

# **Verification and Validation Report**

## **GHCN-Monthly Version 4 Mean Temperature**



## **1. Introduction**

This document is separated into two sections. This first concerns verification conducted through software code reviews of the GHCN-M Phase 1 (quality control) and Phase 2 (Bias Correction) processes. The second describes validation designed and executed to confirm the Phase 1 and Phase 2 processes operate as intended.

## **2. Verification**

### **2.1 Phase 1 process (QC)**

A code review of the Phase 1 process was completed on 2-6 June 2014 by Art Burden, Byron Gleason, Claude Williams, and David Wuertz. A review of the “Understand” Tool was conducted and review of output for GHCN-M phase one software. A code review examined ghcnm\_edit.f95, ghcnm\_qc.f95, ghcnm\_process, and main\_mo\_over\_mo\_dup\_check.pro (IDL) software.

The review identified the following issues of concern:

- A) Code was found to be well written but it was noted that the software would benefit from some additional commenting to provide clarification on several “local” matters within the code, including but not limited to variable descriptions.
- B) All pieces of code were found to need a proper header.
- C) Good coding practices recommends that only one variable per line be included in the declarations section.
- D) I/O interactions were found to need more checking, e.g. “inquire”s and graceful exit for non-existent but required input files).
- E) A few unused variables were found, and it was recommended that they be eliminated.
- F) Units need to be assigned to a name
- G) Recommendations included a need for better documentation of the isolated value check algorithm, including information on variables like ISO\_MASK.

- H) The addition of a lookup table for the spatial z-score check was recommended.
- I) The reviewers recommended parameterizing all file units.
- J) For consistency and ease of maintenance the reviewers recommended replacing the IDL code (Month-over-month duplicate check) with Fortran.

Changes to the Phase 1 software were made to address each of the above recommendations. Changes for one recommendation listed below remains under consideration.

- K) All floating point comparisons were found to be written in old style. Ideally these should be considered for the new comparison using "abs" and an "epsilon" parameter.

## **2.2 Phase 2 process (Bias Corrections)**

In 2014, refactoring of the PHA code began under the guidance of the GHCN-M team with leadership provided by GST contractor Diana Kantor. Refactoring was completed in May 2016 and a period of final testing and acceptance began. During this refactoring, code was broken out into multiple small subroutines and functions, with unit tests added for each one. Output files from the original code were compared against output files for the refactored code to ensure that the files were identical. Refactoring of program ushcn\_dist.v5a.combo.f (getting each station's closest neighbors) was completed, and during the process, two bugs were discovered:

- 1) Erroneous stations were being appended to the end of a station's "closest neighbors" list when 99 legitimate close neighbors could not be found. This was the result of an array not being properly reinitialized in between stations. The arrays that collect the closest neighbors information were not being nulled out in between target stations. When there were not 99 neighbors within the 5000km distance range of a target station, neighbors from the previous station's neighbor arrays were left in place. Furthermore, because the number of neighbors was passed into the sort subroutine, those erroneous stations remain unsorted at the end of the arrays. Details on this software bug are available in Trac ticket:

<http://crntools.dev.ncdc.noaa.gov/trac/pha/ticket/1>.

- 2) Composite stations were incorrectly being included as neighbors of USHCN stations in the distance files. This was the result of a logic error in an "if" statement. Details on this software bug are available in Trac ticket: <http://crntools.dev.ncdc.noaa.gov/trac/pha/ticket/2>. A similar refactoring process was then completed for the ushcн\_corr.v5a.combo.f (getting each station's best neighbors by correlation).

**RESPONSE TO FINDINGS:** This refactored code was incorporated into the GHCN-M PHA refactor project and applied to the v4 software.

Following completion of the PHA Refactor, a Test Readiness Review (TRR) was held with the CWC Science Council on May 12, 2016 and the test plan accepted. Final testing was conducted through September 2016 to identify any previously unidentified coding errors. During this testing phase the original scientific PHA code was subjected to full error checking using the Lahey compiler (LF95 –chk). This full checking took approximately two weeks to complete and led to the identification of two minor bugs in the scientific PHA code. The software on the scientific development side and the refactored PHA code were both updated and an additional 30-day testing period was completed. No additional problems were identified.

