



Workshop on climate data management,
data sharing and exchange.

DGM-WMO, 4-5 and 8 November 2021



Overview on Assessing Data Quality and Homogeneity

Driss BARI

National Center of Climate
Moroccan Meteorological Service, Casablanca, Morocco
bari.driss@gmail.com

Workshop on climate data management, data sharing and exchange
DGM-WMO 4-5 and 8 November 2021

- 1 Rationale
- 2 Climate Data Quality Control : Concepts
- 3 Climate Data Quality Control: Tools
- 4 Climate Data Homogeneity: Concepts
- 5 Climate Data Homogeneity: Tools



- 1 Rationale
- 2 Climate Data Quality Control : Concepts
- 3 Climate Data Quality Control: Tools
- 4 Climate Data Homogeneity: Concepts
- 5 Climate Data Homogeneity: Tools





World Meteorological
Organisation

WMO No. 100

**Guide to Climatological
Practices**

Edition 2018

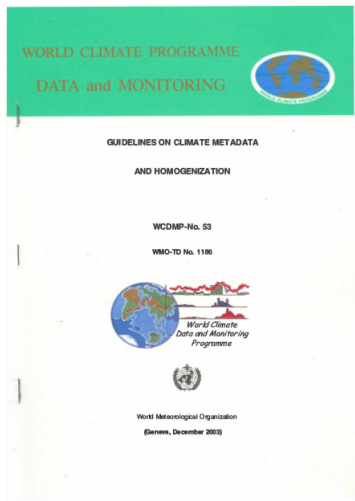


World Meteorological
Organisation

WMO No. 1238

**Manual on the High-quality
Global Data Management
Framework for Climate**

Edition 2019

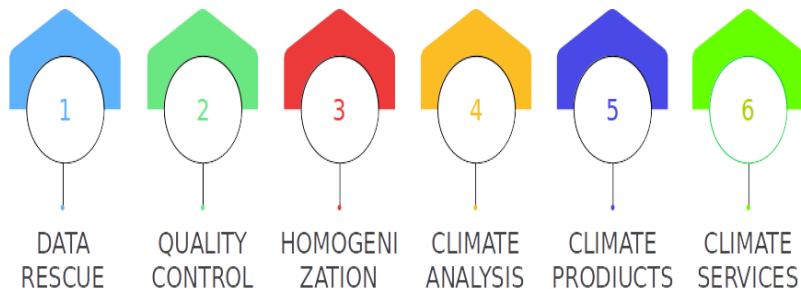


World Meteorological
Organisation

WMO-TD No. 1186

**Guidelines on Climate
Metatdata and
Homogenization**

Edition 2003



- 1 Rationale
- 2 Climate Data Quality Control : Concepts
- 3 Climate Data Quality Control: Tools
- 4 Climate Data Homogeneity: Concepts
- 5 Climate Data Homogeneity: Tools



Data quality control

The process of ensuring that errors in the data are detected and flagged. It involves checking the data to assess **representativeness** in time, space and **internal consistency**, and flagging any potential errors or inconsistencies.

The purpose of quality control is to ensure that meteorological and climate data available to potential users are sufficiently reliable to be used with confidence. Quality control is therefore part of the overall data quality assessment.



Data quality assurance

It refers to the processes for maintaining a desired level of quality in a dataset or collection. Data verification, quality control and validation are important steps in supporting defensible products and decisions. **Data quality assurance is required across the whole data life cycle and should also include ensuring effective transmission and secure management of the data.**

Any available details about the exact techniques applied will be a great help for the future data user if provided, as well as information on the data that fail the tests and the period which the tests have been run for.



- **Metadata errors** often manifest themselves as data errors. For example, an incorrect station identifier may mean that data from one location apparently came from another; an incorrect date stamp may mean the data appear to have been observed at a different time.
- **Data errors** arise primarily as a result of **instrumental, observer, data transmission, key entry and data validation** process errors, as well as changing data formats and data summarization problems.



