



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

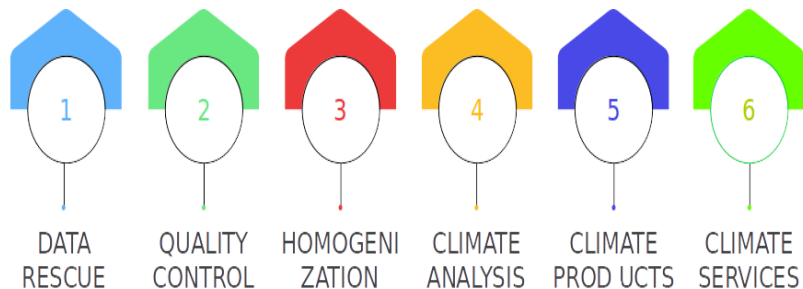
National Center of Climate
Moroccan Meteorological Service, Casablanca, Morocco
bari.driss@gmail.com

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

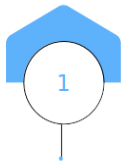
- 1 Processing of Climate Data
- 2 Climate Data Quality Control : Concepts and Tools
- 3 Climate Data Homogeneity: Concepts and Tools

- 1 Processing of Climate Data
- 2 Climate Data Quality Control : Concepts and Tools
- 3 Climate Data Homogeneity: Concepts and Tools

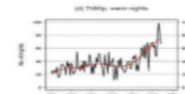
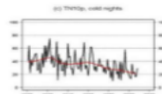
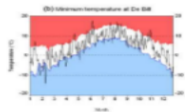
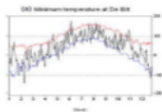
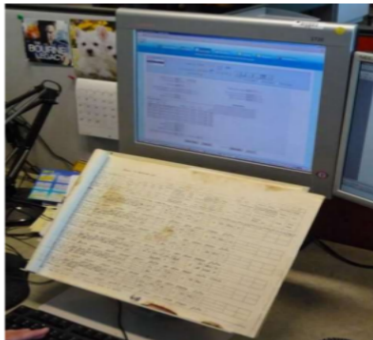
Processing of Climate Data



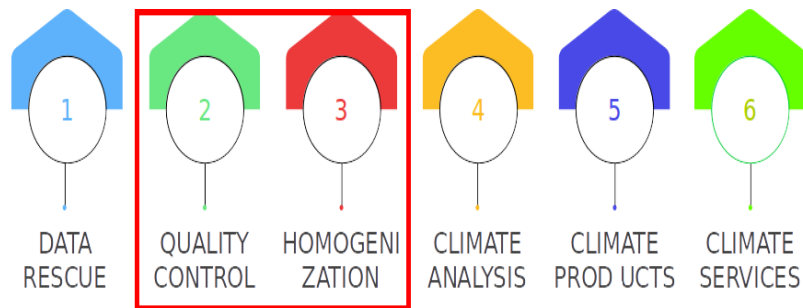
Processing of Climate Data



DATA
RESCUE



Processing of Climate Data



- 1 Processing of Climate Data
- 2 Climate Data Quality Control : Concepts and Tools
- 3 Climate Data Homogeneity: Concepts and 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

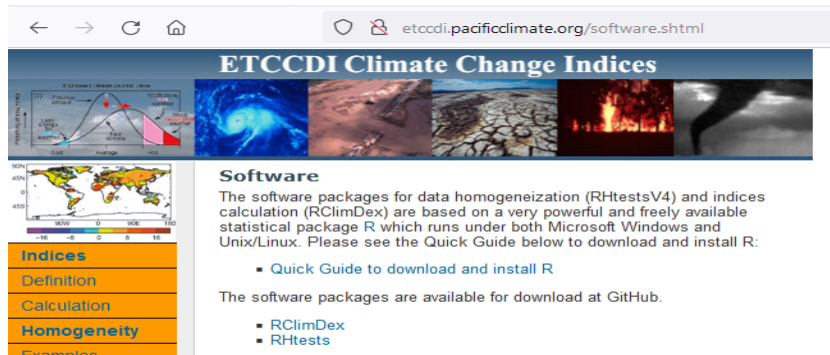
It refers to the processes for maintaining a desired level of quality in a dataset or collection. Data verification, quality control and validation are important steps in supporting defensible products and decisions. **Data quality assurance is required across the whole data life cycle and should also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

- **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.
- **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.

