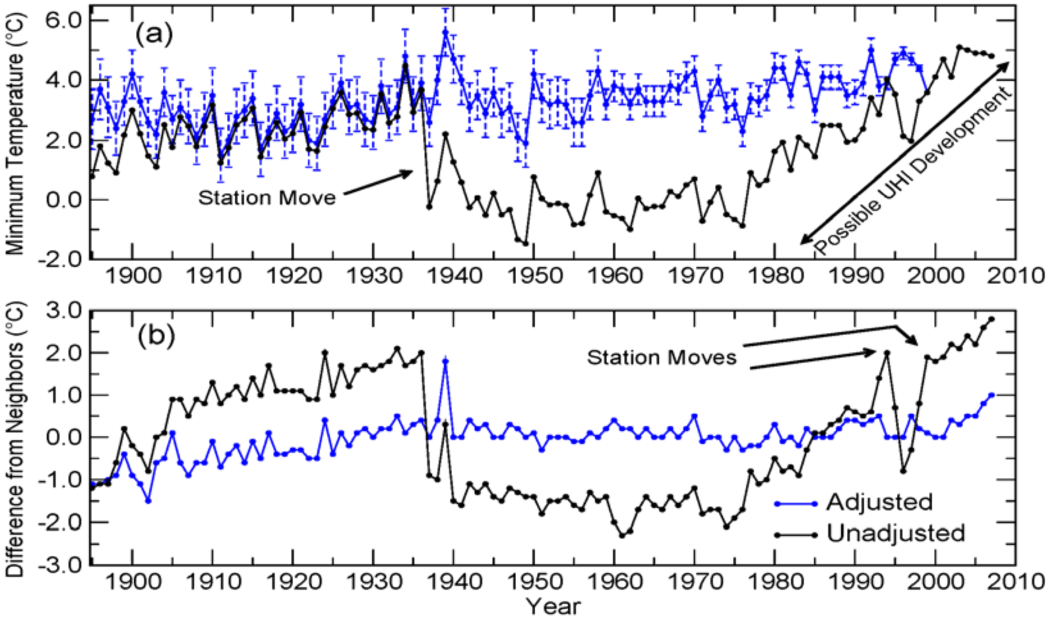

Figure 2. (a) Mean annual unadjusted and fully adjusted minimum temperatures at Reno, Nevada. Error bars depict a measure of the cumulative uncertainty (b) Difference between minimum temperatures at Reno and the mean from its 10 nearest neighbors. (Menne et al. 2009).

2.b. Quality Control

Improved quality control procedures will also be included in the new version of GHCN-M. They will be based on algorithms developed for the GHCN-Daily global temperature and precipitation datasets (Durre et al. 2009) and recently applied to quality control of USHCN-Monthly Version 2 data (Menne et al. 2009). Nineteen checks were developed to detect duplicate data, climatological outliers, and internal, temporal, and spatial inconsistencies. Thresholds were selected based on manual review of random samples of flagged values and the performance of each check was evaluated using the method described in Durre et al. (2008). Checks based on this method have a low false-positive rate and a low miss rate, both less than 5%.

Version 2.5 of the GHCN-M monthly mean temperature dataset will be distributed as a beta release in February 2010. Monthly mean maximum and minimum temperature data will be released later in the year.

3. Cooperative Summary of the Day

The Organic Act formally established the U.S. Cooperative Observer Program (COOP) in 1890. Observations provided by more than 8,000 volunteer observers continue to this day. They typically consist of daily maximum and minimum temperatures, snowfall, snow depth, and 24-hour precipitation totals. Since its inception, volunteers have recorded daily observations on paper forms and mailed them at the end of each month to a responsible National Weather Service office. Although the NWS delivered observer forms to NCDC for many years, for the past decade a federal contractor in Kentucky has been responsible for keying and data transfer to NCDC. Photocopies of the observer forms are maintained and publicly accessible via NCDC's Environmental Document Access and Display System (EDADS).

3.a. Quality Control

After ingest at NCDC, data are subjected to a series of quality control procedures involving manual review and editing of suspect data (Reek and Crowe, 1990). The entire process from observation to acquisition, quality control, and archive takes five to six months to complete and culminates in the addition of a month of daily observations to NCDC's DSI-3200 dataset. Although advances in quality control techniques have been made at various times since the network's inception (see e.g., Angel et al. 2003), a number of problems associated with acquisition and quality control persist to this day. These include an unusually high rate of false-positives (good data values erroneously identified as invalid during the quality control process) and quality assurance methods that vary throughout the period of record resulting in inconsistent levels of data quality.