Types of Data Quality Control Tests

- **Format tests** : Checks should be made for repeated observations or impossible dates, etc.
- **Completeness tests** : For some elements, missing data are much more critical than for others. Total monthly rainfall amounts may also be strongly compromised by a few days of missing data, particularly when a rain event occurred during the missing period.
- **Consistency tests** : The four primary types of consistency checks are **internal, temporal, spatial and summarization**.
- **Tolerance tests** : set upper or lower limits to the possible values of a climatological element (such as wind direction, cloud cover, and past and present weather)



Internal consistency tests

Internal consistency relies on the physical relationships among climatological elements. All elements should be thoroughly verified against any associated elements within each observation.

- psychrometric data should be checked to ensure that the reported dry bulb temperature equals or exceeds the reported wet bulb temperature
- the relationship between visibility and present weather should be checked for adherence to standard observation practices.
- a maximum value must be equal to or higher than a minimum value.
- sunshine duration is limited by the duration of the day
- global radiation cannot be greater than the irradiance at the top of the atmosphere
- wind direction must be between 0° and 360°
- precipitation cannot be negative



Temporal consistency tests the variation of an element in time. This change usually depends on the element, season, location and time lag between two successive observations.

- A temperature drop of 10°C within one hour may be suspect, but could be quite realistic if associated with the passage of a cold front or onset of a sea breeze.
- A lack of change could indicate an error. For example, a series of identical wind speeds may indicate a problem with the anemometer.



Spatial consistency and Summarization tests

Spatial consistency

It compares each observation with observations taken at the same time at other stations in the area.

Summarization tests

By comparing different summaries of data, errors in individual values or in each summary can be detected.

For example, the sums and means of daily values can be calculated for various periods such as weeks, months or years. Checking that the total of the twelve monthly reported sums equals the sum of the individual daily values for a year provides a quick and simple cross-check for an accumulation element like rainfall.



- 1 Rationale
- 2 Climate Data Quality Control : Concepts
- 3 Climate Data Quality Control: Tools**
- 4 Climate Data Homogeneity: Concepts
- 5 Climate Data Homogeneity: Tools



← → ↻ 🏠 etccdi.pacificclimate.org/software.shtml

ETCCDI Climate Change Indices

ETCCDI Climate Change Indices

Indices
Definition
Calculation
Homogeneity
Examples

Software

The software packages for data homogeneity (RHtestsV4) and indices calculation (RClimDex) are based on a very powerful and freely available statistical package R which runs under both Microsoft Windows and Unix/Linux. Please see the Quick Guide below to download and install R:

- Quick Guide to download and install R

The software packages are available for download at GitHub.

- RClimDex
- RHtests

<http://etccdi.pacificclimate.org/software.shtml>

Expert Team (ET) on Climate Change Detection and Indices (ETCCDI)



The **EXTRAQC** routines are a set of R-coded functions for quality control. They have been integrated by Enric Aguilar et Marc Prohom (Spain) into the widely used ETCCDI's software R-Climdex.

EXTRAQC routines focus mainly on temperature data and include the following tests:

- Duplicate dates control
- Rounding problems evaluation
- Out of range values, based on fixed threshold values
- Outliers, based on Interquartile Range exceedance
- Interdiurnal differences based on fixed threshold values
- Coherence between maximum and minimum temperatures ($T_{\max} > T_{\min}$)
- Consecutive equal values control



<http://www.c3.urv.cat/softdata.php>

Q.C Software

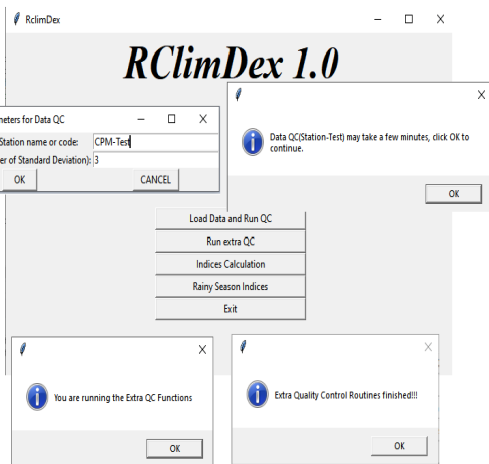
RclimDex-extraqc

 **Manual**



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/)

