Homogenization of climatic series with Climatol

Version 3.1.1

https://CRAN.R-project.org/package=climatol

Jose A. Guijarro

State Meteorological Agency (AEMET), Balearic Islands Office, Spain

Version of this guide: 1.3.1 (August 2018)

Versión española disponible en http://www.climatol.eu/homog_climatol-es.pdf

Contents

1. Introduction

The series of meteorological observations are of capital importance for the study of climate variability. However, these series are frequently contaminated by events unrelated to that variability: errors in the observations or in their transmission, and changes in the instrumental used, in the location of the observatory or in its environment. These last they can produce sudden changes, like a fire burning an adjoining forest, or gradual, as the subsequent recovery of vegetation. These alterations of the series, called inhomogeneities, mask the real changes of climate and may mislead the conclusions derived from the study of the series.

This problem has been addressed many years ago by developing homogenization methodologies that allow to eliminate or reduce as much as possible these unwanted alterations. Initially they consisted of comparing a problem series with another supposedly homogeneous, but as this assumption is very risky, many methods began building composite reference series, by averaging others selected for their proximity or high correlation, thus diluting its possible inhomogeneities. As this does not guarantee the homogeneity of the composite reference, other methods proceed to compare all series available in pairs, so that the repeated detection of a inhomogeneity allows to identify which is the erroneous series. Reviews of these methodologies can be seen in the works by Peterson et al. (1998) and Aguilar et al. (2003).

There are many software packages that implement these methods so that they can be used by the climatological community (http://www.climatol.eu/tt-hom/index.html). The COST Action ES0601 (Advances in homogenisation methods of climate series: an integrated approach, HOME) funded an international effort to compare them (Venema et al., 2012). Later the MULTITEST project (http://www.climatol.eu/MULTITEST/) made another comparison of the updated methods that could be executed in fully automatic mode. Homogenization efforts had been focused on monthly series so far, mainly of temperature and precipitation, but there has been a growing interest in addressing the homogenization of daily series, necessary for the study of the variability of the extreme phenomena, and currently the European INDECIS project is trying to apply several methods to daily series of diverse climatic variables.

The R package *Climatol* (https://CRAN.R-project.org/package=climatol) contains functions for quality control, homogenization and infilling of the missing data in a set of series of any climatic variable. The standard documentation of the package provides detailed information about each of its functions and their control parameters, as well as brief examples of their application. This manual is a complement to that documentation which, without giving many details about each of the available functions, explains the fundamentals of the methodologies used, and then provides a practical guide on how to approach the homogenization of daily or monthly series of different variables.

2. Methodology

In its beginnings, this program was focused on infilling the missing data by estimates calculated from the closest series. This was done by adapting the method from Paulhus and Kohler (1952) to infill daily rainfall data by averaging neighboring values, normalized by dividing them by their respective average rainfall. This method was chosen for its simplicity and for allowing the use of nearby series even if they did not have a common period of observation with the problem series, which would preclude the adjustment of regression models.

In addition to normalizing the data through a division by their average values, *Climatol* also offers the possibility of subtracting the means or applying a full standardization. So, letting m_X and s_X be the average and standard deviation of a X series, we have these options for their normalization:

1. Remove the mean: $x = X - m_X$
2. Divide by the mean: $x = X/m_X$
3. Standardize: $x = (X - m_X)/s_X$

The main problem with this methodology is that means (and standard deviations in the third case) of the series in the study period are unknown when the series are not complete, which is most often the case in real data bases. Then *Climatol* first calculates these parameters with the available data in each series, infill the missing data using these provisional averages and standard deviations, and recalculates them with the infilled series. Then data the originally missing data are recalculated using the new parameters, which will lead to new means and standard deviations, hence repeating the process until no average changes when rounded up to the initial precision of the data.

Once the means become stable, all data are normalized and estimated (whether existing or missing, in all of the series), by means of the simple expression:

$$\hat{y} = \frac{\sum_{j=1}^{j=n} w_j x_j}{\sum_{j=1}^{j=n} w_j}$$

in which $\hat{y}$ is a data item estimated from their corresponding nearest n data available at each time step, and w_j is the weight assigned to them.

Statistically, $\hat{y}_i = x_i$ is a linear regression model called *Reduced Major Axis* or *Orthogonal Regression*, in which the line is adjusted by minimizing the distances of the points measured perpendicular to it (type II regression) instead of place in the vertical direction (type I regression) as it is usually done (figure 1), whose formulation (with standardized series) is $\hat{y}_i = r \cdot x_i$, where r is the correlation coefficient between the series x e y. Note that this type of adjustment is based on the assumption that the independent variable x is measured without error (Sokal and Rohlf, 1969), assumption that does not hold when both are climatic series.

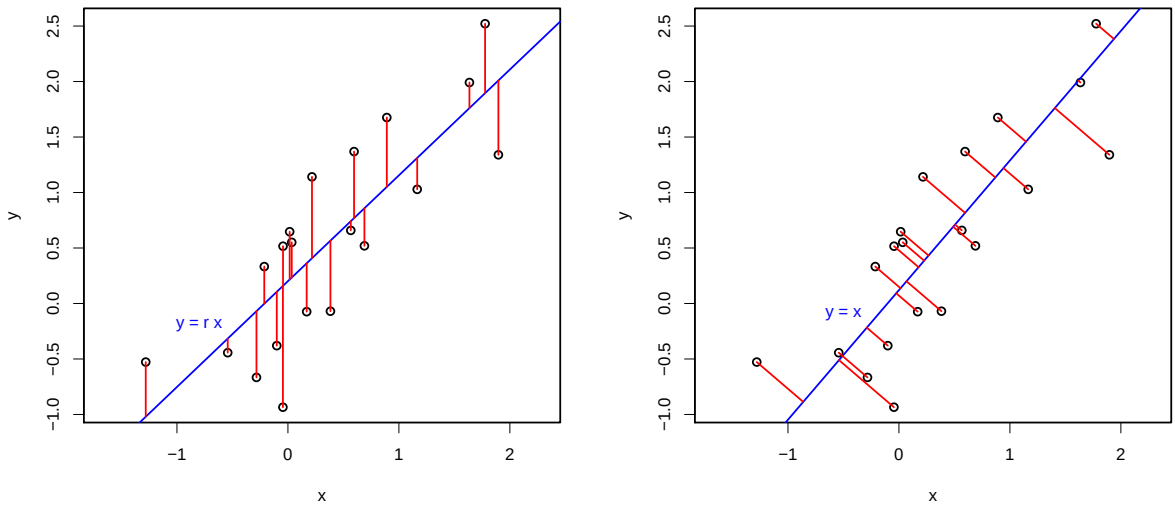

Figure 1: In red, deviations from the regression line (blue) minimized by least squares in regression types I (left) and II (right).