### **3. Validation**

#### **3.1 Phase 1 Process (QC)**

The GHCN-M v4 data are put through a rigorous suite of quality control checks each time the update system is run in operations. The automated suite of quality control programs is fully responsible for conducting each test and setting the appropriate quality control datum flag for all test failures (refer to the GHCN-M v4 Dataset Description Document for a description of these tests).

To ensure the quality control system is operating as intended, a software program external to the QC suite is used to test the quality control software for accuracy in two ways. First, the test program is used to analyze a known suite of metadata and data, whereby the quality control status of certain data within the suite (either good or bad) is known and has been independently verified. This known and independently verified test dataset is referred to as a "golden dataset". It assists the GHCN-M scientists in ensuring the programmed quality control checks work as expected. Second, this golden dataset and test program are continually evolving during the development process, as new data are added to the golden dataset and tests are performed. In this way the developer can retest (regression testing) the quality control program to ensure that new development does not break, change, harm, or in any way cause the program to take a step backwards, or "regress", in terms of functionality and accuracy.

Validation of the GHCN-M Phase 1 process was conducted using a Golden Dataset that consisted of GHCN-M v3 data from the U.S. state of Texas and the South American country of Chile. The data set contains 70 stations and just over 377 Kilobytes of data. Data from the State of Texas were chosen due to its large geographical extent, relative high station density, and diversity of climate regimes. Chile was chosen due to its large latitudinal extent, its location in the Southern Hemisphere, its proximity to Ocean, and varied geographical features (coast to mountains).

The primary purpose of the golden dataset is to assist GHCN-M scientists and developers in verifying the functionality and accuracy of the quality control steps within the GHCN-M processing system. Data values of known accuracy and inaccuracy are evaluated by executing the quality control software. Successful completion of all quality control checks requires that each of the known invalid observations be correctly identified as invalid. If all invalid observations are correctly identified as such the test completes successfully.

There are 14 invalid observations in the golden data set that should be identified as invalid by the quality control software. Two or more invalid observations are specific to

each quality control check. The set of invalid observations and the associated check that identifies each is listed below.

- 1) Duplicate year, identify duplicate years
- 2) Duplicate year, identify duplicate years
- 3) Isolated value, identify gaps in record
- 4) Isolated value, identify gaps in record
- 5) Climatological Outlier, slightly positive z-score
- 6) Climatological Outlier, gross negative z-score
- 7) Climatological Outlier, gross positive z-score
- 8) Climatological Outlier, slightly negative z-score
- 9) Spatial Outlier, positive value
- 10) Spatial Outlier, negative value
- 11) Spatial Outlier (Z-Score), Chile
- 12) Spatial Outlier (Z-Score), Texas
- 13) Consecutive duplicate months, Chile
- 14) Consecutive duplicate months, Texas

A sample of the test procedure process is as follows.

- 1) compile main program
- 2) execute main program using golden dataset as input, output is quality controlled golden dataset
- 3) compile test program
- 4) execute test program using quality controlled golden dataset as input.

Resultant output for successful test will be as follows:

PASSED: TEST #001  
PASSED: TEST #002  
PASSED: TEST #003  
PASSED: TEST #004  
PASSED: TEST #005  
PASSED: TEST #006  
PASSED: TEST #007  
PASSED: TEST #008  
PASSED: TEST #009  
PASSED: TEST #010  
PASSED: TEST #011  
PASSED: TEST #012  
PASSED: TEST #013  
PASSED: TEST #014

**RESULTS:** Findings of Passed Tests verify that the quality control process is performing as intended. That is, data values known to be invalid are flagged as invalid by the quality control process.

If a failure for a test had occurred, an example result would be "FAILURE: TEST #002", and the developer would immediately be made aware of a potential new coding error that has "regressed", and been reintroduced into the software.

It is intended that developers use the test program routinely, but not necessarily at every compilation, to monitor their developmental performance and to avoid software programming errors. At a minimum the validation is performed any time a software change is made through a major, moderate or minor version upgrade.

## **3.2 Phase 2 Process (Bias Corrections)**

Identifying inhomogeneities and estimating adjustments with the Pairwise Homogeneity Algorithm (PHA) relies on a selection of choices for all steps in the PHA process from how to define target and reference series to the particular statistical breakpoint tests applied and mechanisms for adjusting each detected break. Ideally an optimum set of

choices is made to create an algorithm that has the best performance in detecting and adjusting each inhomogeneity.