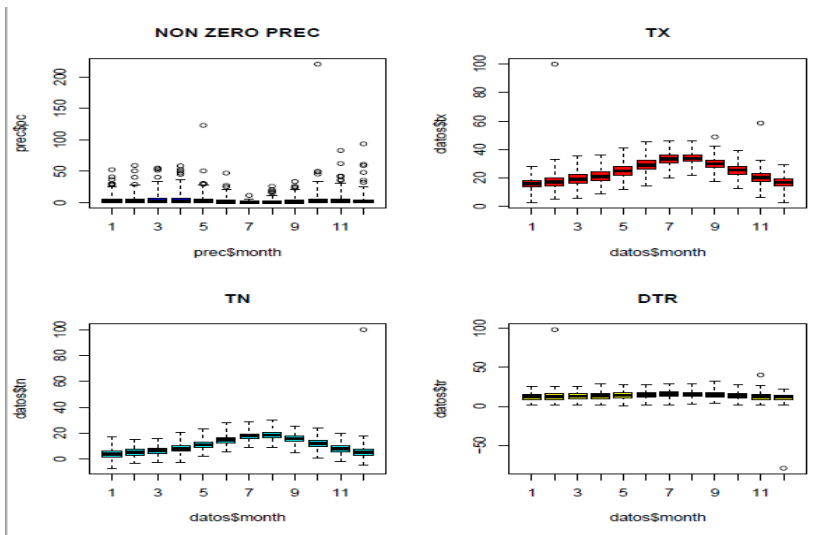
[Download](#)



The screenshot displays the RclimDex 1.0 software interface with several overlapping dialog boxes:

- RclimDex 1.0**: The main application window, showing the title and a menu with options: "Load Data and Run QC", "Run extra QC", "Indices Calculation", "Rainy Season Indices", and "Exit".
- Set Parameters for Data QC**: A dialog box for configuring data parameters. It contains two input fields: "Station name or code:" with the text "CPM-Test" and "Criteria(number of Standard Deviation):" with the value "3". It has "OK" and "CANCEL" buttons.
- Data QC(Station-Test) may take a few minutes, click OK to continue.**: An information dialog box with an "OK" button.
- You are running the Extra QC Functions**: An information dialog box with an "OK" button.
- Extra Quality Control Routines finished!!!**: An information dialog box with an "OK" button.

EXTRAQC : Example



EXTRAQC : Example

Occurrence of 4 or more equal consecutive values

2001	5	15	0	26.7	8.7
2001	5	16	0	28.7	12.4
2001	5	17	0	29.8	14.3
2001	5	18	0	26.1	13.5
2001	5	19	0.9	15.5	13.6
2001	5	20	0	22.8	8.6
2001	5	21	0	27.2	8.8
2001	5	22	0	27.2	10.4
2001	5	23	0	27.2	11.5
2001	5	24	0	27.2	9.5
2001	5	25	0	27.2	10.9
2001	5	26	0	27.2	11.8
2001	5	27	0	27.2	13.7
2001	5	28	0	33.2	12.6
2001	5	29	0	28	13
2001	5	30	0	30.5	13.9

Jumps : the temperature difference with the previous day is greater or equal than 20 °C

2015	12	2	0	19.4	4.1
2015	12	3	0	20.6	99.9
2015	12	4	0	19.4	5.9
2015	12	5	0	19.6	3.1
2015	12	6	0	19.7	3.7
2015	12	7	0	19.5	5.8

Maximum temperature is lower than minimum temperature

2015	12	2	0	19.4	4.1
2015	12	3	0	20.6	99.9
2015	12	4	0	19.4	5.9
2015	12	5	0	19.6	3.1

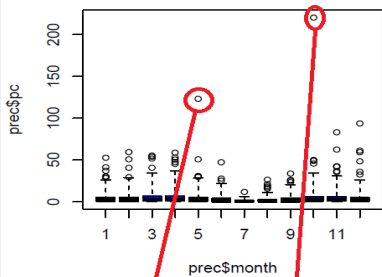
Too large : precipitation values exceeding 200 mm and temperature values exceeding 50 °C.