The series estimated from the others serve as references for their corresponding observed series, so the next step is to obtain series of anomalies by subtracting the estimated values from the observed ones (always in normalized form). These series of anomalies will allow:

- Control the quality of the series and eliminate those anomalies that exceed a preset threshold.
- Check their homogeneity by applying the *Standard Normal Homogeneity Test* (SNHT: Alexandersson, 1986).

When the SNHT statistics of the series are greater than a prescribed threshold, the series is split at the point of maximum SNHT, moving all data before the break to a new series that is incorporated into the data pool with the same coordinates but adding a numerical suffix to the code and name of the station. This procedure is done iteratively, splitting only the series with the higher SNHT values at every cycle, until no series is found inhomogeneous. Moreover, as SNTH is a test originally devised for finding a single break-point in a series, the existence of two or more shifts in the mean of similar size could mask its results. To minimize this problem, SNTH is applied in a first stage to stepped overlapping temporal windows, and after that a second stage is devoted to apply SNHT on the complete series, which is where the test exhibits more power of detection. Finally, a third stage is dedicated to infill all missing data in all homogeneous series and sub-series with the same data estimation procedure. Therefore, although the underlying methodology of the software is very simple, its operation becomes complicated through a series of nested iterative processes, as shown in the flow-chart displayed in Figure 2, and the processing time can vary from seconds to hours (or even days, when dealing with hundreds of stations and many decades of daily data).

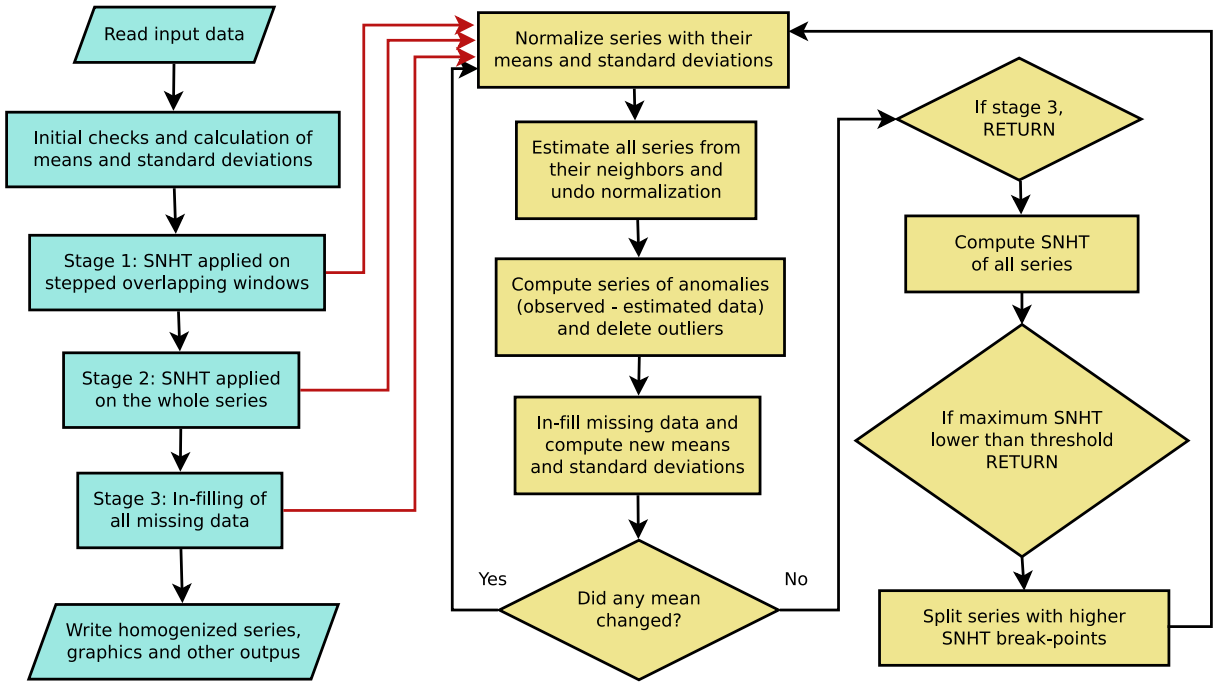

Figure 2: Flow-chart of the *Climatol* operation, showing its nested iterative processes.

Although SNHT thresholds have been published for different series lengths and levels of statistical significance, experience shows that this test can give very different values depending on the climatic variable studied, the degree of correlation between the series and their temporal frequency. *Climatol* uses a default value of SNHT = 25, appropriate for monthly values of temperature, although a little conservative, trying not to detect false jumps in the average at the expense of disregarding those of minor importance. However, for other variables and in particular for daily values, it is necessary to raise considerably that threshold to avoid an excessive number of splits in the series. The same happens with the threshold to reject anomalous data, established by default in 5 standard deviations, since with daily data of precipitation, given its great spatial variability, 20 or more may be needed. Therefore, instead of setting these thresholds according to levels of significance, impossible to establish with general validity, users have the option to choose them subjectively, after inspecting the histograms of the values found after a first application of *Climatol* to their concrete problem.

3. Homogenization procedures

After having exposed the methodology followed by the *Climatol* package, this section will be dedicated to illustrate its practical application through some examples.

3.1. Preparation of the input files

Climatol only needs two input files, one with a list of coordinates, codes and names of the stations, and another with the all the data, in chronological order and from the first station to the last one. As no temporal references appear in the data file, all data must be present, for the whole period of study, with missing data represented with `NA` or with other distinctive code. Moreover, to avoid complications, the period of study should begin in January of the first year (in day 1 when dealing with daily data) and end in December of the last year (in day 31 when processing daily data). Both files share the base name `VAR_yyyy-YYYY` where `VAR` is an acronym of the name of the studied variable, `yyyy` the first year and `YYYY` the last year of the data, but they have different extensions: `dat` for the data and `est` for the stations. Both are plain text files, so Windows users can associate to open them with Notepad or any other plain text editor. (If you edit them with LibreOffice or Word, take care to save them as plain text files to avoid problems.)

Only for the purpose of running the following examples, these files can be generated in the working directory by means of these commands (anything after # is a comment):

```
library(climatol) # load the functions of the package
data(Ttest) #load the example data into R memory space
write(dat, 'Ttest_1981-2000.dat') #save the data file
#save the stations file:
write.table(est.c, 'Ttest_1981-2000.est', row.names=FALSE, col.names=FALSE)
rm(dat, est.c) #remove the loaded data from memory space
```

These files contain 20 years of daily test temperature data for 12 invented stations. They can be inspected to see their structure. The first lines of the station file `Ttest_1981-2000.est` are:

```
-108.035 44.38 1169.5 "WY003" "Small Horn"
-108.9006 44.4139 1599.6 "WY018" "Narrow Canyon"
-108.5931 44.8919 1251.2 "WY020" "Wide Meadows"
-108.3906 44.4972 1355.8 "WY027" "Greenbull"
```

As you can see, every line has, in free space separated format, the coordinates X, Y, Z of the station, followed by the code and the name. Normally X and Y are the longitude and latitude, in degrees with decimals (not in degrees, minutes and seconds) and with appropriate signs to indicate West, East, North or South. Z is the elevation in meters.