To assess the performance of the specific set of parameters selected in the PHA algorithm, a set of plausible analogs was created from which the truth was known a priori. The analog worlds share the likely principal characteristics of the raw data such as spatio-temporal sampling structure, noise and bias characteristics. The PHA algorithm was then run against the suite, allowing a quantifiable appraisal of algorithm strengths and weaknesses.

The PHA algorithm, as described in Menne and Williams (2009), was evaluated against eight analog datasets. A large-scale (contiguous U.S.) long-term trend metric is used as the measure of performance.

To ensure plausible geographical data structures and teleconnections the analog worlds were derived from gridded output from Global Climate Models (GCMs). A range of climate model runs were downloaded from the World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model dataset (Meehl et al., 2007) and these were each sub-sampled in space and time to the observational mask. Although not the real world, GCMs do mimic many of the gross characteristics and use of a range of models, mitigating against any issues that may be introduced through non-plausible characteristics in any single model. Because the models are at much coarser resolution than the typical station, separation climatological offsets and white noise were applied in each case before any further steps. This ensures that nearby 'stations' arising from the same GCM gridpoint are non-identical and mimics likely real-world physical offsets due to local environment and elevation as well as random errors.

Five principal break structures were assigned (Perfect data; Big breaks good metadata; Mixed break sizes some clustering; Very many mainly small breaks, Clustering and sign bias). For the last of these, four distinct analogs were created that while sharing the

exact same breaks differed in their underlying climate change signal and interannual variability, bringing the total number of analog worlds to eight.

**RESULTS:** One hundred randomized versions of the PHA were compiled using different values for the parameters listed in Table 1, and the 100 different versions of the PHA were run on the eight analog datasets to assess the parametric uncertainty. Figure 1 provides an example of the results for one of the eight analog worlds. The contiguous U.S. trend produced from the data corrected using the operational (default) configuration of the PHA is very near the “true” trend computed from the homogenous data (before breaks were added). Similar conclusions are reached from the other analog worlds. The operational PHA, as well as many of the other different randomized versions of the PHA is able to move the trend more than 95% toward the true climate signal.

The eight analogs provide a measure of confidence that the pairwise algorithm will adjust monthly temperature series such that their regional mean is moved closer to the true value, even when the series contain pervasive errors with a sign bias that are clustered in time - regardless of the underlying climate forcings. Likewise, based on the other analogs, there is no evidence that the pairwise algorithm will move the trend away from truth when there is no sign bias to the errors, or when there are no errors at all.

Based upon performance against the analogs it can be concluded that the algorithm is better than 92% of the randomly detuned versions, balances type1 and type2 errors, and highly unlikely to consistently make incorrect inferences if the real world data are biased. Complete details on this analysis are available (Williams et al. 2012).

## 4. REFERENCES

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor (2007), The WCRP CMIP3 multi-model dataset: A new era in climate change research, *Bull. Am. Meteorol. Soc.*, 88, 1383–1394, doi:10.1175/BAMS-88-9-1383.

Menne, M. J., and C. N. Williams (2009), Homogenization of temperature series via pairwise comparisons, *J. Clim.*, 22, 1700–1717, doi:10.1175/2008JCLI2263.1.

Williams, C. N., M. J. Menne, and P. W. Thorne (2012), Benchmarking the performance of pairwise homogenization of surface temperatures in the United States, *J. Geophys. Res.*, 117, D05116, doi:10.1029/2011JD016761.

## 5. Tables

Table 1. System Parameters used in the bias correction process.

