User's guide of the *climatol* R Package (version 4.1)

Jose A. Guijarro (jaguijarro21@gmail.com)

2024-04-05*

Contents

Foreword

This guide is a complement to the R standard manual included in the *climatol* package itself, which is where all aspects of each function are detailed (description of the parameters, supplementary information and examples). Here it will be explained how use those functions for quality control, homogenization and filling of missing data of a set of climate series, and how to obtain derived products from them. Examples of functions that generate various graphs useful in climatology will also be shown, but the information on its use must be found in the standard manual.

The main new features added to the *homogen* function in versions 4.* are the following:

- Algorithms have been introduced to automatically apply the more convenient parameters, depending on the studied variable and its temporal resolution, thus simplifying its use.
- Initial quality checks have been added, automatically discarding any extremely anomalous data that could compromise the subsequent process when using them as references for their neighboring series. Too long runs of identical data are also deleted.
- In the case of daily rainfall there is now the possibility of dissaggregating accumulated data when the observer could not read the rain gauge for a few days.
- Data transformation is not an option anymore, since tests carried out within the MULTITEST project showed that the normal ratio normalization of highly skewed data yielded much better results than their logarithmic or root transformations.
- If reanalysis series are added to serve as references, by default they will weight less than the observed series.
- The list of outliers now includes data that, not being deleted, can be considered suspect values.
- The text file that collects the messages issued during the entire process contains at the end the tails of the distributions of the final anomalies and homogeneity tests as a complement to the histograms displayed at the end of the graphics file.
- By default, the homogeneity test is still the SNHT, but the Cucconi test can be chosen as an alternative for simultaneously detecting changes in the mean and in the variance of the series of anomalies.

In addition to these changes to the *homogen* function, this version adds functions for:

- Restore in the homogenized data of the results file `*.rda` the rejected anomalous data that the user considers were correct, produced by local phenomena that should not have affected the neighboring series.
- Convert existing series in CSV, xls/xlsx and SEF[1] files to *climatol* format.
- Prune any excess daily sunshine hours resulting from the homogenization of those serie.
- Obtain, for each daily or sub-daily series, monthly quantiles of extreme values, of increments between consecutive values, and of lengths of sequences of identical values. These quantiles can be used to implement suspicious value alerts in Climate Data Management Systems.
- Generate various graphics. To the already existing wind roses and Walter and Lieth climograms are added graphs of precipitation Intensity-Duration-Frequency, running trends in periods of different lengths, isopleths on a month-hour diagram, and meteograms.

The latest minor updates (4.1.*) of this package can be found at https://climatol.eu, along with this user's guide and some help videos.

This guide is structured in two chapters. The first and foremost is dedicated to the homogenization of series and interpretation of the results, while the second shows examples of the graphic functions previously discussed.

[1] Station Exchange Format, the Copernicus Climate Change Service file format for Data Rescue projects.

1 Homogenization of series

1.1 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. The latter can produce sudden changes, like a fire burning an adjoining forest, or gradual changes, 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), Aguilar *et al.* (2003) and Venema *et al.* (2012), as well as the guidelines on homogenization (WMO, 2020).

There are several software packages that implement these methods so that they can be used by the climatological community. 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/) (Guijarro *et al.*, 2023) 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 the European project INDECIS homogenized the series of eight essential climatic variables from the ECA&D database.

The comparisons carried out within the framework of the aforementioned MULTITEST project showed that *climatol* yielded results ranking within the best that can be obtained by other methods. Furthermore, while other programs are poorly tolerant to missing data, *climatol* was designed to be able to use very short or fragmented series, thus taking advantage of all the climatic information available in the study area.

1.2 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 databases. 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 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 perpendicularly to it (model II regression) instead of in the vertical direction (model 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.