The first lines of the data file `Ttest_1981-2000.dat` are:

```
-1.8 2.7 0.4 8 2.4
 1.4 1.2 3.3 1.5 0.7
```

```
-0.8 -0.6 4 2.6 -1.6
-4.8 -3.1 -0.8 -0.6 -4
```

These 20 data are the mean temperatures of the first 20 days of January 1981 in the first station (Small Horn). The following lines of the file contains the remaining data of this station until December 31, 2000, followed by all data for the other stations listed in the `Ttest_1981-2000.est` file.

To help preparing the input files with this format, *Climatol* provides some utility functions (see the R documentation for details about their use):

- `db2dat` generates the files by retrieving the series from a database through an ODBC connection.
- `daily2climatol` can be used when every station has their daily data stored in individual files.
- `rclimdex2climatol` can convert files in RClimDex format.

3.2. First exploratory analysis of the data

The homogenization function of *Climatol* is called `homogen`, and its most trivial application only requires the specification of three parameters: the variable acronym, and the initial and final years of the studied period:

```
homogen('Ttest', 1981, 2000)
```

This command can be applied whether the data are daily, monthly, bimonthly, quarterly, semi-annual or annual: the function will guess the frequency from the number of data in the file. But as explained in the methodology section, thresholds for outlier rejection and break-point detection can be very different depending on the data periodicity and cross-correlation of the series. Therefore, it is advisable to make a first run in exploratory mode:

```
homogen('Ttest', 1981, 2000, expl=TRUE)
```

Now we can open the output file `Ttest_1981-2000.pdf` to revise its various diagnostic graphics. First we see the data availability, in all stations and globally (Figure 3). Ideally, there should be 5 or more data available at every time step, or a minimum of three, levels marked with dashed green and red lines in the right part of the figure, but the function will work except when no data is available in any station at one or more time steps, situation that halts the process with an error message. In this case, series with data at that "orphan" time steps should be added to the data-base, or the period of study should be reduced to avoid that condition.

When working with zero-limited variables with a skewed probability distribution (as precipitation or wind speed), the average ratio normalization (set with `std=2`) is preferred to the default standardization.

It is important to run these exploratory analysis on the original data for a reliable quality control, since the detection of outliers in derived series may mask the observation errors. For example, if there is a 10°C error in a daily maximum temperature, it will be lowered to 5°C in the daily mean if calculated as $(T_{max} + T_{min})/2$, and to around 10/30=0.33°C in the monthly maximum mean or 5/30=0.17°C in the monthly mean.

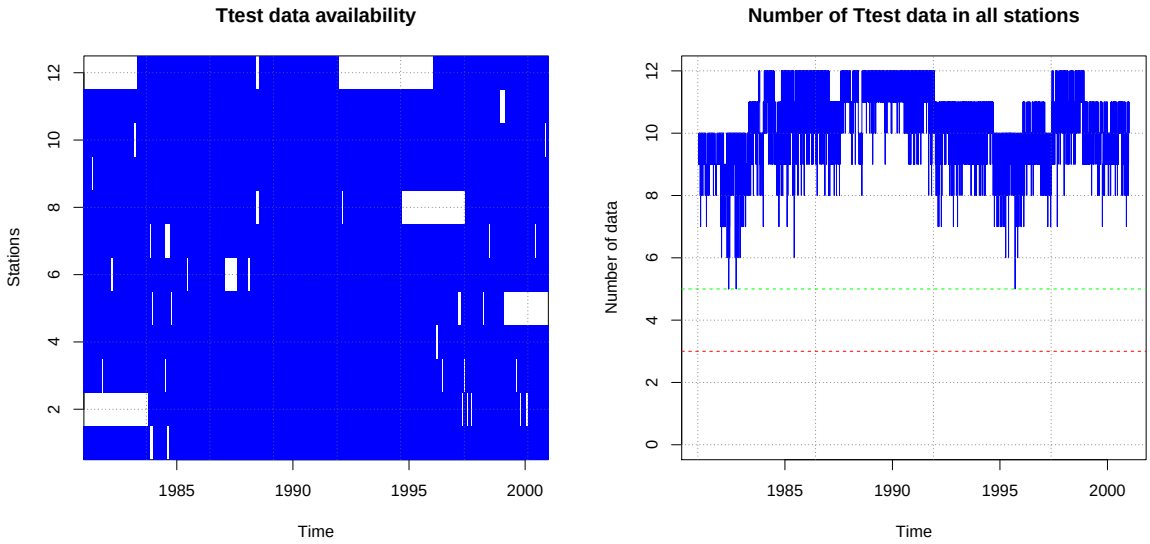

Figure 3: Data availability, by stations (left) and globally (right).

The following graphics show box-plots of the data at every station and a histogram of all data. The presence of very anomalous values would be evident in these plots, allowing the user to take corrective actions. Also the frequency histogram will be useful to decide if the probability distribution is near normal or very skewed. In the second case, it can be preferable to use the normal ratio normalization of the data (using the parameter `std=2`) rather than the default standardization.

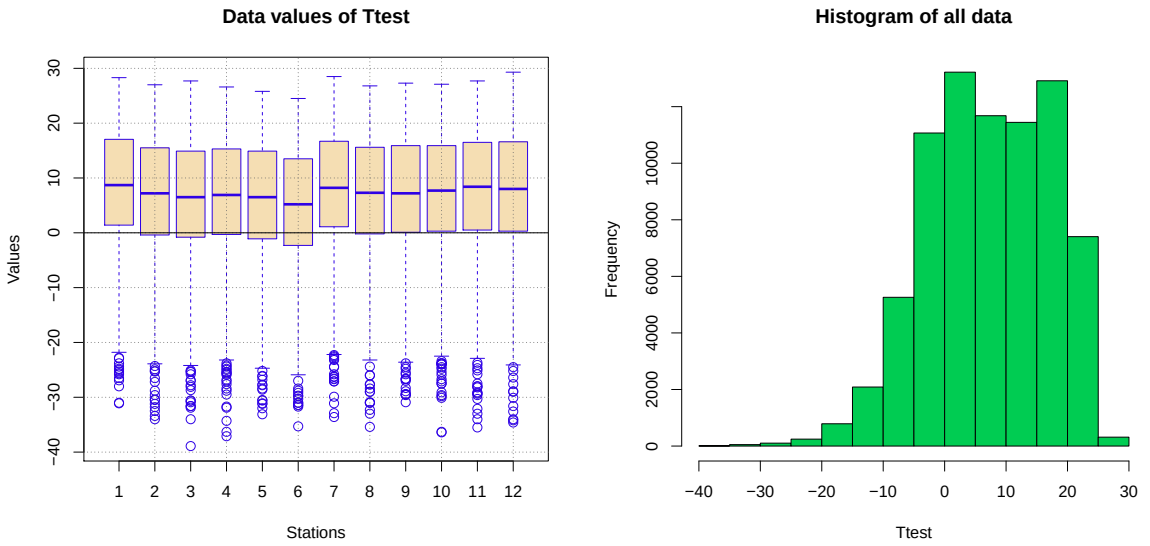

Figure 4: Box-plots of the data at every station (left) and histogram of all data (right).