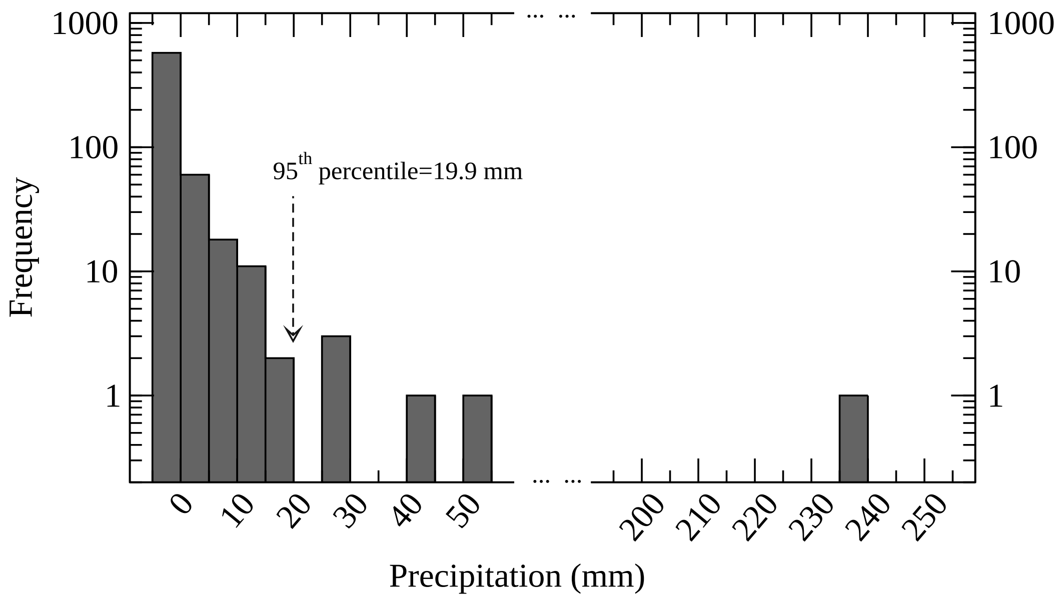

Figure 3. Percentile-based climatological outlier check. One of the 19 checks performed as part of the Summary of the Day quality control process. Histogram of daily precipitation totals reported between August 6 and September 3 throughout the 1966-1990 period of record at Gold Hill, Utah, showing an outlier flagged by the outlier check (Durre et al. 2009)

To address these problems, new quality control procedures have been established which follow the paradigm established in Durre et al. (2008). The application of this strategy has already been completed as part of development and operational processing of NCDC's Global Historical Climatology Network-Daily (GHCN-D) dataset (Durre et al. 2009). The strategy involves complete automation in the form of a robust and reliable quality control system, in which data are analyzed consistently and objectively. Manual intervention is used extensively *prior to* implementation of the quality control algorithms to ensure the validity of thresholds and logic in the system's decision- making. This differs from the traditional semi-automated process for COOP data, where decisions made by automated procedures are manually evaluated as part of the operational quality control process and sometimes overridden.

As part of the new strategy, thorough documentation of the system's performance is required before implementation, including an empirical assessment of false-positive and flag rates, information on types of errors removed and detected, as well as conditions under which errors might remain. Documentation on the processes and thresholds applied in the quality control process are made available to users to aid them in making an informed decision about how to appropriately apply the data.

Advantages to this process over traditional methods involving manual intervention include the removal of the subjective component intrinsic to any process with a human interface. This method also provides a consistent set of quality control checks throughout the period of record, instead of antecedent practices that introduce new quality measures at various times throughout the period of record. Most importantly, the ability to process the entire period of record makes it possible to apply quality control retrospectively as new methods are developed and to do so in a consistent manner throughout the life of the data.

Before the end of 2010, subjective assessments of data quality through manual quality control procedures will no longer be part of operational processing of Cooperative Summary of the Day data. Quality control will take place solely through the GHCN-Daily quality control process, and these data which have comprised DSI-3200 for decades will be available only as part of GHCN-Daily. This process will include nineteen checks designed to detect duplicate data, climatological outliers, and internal, temporal, and spatial inconsistencies (Durre et al. 2009). An example of a climatological outlier check is shown in Figure 3.

3.b. Improvements in Timeliness

In addition to providing a dataset with objective and consistent quality control throughout the period of record and a low false-positive rate, the automated nature of these processes will allow NCDC to perform quality control checks twice daily as new data arrive.

This will reduce by several months the time required to quality control Cooperative Summary of the Day data and enable NCDC to provide associated climate products with much greater timeliness than has been the standard.

While these changes will provide significant improvements to data quality and timeliness, the continued use of paper forms for data entry will remain a complicating factor for several more years. In addition to delays in data acquisition, data quality problems associated with the use of paper forms will likely persist. These problems are due to factors including keying errors, indecipherable handwriting, and errors of transcription, such as observers writing data into the wrong column on the form.

