The file format for the parallel data initiative

Victor Venema and Renate Auchmann

Preliminary discussion version: 29 August 2014

This document describes the file format and the directory structure of the database with parallel climatological and meteorological measurements of the parallel data initiative. The file format is based on the format of the COST action HOME.

The main design considerations for the data format are as follows:

We would like to have one file per measurement (because a parallel may have a varying number of measurements of the same variable per station and multiple measured variables).

For clarity, the data (with flags) and the metadata will be in separate files. Next to the parallel measurements, also related measurements should be stored. For example, to understand the differences between two temperature measurements, additional measurements (co-variates) on, for example, insolation, wind or cloud cover are important. And to understand the differences between two rain gauges, information on the wet-bulb temperature and wind are helpful.

For ease of reading the data and give the limited manpower at the moment, an ASCII format will be used and the database will simply be a directory tree. Later a scientific data format such as NetCDF may be added. A real database may be useful, especially to make it easier to select the parallel measurements one is interested in, but is too ambitious for the moment.

Also metadata needs to be stored and should be machine readable as much as possible. Without meta-information on how the parallel measurement was performed, the data is not useful. A description of the data format for the metadata will follow; we are still working on that. This will be included in a later version of this document.

We are interested in parallel data from any source, variable and temporal resolution. High resolution (sub-daily) data is very important for understanding the reasons for any differences. There is probably more data, especially historical data, available for coarser resolutions and this data is important for studying non-climatic changes in the means.

However, we will scientifically focus on changes in the distribution of daily temperature and precipitation data in the climate record. Thus, we will compute daily averages from sub-daily data and will use these to compute the indices of the Expert Team on Climate Change Detection and Indices (ETCCDI), which are often used in studies on changes in “extreme” weather.

Following the principles of the ISTI, we aim to be an open dataset with good provenance, that is, it should be possible to tell were the data comes from. For this reason, the dataset will have levels with increasing degrees of processing, so that one can go back to a more primitive level if one finds something interesting/suspicious.

For this same reason, the processing software will also be made available and we will try to use open software (especially the free programming language R, which is widely used in statistical climatology) as much as possible.

It will be an open dataset in the end, but as an incentive to contribute to the dataset, initially only contributors will be able to access the data. After joint publications, the dataset will be opened for academic research as a common resource for the climate sciences. In any case people using the data of a small number of sources are requested to explicitly cite them, so that contributing to the dataset also makes the value of making parallel measurements visible.

1. Directory structure

In the main directory there will be sub-directories called: `data`, `documentation,` and `software` and in the end also: `results` (quicklooks and tables) and `articles`.

In the sub-directory `data` (and `results`) there are sub-directories for the data sources with names `d###`; with `d` for **d**ata source and ### is a running number of arbitrary length.

In these `d###` directories there are up to 5 sub-directories with the data in various levels of processing (see below) and one directory with "additional" metadata such as photos and maps that cannot be copied in every level directory.

In the level 0 and level 1 directories, the data native files are directly stored in this directory. However, because one data source can contain more than one station, in the levels 2 and higher there are sub-directories for the various stations. These sub-directories will be called `s###`; with `s` for **s**tation.

The parallel data can be available at 5 levels, which are detailed in the sub-sections below:
0: Original, raw data (e.g. images)
1: Native format data (as received)
2: Data in a standard format at the original resolution
3: Daily data
4: ETCCDI indices

In levels 2, 3 & 4 we will provide information on outliers and inhomogeneities. Especially for the study of extremes, the removal of outliers is important.

Longer parallel measurements may contain inhomogeneities. We will thus detect breaks and provide their date and size as metadata, so that the user can work on homogeneous subperiods if interested. This detection will probably be performed at monthly or annual scales with one of the HOME recommended methods.

Because parallel measurements will tend to be well correlated, it is possible that statistically significant inhomogeneities are very small and climatologically irrelevant. Thus we will also provide information on the size of the inhomogeneity so that the user can decide whether such a break is problematic for this specific application or whether having longer time series is more important. In case of two measurements, we can only provide the information that one of them has an inhomogeneity. In case of more than two measurements, we will try to attribute the inhomogeneities to a specific time series.

1.1 Level 0 - images

If possible, we will also store the images of the raw data records. This enables the user to see if an outlier may be caused by unclear handwriting or whether the observer explicitly wrote that the weather was severe that day.

In case the normal measurements are already digitized, only the parallel one needs to be transcribed. In this case the number of values will be limited and we may be able to do so. Both Bern and Bonn have facilities to digitize climate data.

1.2 Level 1 – native format

Even if it will be more work for us, we would like to receive the data in its native format and will convert it ourselves to a common standard format. This will allow the users to see if mistakes were made in the conversion and allows for their correction. Submitting data in both the native and the standard format is naturally also possible and appreciated. This directory may contain sub-directories, for example if the original data is in Microsoft XLS, a subdirectory may exists with text files of these tables.

1.3 Level 2 – standard format

In the beginning our standard format will be an ASCII format. The format is described in §2 below. Later on we may also use a scientific data format such as NetCDF. This directory may contain sub-directories. For example directories with monthly or annual means for use by the homogenization methods.

1.4 Level 3 - daily data

We expect that an important use of the dataset will be the study of non-climatic changes in daily data. At this level we will thus gather the daily datasets and convert the sub-daily datasets to daily.