The next graphics are focused on the correlations between the series and their classification in groups of similar variability, which are plotted on a map (Figure 5). Correlations are generally lower when the distance between stations is greater, as in this example. The higher the correlations, the higher the reliability of the homogenization and missing data infilling. In particular, correlations should always be positive, at least within reasonable distances. Otherwise there are probably geographic discontinuities producing climate differences (e.g., a mountain ridge can produce opposite precipitation regimes). This can be confirmed by the map of stations, where groups of similar variability would be located in distinct areas. In areas of complex topography and/or low density of stations, correlations may be far from optimal. In this situation, individual estimated missing data will be affected by important errors, but their overall statistical properties are expected to be acceptable.

To avoid processing too large correlation matrices, the number of series used for this cluster analysis is limited by default to 100, and a random sample of this size is used when the number of series is greater, but the user can change this number.

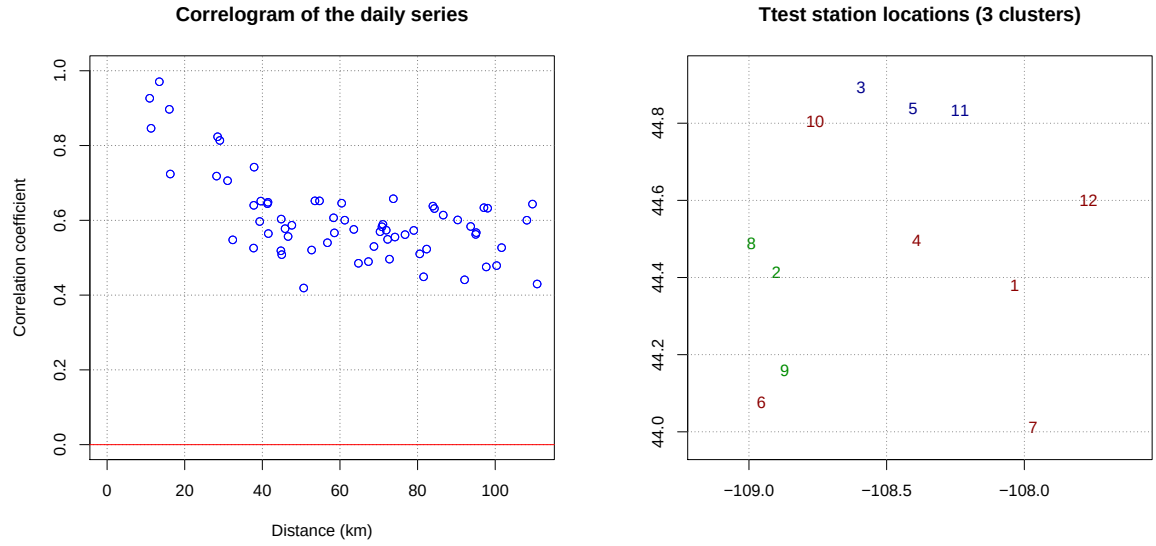

Figure 5: Correlogram of the series (left) and map of the stations (right; colors identify groups of stations with similar variability).

After these initial graphics dedicated to check the data, the following pages of the document display plots of standardized anomalies. In normal operation this plots are shown for each of the stages: 1, detection on overlapping windows; 2, detection on the whole series; and 3, final anomalies of the homogenized series. The plots of the first two stages show the anomalies series of the detected inhomogeneities, marking the break-points where the series are split, but in this exploratory mode the first two stages are skipped, and only the anomalies of all original series are shown.

Figure 6 displays two of these plots. The series of the left one seems quite homogeneous, with a maximum SNHT of 12 over stepped overlapping windows marked in green over a dashed line of the same color at the point where that maximum is reached, and a maximum SNHT of 17 over the whole series under a black line at its corresponding time step. On the contrary, the series on

the right is clearly inhomogeneous, with maximum SNHT of 117 and 1561 reached at the same point. Two additional lines at the bottom inform about the minimum distance of neighbor data (in green) and the number of used reference data (in orange), both using the logarithmic scale on the right axis.

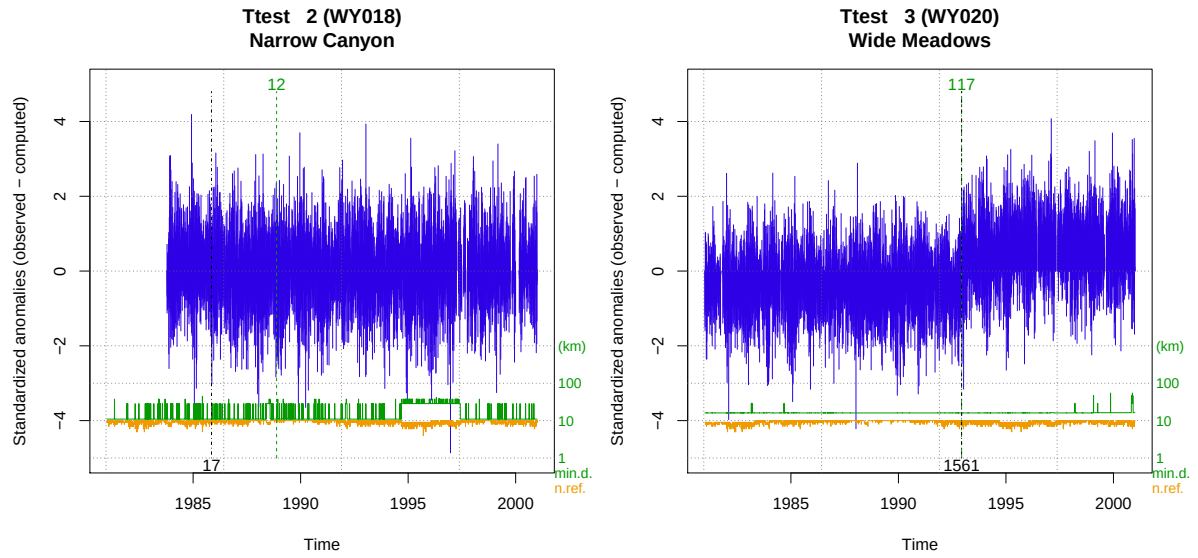

Figure 6: Anomalies of a homogeneous series (left) and a very inhomogeneous one (right).