Algorithm Step	Keyword Name	Optimum Value	Functional Description
Choosing Neighbors	NEIGH_CLOSE NEIGH_CORR  CORR_LIM  MIN_STNS  NEIGH_FINAL	100 1diff  0.1  7  40	Maximum # nabor series to consider Method used for ranking nabors based on degree of similarity (1diff=calculate corr using first differences) Minimum correlation coeff with tgt to qualify as a nabor Minimum # of nabors with coincident data Final (max) # of nabors per tgt station
Resolving breaks in difference series	SNHT_THRES BIC_PENALTY	5 BIC	SNHT significance threshold (in percent) Penalty function used to determine form of the break; BIC=Bayesian Information Criterion
Identify the series causing the break	SHF_META  AMPLOC_PCT  CONFIRM	1  92  2	Toggle for metadata (1=identify undocumented breaks and exploit metadata) Confidence window table used to coalesce changepoints Number of target-nabor difference series with coincident breaks required to implicate the target as the source of the break
Estimating the magnitude of the break	ADJ_MINLEN  ADJ_MINPAIR  ADJ_OUTLIER  ADJ_WINDOW  ADJ_FILTER  ADJ_EST  NS_LOOP	18  2  1  24  Conf  Medi  1	Min length of data period that can be adjusted (months) Min # of pairwise estimates of break size req'd to determine size of adj. Toggle to test and remove outliers using the Tukey outlier test Min # of months before and after a break in the difference series necessary to calculate breakpoint size Outlier filtering method for the pairwise break estimates Method used to determine the adjustment factor from the multiple pairwise estimates Toggle to merge data segments when break size is statistically insignificant (this loop increases the length of the homogenous segments available to estimate other breakpoint sizes in data sparse periods)

## 6. FIGURES

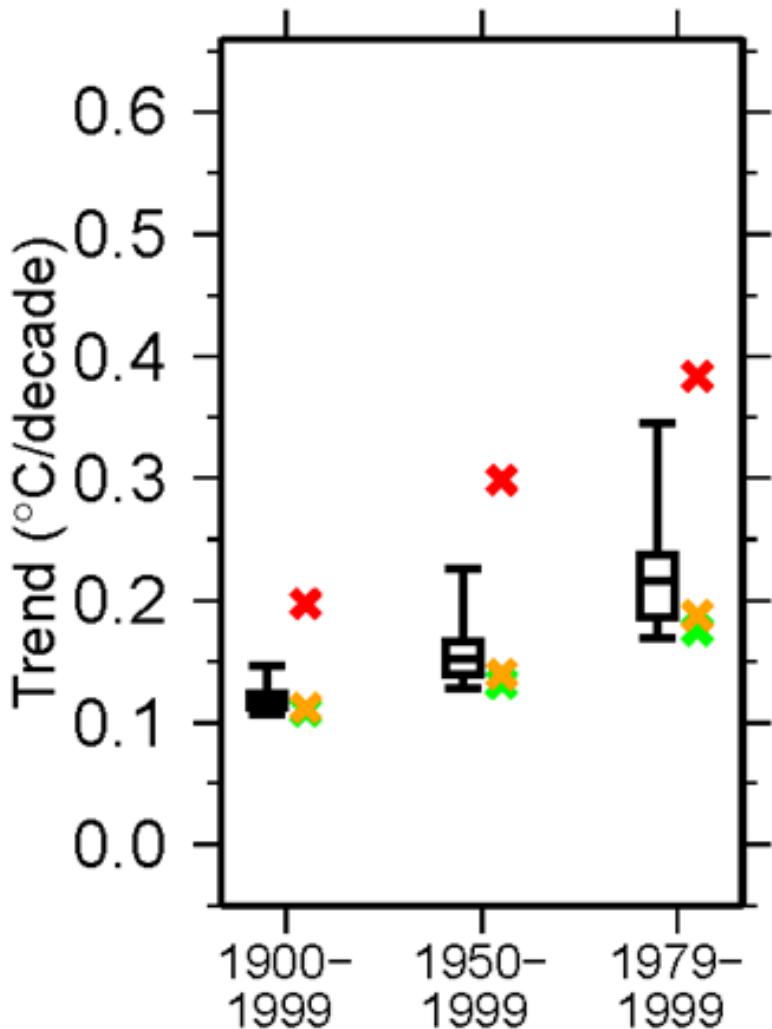


Figure 1. Box plot depicting the range of CONUS average trends for three different summary periods produced by the 100 randomized versions of the pairwise homogenization algorithm (from Analog World 4; Clustering and sign bias). The magnitude of the CONUS average trends based on the raw input data are given by the red “X,” the magnitude of the true (homogeneous) trends are given by the green “X.” The magnitude of trends produced by the default version of the homogenization algorithm is shown by the yellow “X.” Whiskers denote the full range, boxes the inter-quartile range and horizontal line within the box the median estimate for the 100 member ensemble. (Williams et al. 2012)