1982	10	25	220.2	20.2	3.3
2012	11	3	0	58.5	18.3
2013	2	13	0	99.9	1.6
2015	12	3	0	20.6	99.9



EXTRAQC : Example

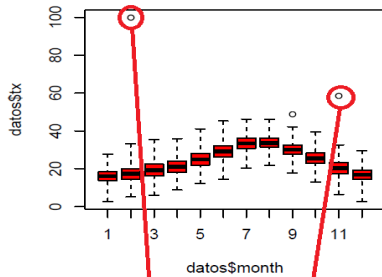
NON ZERO PREC



1982	10	24	0	18.5	5.4
1982	10	25	220.2	20.2	3.3
1982	10	26	0	22.4	4
1982	10	27	0	22.9	3.4
1982	10	28	0	21.5	5.5
1982	10	29	0.3	20.5	4.8

1968	5	7	0	19.8	2.2
1968	5	8	7.7	19.1	4.8
1968	5	9	23.6	12.2	9.1
1968	5	10	123	14	8.3
1968	5	11	2.2	18.5	11.1
1968	5	12	0	26.2	7.1
1968	5	13	0	23.6	11.1

TX



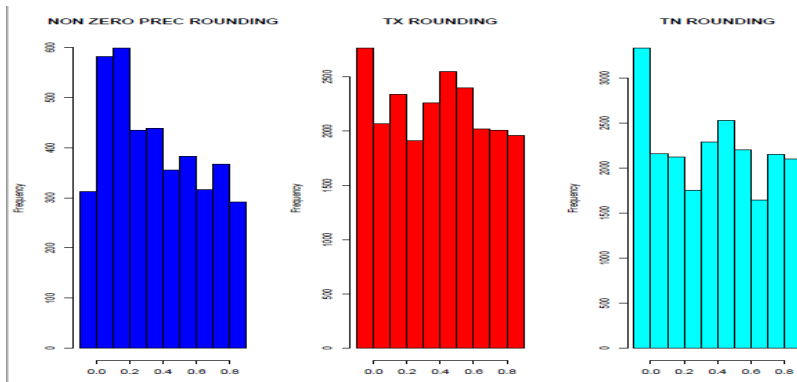
2013	2	10	0	18.1	-0.4
2013	2	11	3	12.6	4.9
2013	2	12	4.3	16.5	5.3
2013	2	13	0	99.8	1.6
2013	2	14	0	19.2	2.8
2013	2	15	0	19.2	4.1
2013	2	16	0	20	3.4
2013	2	17	0	20.8	4.6

2012	11	2	0	24.2	14.1
2012	11	3	0	58.5	18.3
2012	11	4	0	28.9	19.7
2012	11	5	0.2	24.9	18.9
2012	11	6	0.6	23.9	15.6
2012	11	7	0.1	18.6	16.2



EXTRAQC : Example

It looks at rounding problems by plotting the values after the decimal point. It shows how frequently each of the 10 possible values (.0 to .9) appears. It is expected that all the values are well represented.



- 1 Rationale
- 2 Climate Data Quality Control : Concepts
- 3 Climate Data Quality Control: Tools
- 4 Climate Data Homogeneity: Concepts**
- 5 Climate Data Homogeneity: Tools



Climate data can provide a great deal of information about the atmospheric environment that impacts almost all aspects of human endeavour. Climate analysis relies on long time series. If we want to assess if a such or such place has warmed or become wetter, we need to examine 50, 60, ... 100 years of data. **However**, for these and other long-term climate analyses to be accurate, the climate data used must be as homogeneous as possible.