After the graphics of anomalies you can find the plots of adjusted series and applied corrections, but as no modification is done to the series in exploratory mode apart from infilling all their missing data, these graphics will be explained later on.

The graphics document ends with histograms of standardized anomalies and SNHT of the final series, and a plot indicating their quality or singularity. The histogram of anomalies (Figure 7) helps in choosing adequate thresholds to reject very anomalous data, assuming they are errors and can be deleted. Our example histogram show some skewness to the left, but it is not very pronounced and therefore all data could be accepting by setting `dz.max=9`, since the default value would deleted those data with absolute anomalies greater than 5 standard deviations.

The histograms of maximum SNHT (whether windowed or complete) are intended to choose the detection thresholds of changes in the mean of the series. If we were processing a big number of series, these histograms would show a high frequency of low values, corresponding to the series fairly homogeneous, and one or more secondary groups of bars due to the inhomogeneous cases. When there is a gap (or a clear minimum) separating these conditions, it is very easy to set a value between them as the threshold for the detection stages. In our case, with only 12 series, frequency bars are separated by several gaps, making the decision difficult. For the windowed stage, setting `snht1=60` seems reasonable, but it is far from clear in the histogram of SNHT applied on the complete series. In this case, visual inspection of the plots of anomalies can help choosing `snht2=70` as an adequate value.

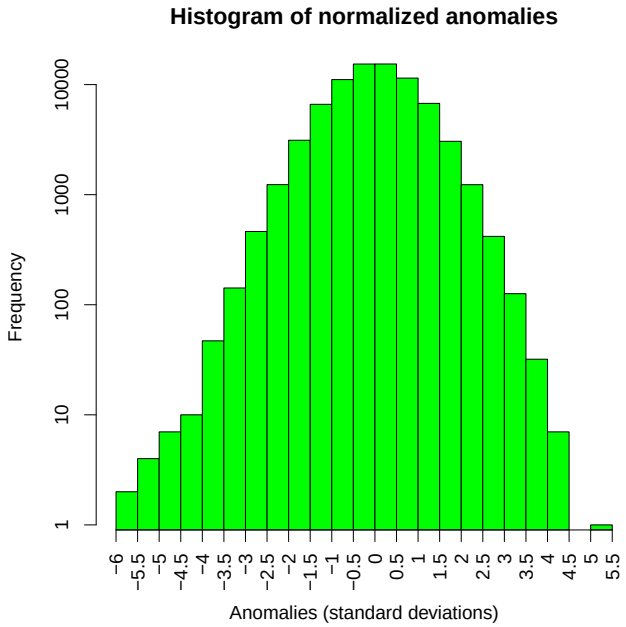

Figure 7: Histogram of anomalies (all data together).

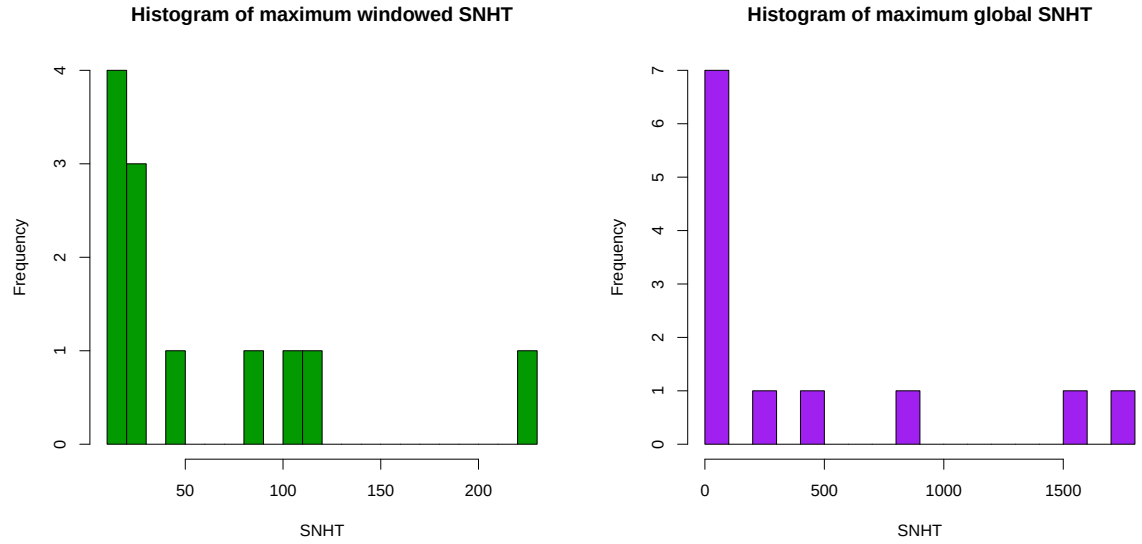

Figure 8: Histograms of the maximum SNHT values found on overlapping stepped windows (left) and on the whole series (right).

The last page of the document shows a plot of station numbers (their order in the `*.est` stations file) according to their final Root Mean Squared Error (RMSE) and SNHT values. RMSE are calculated by comparing the estimated and the observed data in every series. A high value may indicate a bad quality of the series, but it could be caused by the station being located in a peculiar site with a distinct micro-climate as well. Anyway, the homogeneous series from stations sharing the common climate of the region will tend to be clustered to the left-bottom part of the plot.

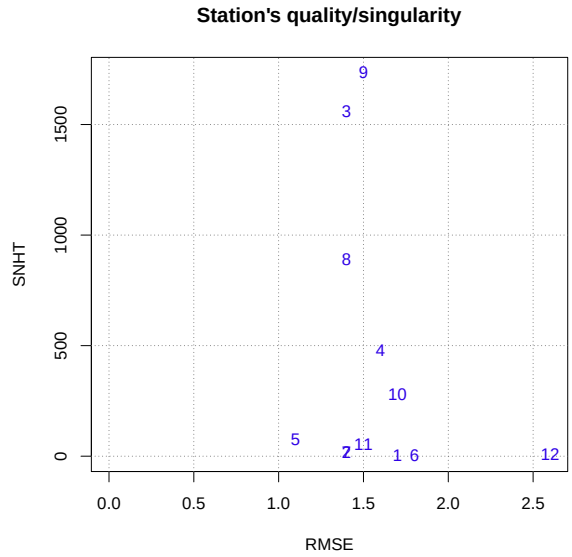

Figure 9: Quality/singularity plot of the final series.

After all these considerations, we would proceed to homogenize the series by applying:

```
homogen('Ttest', 1981, 2000, dz.max=9, snht1=60, snht2=70)
```

But since our example is based on daily data and these kind of series exhibit a high variability that lowers the efficiency of the detection of their inhomogeneities, it is better to aggregate them and homogenize the monthly series first. *Climatol* helps in obtaining monthly data form the daily series by means of the `dd2m` function, that we can apply here in this way:

```
#(With precipitation, add the parameter valm=1 to
# calculate monthly totals instead of averages)
dd2m('Ttest', 1981, 2000)
```

This command saves in `Ttest-m_1981-2000.dat` and `Ttest-m_1981-2000.est` the monthly series, ready to be homogenized. (The suffix `-m` has been added to the name of the variable to avoid overwriting the original daily series.)