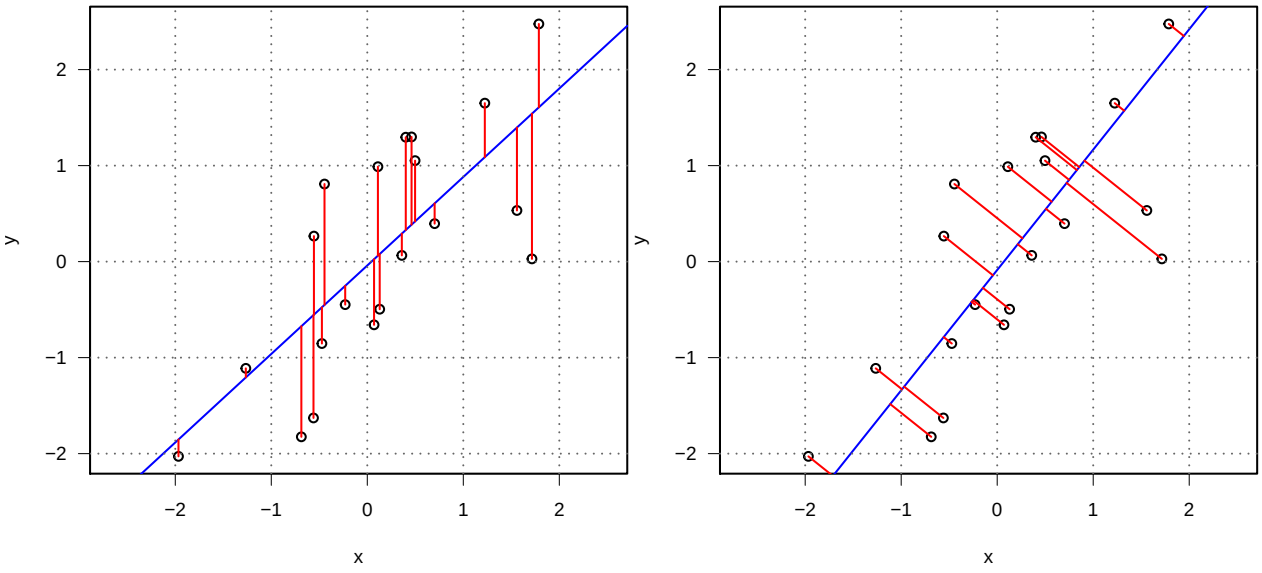


Figure 1: In red, deviations from the regression line (blue) minimized by least squares in regression models I (left) and II (right).

The series estimated from the others serve as references for their corresponding observed series, so the next step is to obtain series of anomalies (spatial) 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 exceeding a preset threshold `dz.max`.
- Check their homogeneity by applying the *Standard Normal Homogeneity Test* (SNHT: Alexandersson, 1986). Alternatively, the Cucconi (1968) test can be chosen).

When the maximum values obtained when applying the test to the series are greater than a preset threshold `inht` *(INHomogeneity Threshold)*, the series is divided by the point of maximum value of the test, passing all previous data to a new series that is added to the others with the same coordinates but adding a numerical suffix to the code and the name of the station. This procedure is performed iteratively, cutting only the series with higher SNHT values in each cycle, until no more inhomogeneities are found. Furthermore, since SNHT is a test originally devised for finding a single breakpoint in a series, the existence of two or more jumps in the mean of a similar size could mask your results. To minimize this problem, SNHT is applied over overlapping time windows in a first pass, and then in a second stage SNHT is applied to the complete

series, which is when the test has more detection power. Finally, a third stage is dedicated to filling in all the missing data. in all series and homogeneous subseries with the same procedure of data estimation explained above. Therefore, although the The underlying methodology of the program is very simple, its operation is complicated by a series of nested iterative processes, as shown in the flowchart of figure 2.