A homogeneous climate time series is defined as one where variations are caused only by variations in climate.

Unfortunately, most long-term climatological time series **have been affected by a number of non-climatic factors** that make these data unrepresentative of the actual climate variation occurring over time.

These factors include changes in:

- instruments,
- observing practices,
- station locations,
- formulae used to calculate means,
- station environment.



Some changes **cause sharp discontinuities** while other changes, particularly change in the environment around the station, can **cause gradual biases** in the data. All of these inhomogeneities can bias a time series and **lead to misinterpretations of the studied climate**. It is important, therefore, **to remove the inhomogeneities or at least determine the possible error they may cause**.

Homogenization

The technique of making time series homogeneous, by application of scientifically sound statistical methods to remove the effects of artificial biases, such as those caused by changes in observational practices, instrumentation, siting, and the like.



Temporal homogeneity of a climate record is essential in climatological research, particularly when data are used to validate climate models, or to assess climate change and its associated environmental and socio-economics impacts. Therefore, **it would be essential to report whether any kind of homogeneity testing has been applied to the data.**

- Which elements have been tested for homogeneity
- During which periods
- On which time scale (daily, monthly, seasonally or yearly)
- Number of inhomogeneities found in each single time-series inhomogeneities, one, two, three inhomogeneities and so on).
- etc.



When assessing the homogeneity of the series we try to identify using statistical techniques and metadata where the heterogeneity of the series has been broken and we attempt to adjust the effect of these ruptures, to improve the quality of our climate inference.

It is almost impossible to be 100% sure about the quality of past data, a homogeneity assessment is always recommended. There is not one single best technique to be recommended. However, the four steps listed below are commonly followed:

1. Metadata Analysis and Quality Control

even in the presence of the most carefully documented metadata, it is advisable to compare what the station history says and what data analysis identifies, as a sort of double check.



2. Creation of a reference time series

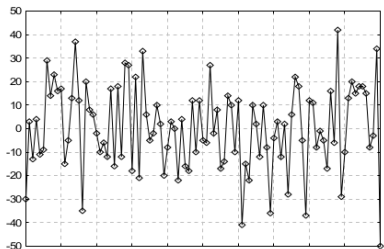
- use relative homogenization, i.e., we compare time series with well correlated neighbours.
- This comparison can be done as pairwise comparisons, averages of several stations or more sophisticated methods, such as PCAs

3. Breakpoint detection

- The idea is to create differences (temperature) or ratios (precipitation) of a candidate series (the one we want to homogenize) towards a reference
- The candidate and the reference share the same climate so the odd features of the candidate minus reference are not due to the evolution of climate



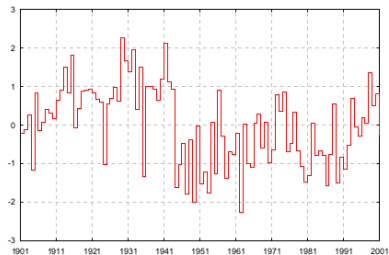
Homogeneity assessment for climate data



Top: Monthly Average of daily minimum temperature for December in Burgos, Spain. Data in 1/10 °C;

Bottom: difference between candidate and normalized reference time series calculated following the Standard Normal Homogeneity Test, using 10 neighbouring stations

The difference between candidate and reference time series (bottom) clearly shows an inhomogeneity in 1941, documented in the metadata as a relocation. The original data (top) mask the inhomogeneity.