3.3. Homogenization of the monthly series

If the user is working with monthly data, there would be no need to use the suffix `-m`, but here we a going to homogenize the monthly series that were obtained form the daily values in the previous sub-section. We could begin with an exploratory application of `homogen` as we did with the daily data, but let us try here the function with its default values:

```
homogen('Ttest-m', 1981, 2000)
```

The inspection of the `Ttest-m_1981-2000.pdf` output graphics reveals that the default values of `snht1=snht2=25` seem appropriate for the monthly values. Most of the graphics have already been formerly discussed. The only difference is that now we have anomaly plots for the detection stages 1 and 2, where the detected shifts in the mean are marked in red (see an example in Figure 10 left). The SNHT histograms at the end of these stages refer to the values of the test after the series have been split at the detected break-points.

After the graphs of anomalies in stage 3, the "final graphics" illustrate the reconstruction of complete series from every homogeneous sub-period. Figure 10 (right) shows an example of a series that was split into two sub-periods. The upper part of the graphic plots the running annual means of the reconstructed series, with original data in black and infilled data in different colors for every resulting series. (Note that in the presence of missing data those running means that cannot be calculated will be missing in the graph.) The lower part display the corrections applied to the series, plotted in different colors. As you can see, the corrections have seasonal variations (constant corrections can be achieved in this case if standard deviations are not used in the normalization by setting `std=1`), and the spikes are due to outlier rejections.

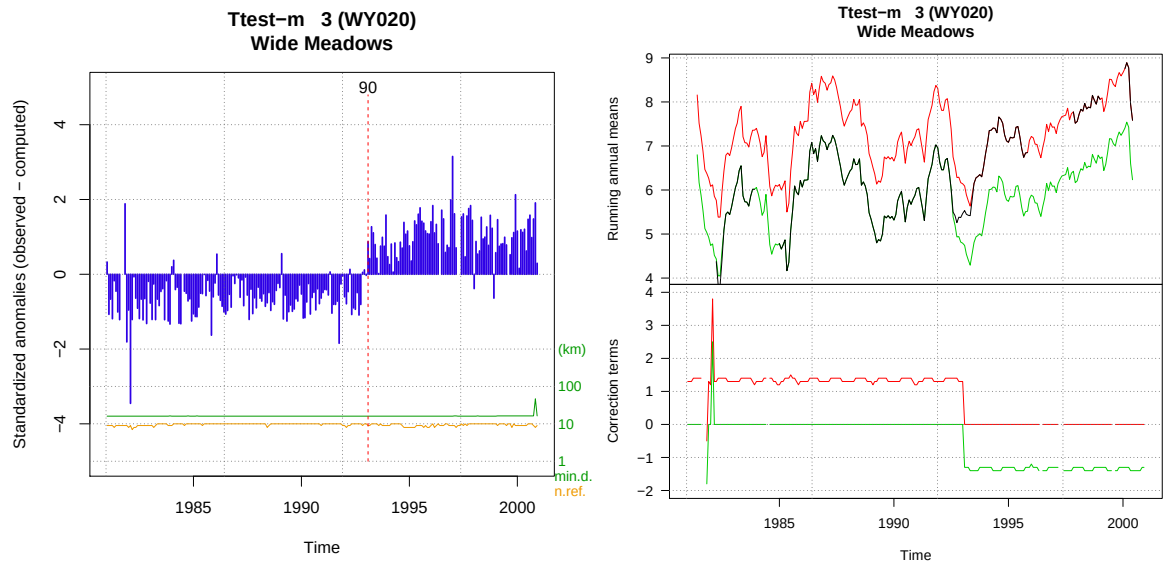

Figure 10: Example of detection of a shift in the mean of a series with SNHT=90 (left) and the reconstruction of complete series from both homogeneous sub-periods (right).

After successfully running the `homogen` function, the user can find the following files in his R working directory (without the suffix `-m` if the original series had monthly or higher periodicity):

- `Ttest-m_1981-2000.txt` : Text file logging all console messages issued during the process. It includes the station clustering and final summaries of SNHT and RMSE values of the resulting series.
- `Ttest-m_1981-2000_out.csv` : Text file (comma separated values) with the list of corrected outliers. Note that the "Suggested" values are only first guesses at the moment of outlier rejection. Therefore, final values can differ and even have multiple values (when the series are split).

- `Ttest-m_1981-2000_brk.csv` : Text file (comma separated values) with the list of break-points and their associated SNTH values.
- `Ttest-m_1981-2000.pdf` : The diagnostic graphics discussed previously.
- `Ttest-m_1981-2000.rda` : An R binary file holding the homogenization results. (See the `homogen` R documentation for more details.)

When the user has direct access to the original data, it is worth inspecting the list of rejected outliers in the `Ttest-m_1981-2000_out.csv` file and check whether they are errors or reliable values. After having corrected the database, input files for *Climatol* could be recompiled and the whole procedure repeated.

Moreover, if there are metadata about historical changes in the observatories, it is very convenient to edit the file `Ttest-m_1981-2000_brk.csv` to adjust the dates of detected break-points to those of the events that can have altered the observations and run the `homogen` function again with the parameter `metad=TRUE` (and `sufbrk=''` if original series were composed of monthly data). But note that not all changes must necessarily have an impact on the climatic variable under study, and that the most common situation is to have incomplete metadata, if not totally absent.

3.4. Adjustment of the daily series with the monthly break-points

If we were studying daily data, now we will adjust them using the break-points detected in their monthly aggregates, with a new application of the function `homogen` with the parameter `metad=TRUE`:

```
homogen('Ttest', 1981, 2000, dz.max=7, metad=TRUE)
```

In this way, `homogen` skips the two detections stages and proceeds to split the daily series by the break-points listed in the `Ttest-m_1981-2000_brk.csv` file, and then resumes processing the third stage of reconstruction of all series from their homogeneous sub-periods by means of its infilling routine. This process creates the usual output files except the `*_brk.csv`, since no break-point detection has been done this time.

4. Obtaining products from homogenized data

The user can load the results of the homogenization into the R memory space for any further manual processing by issuing the command:

```
load('Ttest_1981-2000.rda')
```

But *Climatol* provides the post-processing functions `dahstat` and `dahgrid` to help in obtaining common products from the homogenized series, either directly from the daily series, or from their monthly aggregates, which can be generated by:

```
#(With precipitation, add the parameter valm=1 to
# calculate monthly totals instead of averages)
dd2m('Ttest', 1981, 2000, homog=TRUE)
```

With the parameter `homog=TRUE` the monthly aggregates will we calculated from the homogenized series, and not with the original series as was done previously. Now the newly created files are `Ttest-mh_1981-2000.dat` and `Ttest-mh_1981-2000.est`, containing monthly aggregates of the adjusted daily series.