To address these problems, NCDC in partnership with the NWS and the Western and High Plains Regional Climate Centers, developed the Weather Coder web-based data input and transmission system (Redmond et al. 2008). The current version of this system, Weather Coder III (WxCoder3), allows climate observers to enter daily observations for near real-time reporting via the web and also incorporates phone-based entries through IV-ROCS (Interactive Voice-Remote Observation Collection System). The WxCoder3 web browser displays the entry system and allows the user to enter observations which are checked for errors by the system during the submission process. Possible errors include such things as daily minimum temperature exceeding maximum temperature. At the end of 2009, approximately 2,500 COOP stations in the 8300 station network had been converted from paper forms to web-based data entry. The long range goal is to convert 80% of the network to WxCoder3 operation by 2012.

While conversion of COOP observations from paper to web-based entry will continue for the foreseeable future, NCDC is nearing completion of changes in operational procedures to replace the existing DSI-3200 quality control process with GHCN-D quality control as described above. Full integration is expected to occur before the end of 2010. At that time all data which have traditionally been provided in DSI-3200 will be available as part of the GHCN-Daily dataset.

4. Summary

This paper provided an overview of improvements to two widely used datasets developed and maintained at NOAA's National Climatic Data Center; the Global Historical Climatology Network-Monthly and the U.S. Cooperative Summary of the Day data. The planned release of an updated version of GHCN-Monthly, Version 2.5, in February 2010 will include improved quality control and homogeneity adjustments for thousands of land surface stations. Advances related to the U.S. Cooperative Summary of the Day data include incorporation of automated data reporting technologies and new quality control procedures developed and implemented as part of NCDC's GHCN-Daily dataset.

Significant research and development are continuing in quality control and homogeneity adjustments. Additional meteorological elements and data sources will be added to the quality control process along with appropriate and more refined checks. Comparisons and evaluations of the inhomogeneity algorithm with others in the global climate community will be used to test the process. Refinements and advances in statistics are continually being assessed to enhance the algorithm and efforts are underway to expand the homogeneity algorithm capabilities to daily data.

Data provided by NOAA's National Climatic Data Center are available online at http://www.ncdc.noaa.gov/.

5. References

Angel, W.E., M.L. Urzen., S.A. Del Greco., and M.W. Bodosky, 2003: Automated validation for summary of the day temperature data (TempVal). *83rd AMS Annual Meeting, combined preprints CD-ROM, 9-13 February 2003, Long Beach CA, 19th Conference IIPS [International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology]*, American Meteorological Society, Boston, Mass., File 15.3, 4 p.

Durre, I., M.J. Menne, and R.S. Vose, 2008: Strategies for evaluating quality assurance procedures. *Journal of Applied Meteorology and Climatology*, 47(6), 1785-1791.

Durre, I., M.J. Menne, B.E. Gleason, T.G. Houston, and R.S. Vose, 2009: Comprehensive Automated Quality Assurance of Daily Surface Observations. *Journal of Applied Meteorology and Climatology*, Submitted.

Easterling, D.R., and T.C. Peterson, 1995: The effect of artificial discontinuities on recent trends in minimum and maximum temperatures. *Atmospheric research*, 37 (1-3), 19-26.

Menne, M.J., and C.N. Williams, Jr., 2005: Detection of undocumented changepoints using multiple test statistics and composite reference series. *Journal of climate*, 18 (20), 4271-4286

Menne, M.J., and C.N. Williams Jr., 2009: Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, 22(7), 1700-1717.

Menne, M.J., C.N. Williams, and R.S. Vose, 2009: The United States Historical Climatology Network Monthly Temperature Data - Version 2. *Bulletin of the American Meteorological Society*, 90(7), 993-1107.

Menne, M.J., I. Durre, R.S. Vose, B. Gleason, and T. Houston, 2009: An Overview of the Global Historical Climatology Network Daily Dataset. *Journal of Climate,* Submitted.

National Climatic Data Center (NCDC), 2009: *Data Documentation for Data Set 3200 (DSI-3200).* Asheville, NC, 19 pp. < http://www.ncdc.noaa.gov/oa/documentlibrary/>

Peterson, T.C., and D.R. Easterling, 1994: Creation of homogeneous composite climatological reference series. *International journal of climatology*, 14 (6), 671-679.

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.

Redmond, K.T., G. McCurdy, G. Kelly, M.J. Brewer, T.W. Owen, and B. Bonack, 2008: Weather coder III: Web-based climate data ingest for NOAA's cooperative volunteer observation network. *24th Conference in IIPS*, New Orleans, LA, 20-24 January 2008, American Meteorological Society, Paper 7C.3.

Reek, T., and M. Crowe, 1990: A rules based geographical information system for the geographical edit and analysis of cooperative network data (GEA), *ASAE International Summer Meeting*, Paper no. 904050, American Society of Agricultural Engineers, St. Joseph, MI.