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)



The screenshot shows the ETCCDI Climate Change Indices website. At the top is a navigation bar with a browser address bar showing `etccdi.pacificclimate.org/software.shtml`. Below the navigation bar is a header section titled "ETCCDI Climate Change Indices" with a blue background. The header contains a graph on the left showing a bell curve with labels for "Previous climate", "Current climate", "Loss of extreme events", "Gain of extreme events", "Average", and "Risk". To the right of the graph are five images: a blue swirling cloud, a desert landscape, a cracked dry lake bed, a fire, and a dark storm cloud. Below the header is a "Software" section. It contains a paragraph about the software packages for data homogeneity (RHtestsV4) and indices calculation (RClimDex), noting they are based on the R statistical package. Below the paragraph is a bullet point: "▪ Quick Guide to download and install R". Further down is another paragraph stating the software is available for download at GitHub, followed by two bullet points: "▪ RClimDex" and "▪ RHtests". On the left side of the website, there is a vertical menu with orange buttons labeled "Indices", "Definition", "Calculation", "Homogeneity", and "Examples". The "Indices" button is highlighted in blue.

ETCCDI Climate Change Indices

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

Indices

Definition

Calculation

Homogeneity

Examples

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

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

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

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.

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

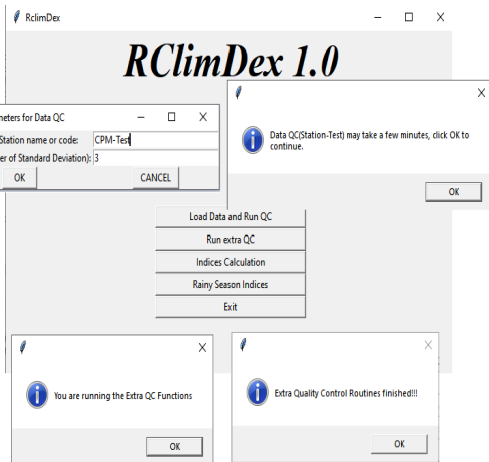
RclimDex-extraqc

 **Manual**



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](#)

[Download](#)



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

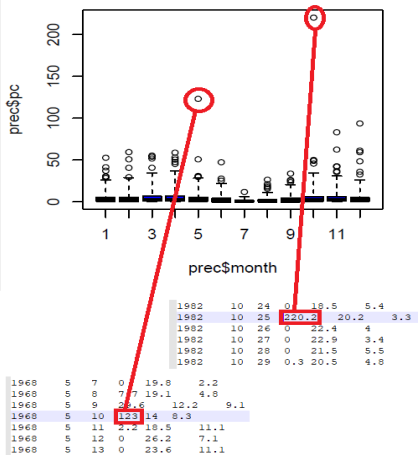
2015	12	2	0	19.4	4.1
2015	12	3	0	20.6	99.9
2015	12	4	0	19.4	5.9
2015	12	5	0	19.6	3.1

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

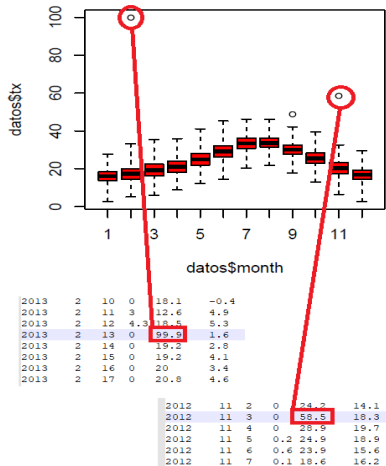
1982	10	25	220.2	20.2	3.3
2012	11	3	0	58.5	18.3
2013	2	13	0	99.9	1.6
2015	12	3	0	20.6	99.9

EXTRAQC : Example

NON ZERO PREC



TX



- 1 Processing of Climate Data
- 2 Climate Data Quality Control : Concepts and Tools
- 3 Climate Data Homogeneity: Concepts and Tools

Background on Homogeneity

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.

Background on Homogeneity

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.