4.1. Homogenized series and statistical summaries

The homogenized series can be obtained in two text CSV files in this way:

```
dahstat('Ttest', 1981, 2000, stat='series')
```

One of the generated files, `Ttest_1981-2000_series.csv`, contains all the homogenized series, and the other, `Ttest_1981-2000_flags.csv`, supplies flags indicating whether the data are observed (0), infilled (1, originally missing) or corrected (2, either because of break-points or outliers).

Statistical summaries are created with the same function. Here are a few examples (more information in the R documentation of `dahstat`):

```
dahstat('Ttest',1981,2000) #means of the daily series
dahstat('Ttest',1981,2000,mh=TRUE) #means of their monthly aggregates
dahstat('Ttest',1981,2000,mh=TRUE,stat='tnd') #monthly trends/p-values
dahstat('Ttest',1981,2000,stat='q',prob=.2) #first quintile (dailies)
```

This function includes parameters to choose a subset of the series, either by providing the code list of the desired stations (as with `cod=c('WY020','WY055')` or specifying that we want the series reconstructed from the last homogeneous sub-period (`last=TRUE`), from the longest sub-period (`long=TRUE`), etc.

4.2. Homogenized gridded series

The other post-processing function, `dahgrid`, provides grids calculated from the homogenized series (disregarding infilled data). But before applying this function, the user must define the limits and resolution of the grid, as in this example:

```
grd=expand.grid(x=seq(-109,-107.7,.02), y=seq(44,45,.02)) #desired grid
library(sp) #load needed package for the following command:
coordinates(grd) <- ~ x+y #convert the grid into a spatial object
```

The R function `expand.grid` has been used to define the sequence of X and Y coordinates, and then `coordinates` (from the `sp` package) is applied to convert the grid, stored as `grd` (any other name could have been used), into an object of class spatial.

Now grids can be generated (in NetCDF format) as in:

```
dahgrid('Ttest', 1981, 2000, grid=grd) #grids with daily time steps
dahgrid('Ttest', 1981, 2000, grid=grd, mh=TRUE) #id. with monthly steps
```

These grids are built in normalized, dimensionless values. You can obtain new grids with temperatures in degrees Celsius by means of externals tools, such as the Climate Data Operators (CDO):

```
#These are not R commands! Example for a linux/unix terminal:
cdo add -mul Ttest-mh_1981-2000.nc Ttest-mh_1981-2000_s.nc \
  Ttest-mh_1981-2000_m.nc Ttest-mu_1981-2000.nc
```

But the new grids in `Ttest-mu_1981-2000.nc` will be based on geometric interpolations only, and therefore better grids of means and standard deviations should be obtained with geostatistical methods in `Ttest-mh_1981-2000_m.nc` and `Ttest-mh_1981-2000_s.nc` before using them to undo the normalization.

5. Additional recipes

The previous examples showed and discussed the more frequent applications of the homogenization functions of *Climatol*. However, questions may arise relative to how to proceed when dealing with other climatic variables or time resolutions. This section is dedicated to answer those questions, and more answers may be added in the future as users ask for further assistance.

5.1. How to modify weights and number of references

The weights w_j given to nearby data to estimate the values of the series depend on the distances d_j through the function $w_j = 1/(1+d_j^2/h^2)$, where h is the distance at which the weight is halved. By default, $h = 100$ km, but it can be changed by assigning another value to the parameter `wd` (weight distance), which is how h in the formula is called within the `homogen` function. By default, `wd=0` in the two first detection stages, meaning that no weighting is applied to the data, since we want to avoid giving too much weight to a very close but potentially inhomogeneous station. But the user is allowed to specify `wd` for the three stages, as setting `wd=c(0,1000,25)`, which gives no weights for the first stage, $h = 1000$ for the second and $h = 25$ for the third. Figure 11 shows the variations of the weights depending on the distance for several values of h (`=wd`).

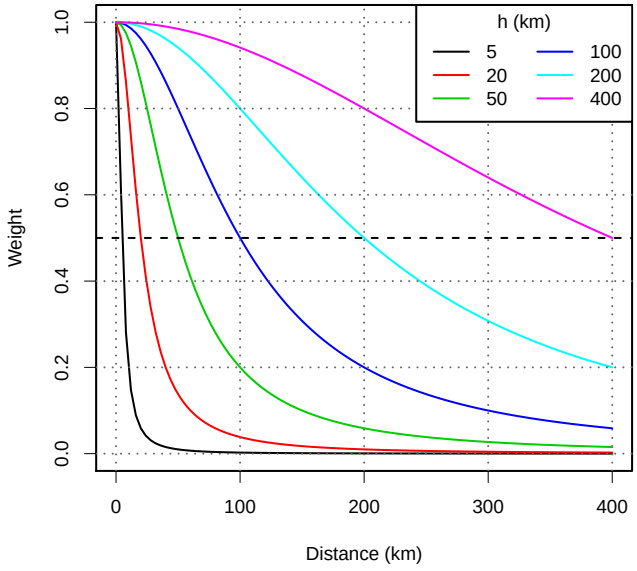

Figure 11: Weight variation for several values of h (`wd` parameter).

As to the number of closest data used at every time step, up to 10 are used by default in the detection stages (if they are available), and 4 in the final series reconstruction stage. This can also be changed with the parameter `nref`, as in `nref=c(8,8,2)`.

The chosen parameters can be optimal or not depending on the final purpose of the series analysis. E.g.: If you want to obtain climate normals, the variance adjustments will have no importance, while they can be crucial if deriving extreme value return periods from the series. In the

latter case, you can limit the variance diminution of the weighted estimates by setting shorter weighting distances, especially in the third stage (e.g.: `wd=c(100,100,15)`), and/or reducing the number of references, even using only one, in the last stage (`nref=c(5, 5, 1)`), as can be preferred when adjusting daily precipitation series.

5.2. How to save results from different runs

If you run `homogen` with different settings to explore which give better results, you can avoid overwriting your previous outputs by renaming them with the help of the `outrename` function. For example, the command

```
outrename('Ttest-m', 1981, 2000, 'old')
```

will rename all output files `Ttest-m_1981-2000*` to `Ttest-m-old_1981-2000*`

5.3. How to change the cutting level in the cluster analysis

Climatol applies a cluster analysis at its initial checks of the data, but the number of clusters is automatic. Looking at the dendrogram near the beginning of the PDF output document, a better cutting level could be chosen. In the example of Figure 12, three groups of stations have been produced at a dissimilarity level of 0.058. But we could prefer to cut the dendrogram at 0.04 to obtain five groups. In that case, we only need to repeat the homogenization command with the addition of the parameter `cutlev=0.04`. By default, up to 100 series are used in the cluster analysis, taking a random sample of this size when the number of studied series exceeds this limit. But if that number is not huge, the user can force to use them all by setting, e.g., `nclust=136`. Also note that the number of clusters is limited to nine.