1.5 Level 4 – ETCCDI indices

Many people use the indices to the ETCCDI to study changes in extreme weather. Thus we will pre-compute these indices. Also in case government policies do not allow giving out the daily data, it may sometimes be possible to obtain the indices.

2. Filename and format for data and quality flags

2.1. Filenames

The filename should be formatted as follows:

pxxvvnnrssssssss.index.txt

For example: pratm04m0306001.mean.txt

p: File contains parallel data.
ra: Raw data (with possible inhomogeneities).
tm: Variable is mean temperature.
04: Running number (this is the fourth setup that measures tm)
m: The file format of the data is monthly
03060001: The station is Aerodrome de Vichy-Charmeil, Charmeil, france.
mean: The file contains average values (other indices also possible in level 4)

Parallel data (p)

We will make sure that no other files will be called p*.txt to make it easier to read all data files in one directory. File ending should be .txt for ease of working with graphics file managers and for file transfers with FTP.

Data status, quality (xx)

These two letters indicate the processing status of the file:
ra: stands for raw data (with possible outliers, missing data and inhomogeneities),
qc: for quality controlled (outliers detected and mentioned in the metadata or corrected),
ho: for homogenized (i.e. outliers removed, missing data filled and detected inhomogeneities removed).
re: for co-variates from reanalysis (future plans).

The status of the parallel measurements will generally be ra or qc, but additional measurements that may explain the differences between the parallel measurements may have been homogenized by the data provider. The status relates to the full dataset. Thus, if more than a few recent years have not been homogenized, the status is quality controlled or raw.

Measured variable (vv)

This section indicates the meteorological variable that is in the data file. We will try to follow standard two-letter conventions, see list below. If abbreviations for other variables are needed, please contact us.

dd: wind direction
ff: wind speed
nn: cloud cover
tm: mean temperature
tn: minimum temperature
tx: maximum temperature
pp: pressure
rr: precipitation
sn: sunshine duration
sd: solar radiation flux down
su: solar radiation flux up
hd: infra-red (heat) radiation flux down
hu: infra-red (heat) radiation flux up

Running number for multiple measurements (nn)
Because there will be multiple measurements of the same variable, which would otherwise have the same filename, the filename also includes a running number. For the parallel measurements they will start with 01, 02, 03, etc. The related measurements (co-variates) will be indicated by 00.

Temporal data format (`r`)
To indicate the data format (see section 2.2) the following letter convention is used. In most cases this will define the averaging period or the frequency with which an index was computed, but for fixed hour measurements (for example, *Mannheimer* hours) the time may be known with more precision as hourly. Please also consider that it is possible that not all files may have the same temporal resolution.

`y`: yearly
`m`: monthly
`d`: daily
`h`: hourly
`M`: minutes
`s`: seconds

Station number (`ssssssss`)
The station number should be specified with 8 digits. Preferably this is the original WMO station number, as this avoids duplicates. If the station does not have a WMO station number, it can also be an internally used number. If this internal number is short enough, the WMO national number is added to obtain a unique number. If the number is still less than 8 digits, zeroes are added in front.

Indices (.index)
From one measurement multiple indices can be computed. We would like to compute the indices of the ETCCDI: FD, SU, ID, TR, GSL, TXx, TNx, TXn, TNn, TN10p, TX10p, TN90p, TX90p, WSDI, CSDI, DTR, Rx1day, Rx5day, SDII, R10mm, R20mm, Rnnmm, CDD, CWD, R95pTOT, R99pTOT, PRCPTOT. (In Rnnmm, the nn stands for an arbitrary threshold of the rain rate) See ETCCDI (2014) for their definitions. To be able to stick to their abbreviations, this part of the file name has a flexible length, in contrast to the rest. In addition to the ETCCDI indices, also percentiles may be computed and indicated as Pnn, with nn indicating which percentile.

Next to these indices, there will be MEAN values (e.g. monthly and yearly means), READ values (value read at specific time, e.g. minimum and maximum temperature and fixed hour measurements), and SUM (for example for daily precipitation sums).

2.2. Climate record data file format

All files are tab-delimited, i.e. between the values, a tab is used as delimiter (no spaces in between, no fixed column width).

The time of the measurement is indicated by integers only; only columns with the time in seconds may (theoretically) be a float. The first column is the year. Consecutive columns are added if necessary: month, day of month, hour, minutes, seconds.

Data should be formatted as a float and put in the second last column. Missing data is indicated with the value: -999.9. The last column is the flag value (an integer, see below).

There will be no missing data periods within the measurement period. We aim to make all files the same length, but in case of large differences in begin or end times, they may differ from measurement to measurement.

2.3. Quality flag file format

The quality flag file has the same format as the data file format, only the data values are substituted by *integer* quality flags. If more than one flag applies, please add the values. The quality flag are:

0 raw (means: not passed QC)
1 missing value
2 outlier detected
4 passed QC
8 reconstructed missing value or outlier
16 homogenized
32 exceptional value. In some cases, the original data is not an error, but has exceptional values and should not be removed in the QC. This may occur for violent, but local thunderstorms for example.

3. References

ETCCDI, 2014: Climate Change Indices, Definitions of the 27 core indices. http://etccdi.pacificclimate.org/list_27_indices.shtml, retrieved on the 28th August 2014.