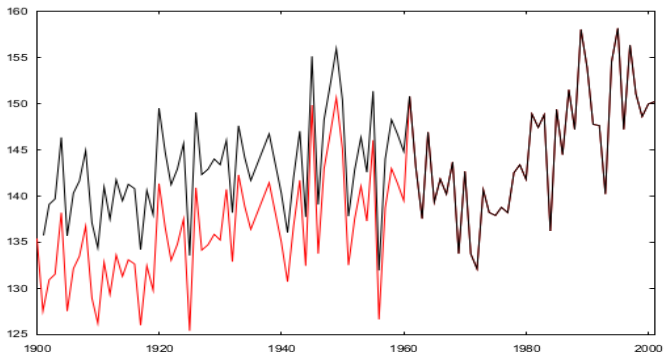
Source : Guidelines on climate metadata and homogenization, Enric Aguilar et al. 2003. WMO-TD No. 1186

4. Data adjustment

- Once the breakpoint identification is finished, the next step is to decide which breakpoints are going to be accepted as real inhomogeneities.
- Data adjustment is the correction applied to data to improve their homogeneity and to make all the observation comparable to the last available data.
- It is always recommended to correct the data to match the conditions of its most recent homogeneous section.



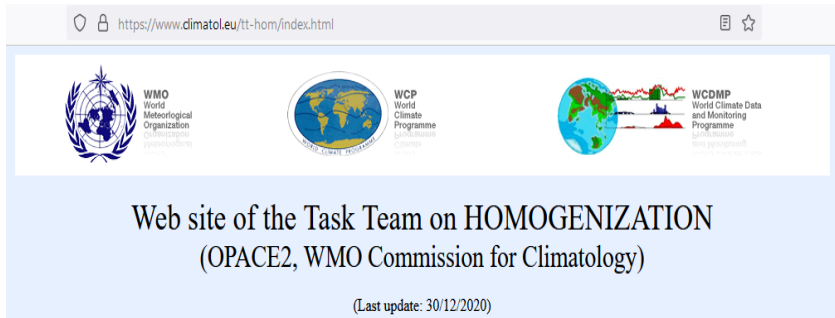
Homogeneity assessment for climate data




Original (red line) and adjusted (black line) annual averages of daily mean temperature for Madrid, Spain. Data in $1/10^{\circ}\text{C}$. Data was adjusted for sudden shifts in mean and artificial trends using an iterative test which compares the mean value of two different periods over a standardized reference time series, calculated from a number of well-correlated reference stations. Inhomogeneous data (red line) show a much larger trend for the 100 years period, as they contain true climate fluctuations plus artificial biases. Figure modified from Aguilar, E (2002) "Homogenizing the Spanish Temperature Series", personal communication to the 7th National Climatology Meeting, Albarracín, Spain.


- 1 Rationale
- 2 Climate Data Quality Control : Concepts
- 3 Climate Data Quality Control: Tools
- 4 Climate Data Homogeneity: Concepts
- 5 Climate Data Homogeneity: Tools**

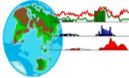




https://www.climatol.eu/tt-hom/index.html

 **WMO**
World Meteorological Organization
المعهد الوطني للأرصاد الجوية
والتنبؤات الجوية

 **WCP**
World Climate Programme
البرنامج العالمي
للمناخ

 **WCDMP**
World Climate Data and Monitoring Programme
البرنامج العالمي
للمناخ والبيانات المناخية

Web site of the Task Team on HOMOGENIZATION
(OPACE2, WMO Commission for Climatology)

(Last update: 30/12/2020)

<https://www.climatol.eu/tt-hom/index.html>

Homogenization software

Package	Version	License	Open source	Operating System	Program type	Primary operation	Availability
ACMANT	4	Freeware	No	DOS/Windows	Executable	Automatic	https://github.com/dpeterfree/ACMANT
AnClim ProClimDB	?	Freeware	No	Windows	Executable	Interactive (and automatic)	https://www.climahom.eu/
Climatol	3.0	GPL	Yes	(Most)	R package	Automatic	https://www.climatol.eu/index.html
GAHMDI HOMAD	?	GPL	Yes	(Most)	R source R/Fortran	Automatic Interactive	mail to andrea.toreti at giub.unibe.ch
GSIMCLI	0.0.1	GPL	Yes	(Most)	Python	Automatic (and interactive)	https://iled.github.io/gsimcli/
HOMER	2.6	GPL	Yes	(Most)	R source	Interactive	https://www.climatol.eu/pub/HOMER2.6.zip
MASH	3.03	Freeware	No	DOS/Windows	Executable	Automatic (and interactive)	https://www.met.hu/en/omsz/rendezvenyek/homogenization_and_interpolation/software/
ReDistribution Test	?	Freeware	Yes	(Most)	R source	Interactive	mail to predrag.petrovic at hidmet.gov.rs
RHtests	4	Freeware	Yes	(Most)	R source	Interactive	https://etccdi.pacificclimate.org/software.shtml
USHCN	52i	Freeware	Yes	Some linux versions	Fortran source	Automatic	ftp://ftp.ncdc.noaa.gov/pub/data/gchen/v3/software/52i/phav52i.tar.gz