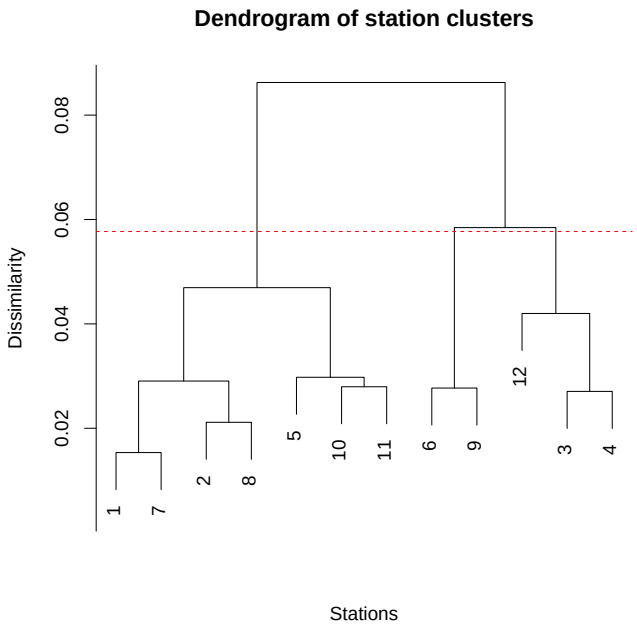

Figure 12: Dendrogram of the stations, based on their correlation coefficients.

5.4. My station coordinates are in UTM

Climatol assumes that coordinates are in degrees if the absolute values of X and Y are lower or equal than 180 and 90 respectively. Otherwise, if the mean of X or Y are greater than 10000 they will be assumed to be in meters, and will be converted to kilometers for the rest of the process.

5.5. How to apply a transformation to my skewed data

The `homogen` function can apply a $log(x+1)$ transformation (`trf=1`) or any root transformation (`trf=2` for square root, `trf=3` for cubic root, etc. Fractional numbers are allowed). By removing their skewness, data could standardized (`std=3`, the default), but benchmarking results with monthly precipitation series during the MULTITEST project showed clearly better results when data were normalized by their average ratio (`std=2`) without the need to apply any transformation to them.

5.6. How to limit the possible values of a variable

Parameters `vmin` and `vmax` can be used to force `homogen` to limit the range of possible values. This can be useful when dealing with relative humidity (set `vmin=0` and `vmax=100`) or any other variable with a truncated range of possible values. Note that `vmin=0` is automatically set when `std=2`, because the average ratio normalization will be normally applied to variables like precipitation or wind speed, which cannot have negative values.

5.7. Can I use reanalysis outputs as reference series?

When data are very fragmented and some time steps of our period of study are void of data in all series, a possible solution is to use series derived form reanalysis products to act as reference series providing data for those critical gaps. These series should be positively correlated with our variable. E.g., temperature at 2 m above ground for temperature series or, if not available, geopotential thickness near the surface or similar for our temperature series. Precipitation series could lack its equivalent in reanalysis, and then some derived variable could be tried (vorticity advection?, vertical velocity?, a combination of them?), but its correlation should be tested before using it as reference series.

Although the appearance of new systems of observation (e.g., satellites) introduce inhomogeneities in the amount of available data assimilated in the models, reanalysis products may be considered in general more homogeneous than the observational series. To use these products as references, series from one or more grid points located in the study domain can be added to the data file `*.dat`, and the coordinates of their corresponding grid points with bogus codes and names appended to the stations file `*.est`. Their codes should begin with an asterisk (example: `*R43`) to skip quality and homogeneity controls of these reliable series.

5.8. Which split series should be retained?

Most homogenization methods return the series adjusted backwards from the last homogeneous sub-period, but *Climatol* yields complete reconstructions from every sub-period (unless it is too short to be reconstructed reliably). Therefore, the user may wonder which one to use in his climate study. The answer depend on the objective of the investigation. To obtain normal values to calculate anomalies of newly incoming data for climate monitoring, those normals should be adjusted to the last homogeneous sub-period. But if the goal is to produce a map of the normal values, all series should be used, since some of them can adjust better to the spatial variability at the scale of the map, while other may be affected by local microclimates and we would obtain a noisier map.

5.9. I have so many long daily series that the process is taking days!

One possibility to shorten the computation time is to apply `homogen` to sub-areas, trying to group those stations sharing the same climate factors. If there are no clear climate discontinuities (e.g., mountain ridges crossing the domain under study), `homogsplit` can produce overlapping sub-areas and homogenize them automatically. The user only needs to supply the dividing X and Y coordinates, but special care should be taken to avoid too few stations in any of the sub-areas (although they can be empty). There is also an increased possibility of becoming void of data at some time steps in some sub-areas, halting the process. (This function can be considered experimental.)

6. Bibliography

Aguilar E, Auer I, Brunet M, Peterson TC, Wieringa J (2003): *Guidelines on climate metadata and homogenization*. WCDMP-No. 53, WMO-TD No. 1186. World Meteorological Organization, Geneve.

Alexandersson H (1986): A homogeneity test applied to precipitation data. *Jour. of Climatol.*, 6:661-675.

Khaliq MN, Ouarda TBMJ (2007): On the critical values of the standard normal homogeneity test (SNHT). *Int. J. Climatol.*, 27:681687.

Paulhus JLH, Kohler MA (1952): Interpolation of missing precipitation records. *Month. Weath. Rev.*, 80:129-133.

Peterson TC, Easterling DR, Karl TR, Groisman P, Nicholls N, Plummer N, Torok S, Auer I, Böhm R, Gullett D, Vincent L, Heino R, Tuomenvirta H, Mestre O, Szentimrey T, Salinger J, Førland E, Hanssen-Bauer I, Alexandersson H, Jones P, Parker D (1998): Homogeneity Adjustments of 'In Situ' Atmospheric Climate Data: A Review. *Int. J. Climatol.*, 18:1493-1518.

Sokal RR, Rohlf PJ (1969): *Introduction to Biostatistics*. 2^{nd} edition, 363 pp, W.H. Freeman, New York.

Venema V, Mestre O, Aguilar E, Auer I, Guijarro JA, Domonkos P, Vertacnik G, Szentimrey T, Stepanek P, Zahradnicek P, Viarre J, Müller-Westermeier G, Lakatos M, Williams CN, Menne M, Lindau R, Rasol D, Rustemeier E, Kolokythas K, Marinova T, Andresen L, Acquaotta F, Fratianni S, Cheval S, Klancar M, Brunetti M, Gruber C, Prohom Duran M, Likso T, Esteban P and Brandsma T (2012): Benchmarking homogenization algorithms for monthly data. *Clim. Past*, 8:89-115.