Background on Homogeneity

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 (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.**

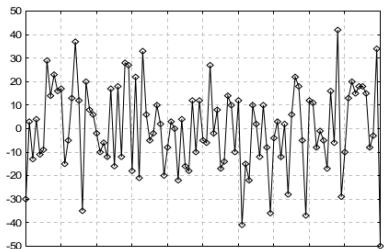
2. Creation of a reference time series

- 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

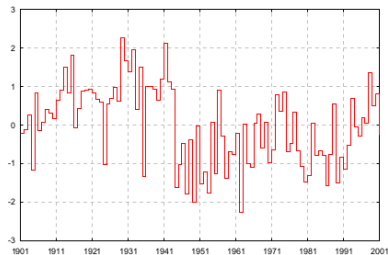
- The idea is to create **differences** (temperature) or **ratios** (precipitation) of a **candidate** series (the one we want to homogenize) towards a **reference**
- The candidate and the reference share the same climate so **the odd features** of the candidate minus reference **are not due to the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



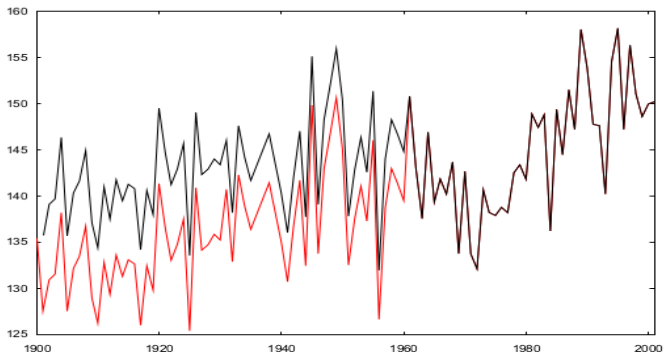
The difference between candidate and reference time series (bottom) clearly shows an inhomogeneity in 1941, documented in the metadata as a relocation. The original data (top) mask the inhomogeneity.

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

4. Data adjustment

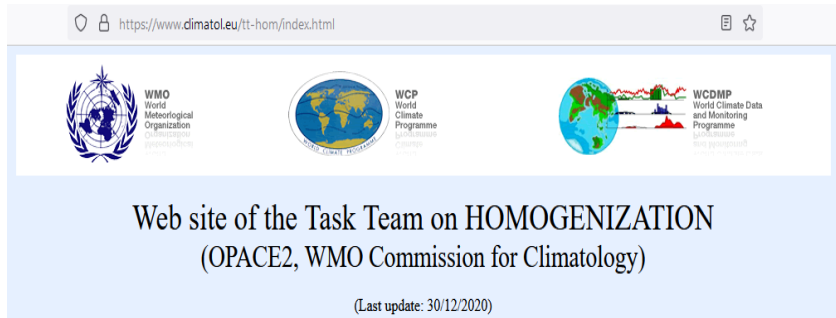
- Once the breakpoint identification is finished, the next step is to decide which breakpoints are going to be accepted as real inhomogeneities.
- Data adjustment is **the correction applied to data to improve their homogeneity and to make all the observation comparable to the last available data.**
- 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 reference 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.

Homogenization software



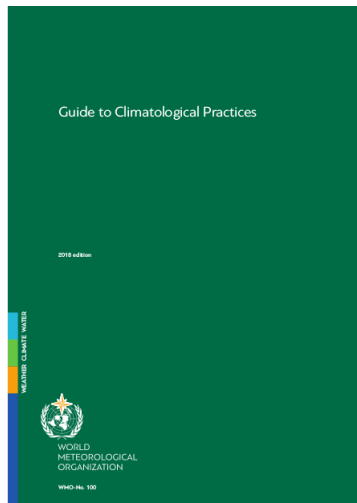
<https://www.climatol.eu/tt-hom/index.html>