Homogenization software

Package	GUI	Time resolution	Input format	Metadata use	Detection method	Ref. series selection	Detection statistic	Climatic variables
ACMANT	No	Monthly & daily	ASCII	No	Reference	Correlation	Caussinus-Lyazhi	Temperature and precipitation
AnClim ProClimDB	Yes	Any	ASCII DBF	Yes	Ref. and pairwise	Correlation & distance	Several	Any
Climatol	No	Monthly & daily	ASCII	Yes	Reference	Distance	SNHT	Any
GAHMDI HOMAD	No	Monthly Daily	ASCII	Yes	Pairwise	Correlation	New method	Any Temperature
GSIMCLI	Yes	Monthly & yearly	ASCII	No	Multiple references	Correlation & distance	User defined	Any
HOMER	No	Monthly	ASCII	Yes	Pairwise	Correlation	Penalized Likelihood	Any
MASH	No	Monthly & daily	ASCII	Yes	Multiple references	Correlation	MLR & Hypothesis test	Any
ReDistribution Test	No	Sub-daily	ASCII	No	Distribution	None	SNHT-like	Wind speed and direction
RHtests	Yes	Monthly & daily	ASCII	Yes	Reference	Correlation	Penalized max. t & F tests	Any
USHCN	No	Monthly	ASCII	Yes	Pairwise	Correlation	MLR	Temperature



Homogenization software

Package	Correction method	Missing data tolerance	Max. number of series	Outputs				Documentation
				Homogenized series	Corrected outliers	Corrected breaks	Graphics	
ACMANT	ANOVA	Very high	4000	Yes	Yes	Yes	No	User's guide
AnClim ProClimDB	Several	User defined	?	Yes	Yes	Yes	Yes	Manuals
Climatol	Missing data filling	Very high	9999*	Yes	Yes	Yes	Yes	User's guide
GAHMDI HOMAD	?	?	?	Yes	No	Yes	Yes	None
GSIMCLI	User-defined & missing data filling	High	9999*	Yes	Yes	Yes	No	Manuals
HOMER	ANOVA	15 year data	?	Yes	Yes	Yes	Yes	User's guide
MASH	Multiple comparisons	30%	500	Yes	Yes	Yes	Yes	User's guide
ReDistribution Test	None	10-20%	?	No	No	Detected breaks	No	None
RHtests	Multi-phase regression	?	1	Yes	No	Yes	Yes	User's guide
USHCN	Multiple comparisons	Very high	9999*	Yes	?	Yes	No	Plain text notes

RHtestsV4

RHtests V4

PMT and t tests:

PMF and F tests:

To adjust daily

Calculate Correlation

Transform Data

FindU.wRef

FindU

Gaussian data:

CalCol

Change Pars

FindUD.wRef

FindUD

QMadj.wRef


Quit

StepSize.wRef

StepSize

QMadj

Current Missing Value Code:	-99.9
Current nominal level of confidence (p.lev):	0.95
Segment to which to adjust the series (ladj):	10000
Current Mq (# of points for evaluating PDF):	12
Current Ny4a (max # of years of data for estimating PDF):	0
Current input Base series filename:	NA
Current input Reference series filename:	NA
Current data directory:	NA
Current output directory:	NA



THANK YOU

Driss BARI
bari.driss@gmail.com