Homogenization software

Overview of the characteristics of homogenization packages

Package	Resolution detection ^a	Detection method	Reference use ^b	Resolution correction ^c	Correction method	Primary operation	Metadata use	Variable ^d	Documentation ^e	Reference
ACMANT ^f	Year, month	Multiple breakpoints	Composite	Year, month, [day]	Joint (ANOVA)	Automatic	No	Any	User guide	Domonkos and Coll (2017)
AnClim	Year, month	Many	Composite, pairwise	Year, month, day	Several	Interactive, automatic	Yes	Any	Manuals	Štěpánek et al. (2009)
Berkeley Earth	Month	Splitting	Composite	n/a	n/a	Automatic	Yes	T	Article	Rohde et al. (2013)
Climatol	Month (serial), day	Splitting	Composite	Month (serial), day	Missing data filling	Automatic	Yes	Any	Manual and user guide	Guijarro (2018)
GAHMDI HOMAD	Month (serial), day	Multiple breakpoints	Selection	Day	Higher-order moment method	Automatic	Yes	T	None	Toreti et al. (2010, 2012)
GSIMCLI	Year, month	Multiple breakpoints	Composite	See footnote ^g	See footnote ^g	Automatic and interactive	No	T, p	Manuals	Ribeiro et al. (2017), Costa and Soares (2009)
HOMER	Year, season, month	Multiple breakpoints	Pairwise, joint	Year, month	Joint (ANOVA)	Interactive	Yes	Any	Basic user guide+courses	Mestre et al. (2013)
iCraddock	Year, season, month	Splitting	Pairwise	Year, season, month, day	Daily: smoothed monthly corrections	Interactive	Yes	Any	None	Craddock (1979), Brunetti et al. (2006)
MASH	Year, season, month	Multiple breakpoints	Composite	Month, [day]	Multiple comparisons	Automatic and interactive	Yes	Any	User guide	Szentimrey (2008, 2014)
ReDistribution Test	Readings	Single breakpoints	No reference	n/a	n/a	Interactive	No (but it is interactive)	Wind	None	Petrovic (2004)
RHtests	Year, month, day	Splitting	Selection or no reference	Year, month, day	Multi-phase regression	Interactive	Yes	Any	User guide+courses	Wang (2008a and b), Wang and Feng (2013)





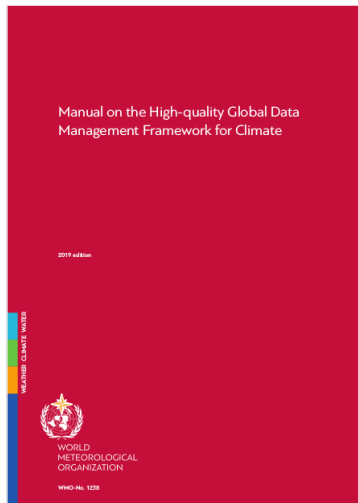
World Meteorological
Organisation

WMO No. 100

Guide to Climatological Practices

Edition 2018

[https://library.wmo.int/
doc_num.php?explnum_id=5541](https://library.wmo.int/doc_num.php?explnum_id=5541)



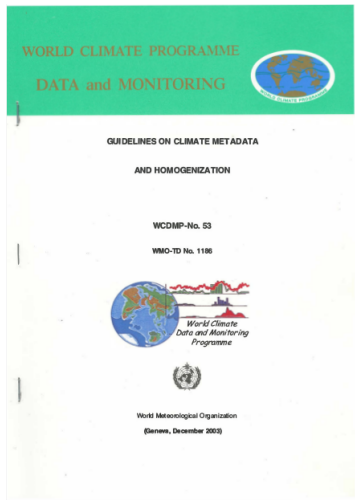
World Meteorological
Organisation

WMO No. 1238

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

Edition 2019

[https://library.wmo.int/doc_num.
php?explnum_id=21686](https://library.wmo.int/doc_num.php?explnum_id=21686)



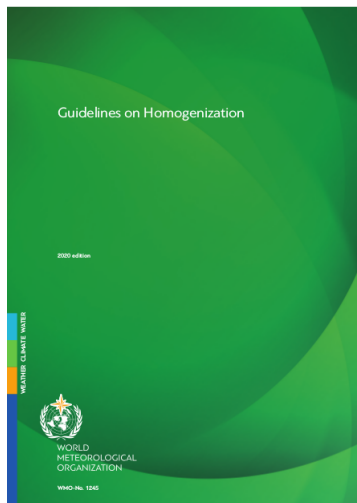
World Meteorological
Organisation

WMO-TD No. 1186

Guidelines on Climate Metadata and Homogenization

Edition 2003

[https://library.wmo.int/doc_num.
php?explnum_id=11635](https://library.wmo.int/doc_num.php?explnum_id=11635)



World Meteorological
Organisation

WMO No. 1245

Guidelines on Homogenization

Edition 2021

https://library.wmo.int/doc_num.php?explnum_id=21756

THANK YOU

Driss BARI
bari.driss@gmail.com