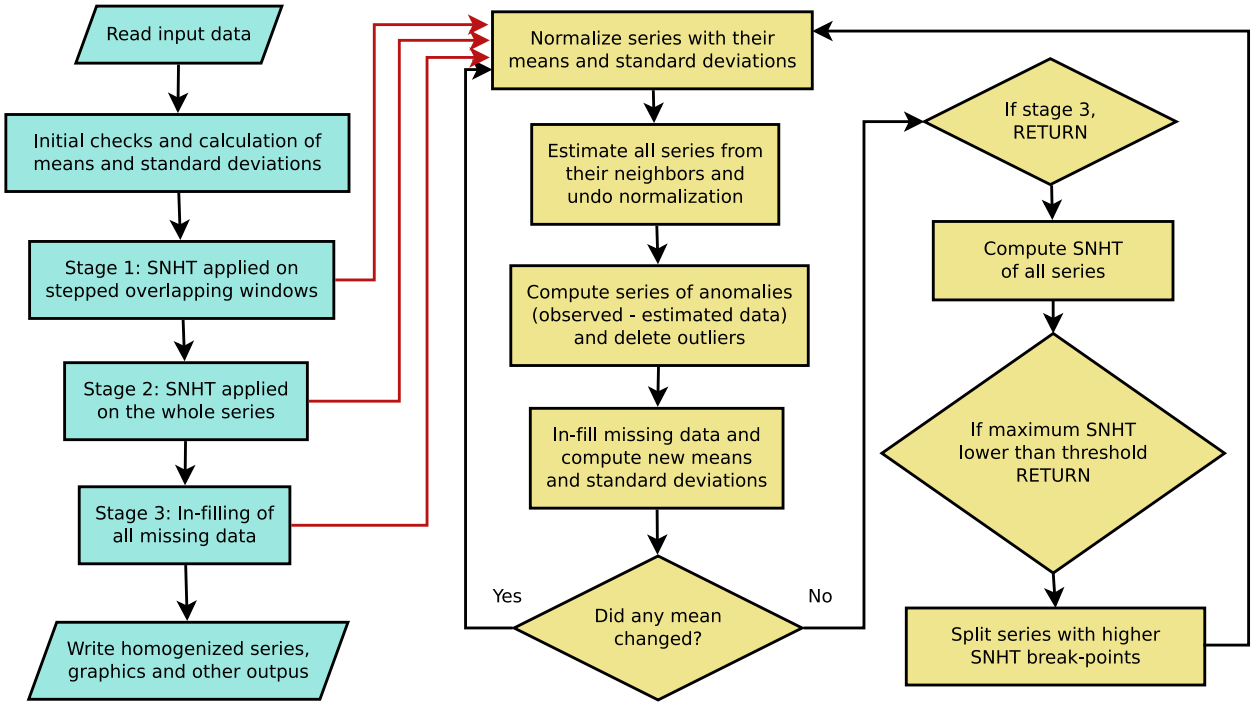


Figure 2: Flowchart of *climatol* operation, showing its iterative processes.

Although SNHT thresholds have been published for different series lengths and statistical significance levels, experience shows that this test can yield very different values depending on the climatic variable studied, the degree of correlation between the series and their temporal frequency. *Climatol* defaults to the threshold value `inht=25`, appropriate for monthly values of temperature, although a little conservative, trying not to detect false jumps in the average at the cost of letting the minor ones pass. for other variables it may be necessary to adjust that threshold, with the help of the final test histograms and the anomaly graphics included in the PDF output file. If one wanted to homogenize daily series directly, the threshold should be around ten times larger, but it is advisable to perform the detection of changes in the mean on the monthly series, and then use the breakpoints (optionally adjusted to the available metadata) to adjust the daily series.

1.3 Homogenization procedure

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

1.3.1 Preparation of the input files

Climatol only needs two input files, one with the list of coordinates, codes and names of the stations, and another with all the data, station by station, in the same order in which they appear in the stations file. None of these files have header lines or row numbers, and their data items are separated by spaces. (Station names must be quoted if they consist in more than one word.) All the series must be complete in the data file, using `NA` or other code for missing data, and lines may have any number of data, since the file will be read sequentially. Furthermore, to avoid problems with the post-processing functions that will be discussed later, the period of study should cover full years, beginning in January (day 1 if they are daily data) of the

initial year and ending in December (the 31st in the case of daily data) of the final year, although this it is not strictly necessary.

Both files share the same basic name `VRB_yyyy-YYYY` (where `VRB` is an abbreviation of the studied variable, and `yyyy` and `YYYY` the first and last years of the data), but have different extensions: `dat` for the data and `est` for the stations. These extensions are not recognized directly by Windows, so users of this operating system will need to indicate that they should be opened with notepad or another editor of plain text, avoiding the use of word processors such as LibreOffice or MS-Word.

This can be illustrated with the examples of the standard documentation of *climatol*. (As a requirement of the CRAN repository is that the examples must run in a few seconds, data in the examples are much smaller than those normally used in real applications):

```
library(climatol) #load the functions of the package
data(climatol_data) #load example data into memory
#show a fragment of the data file:
write(Temp.dat[41:80,4],stdout(),ncolumns=10)
```

```
20.4 23 25.8 26.1 24.8 18.9 14.9 11.7 10.6 8.5
12.6 14.2 19.3 22.4 26.4 26.1 NA NA 15.9 12.5
13.2 NA 12.6 17 NA NA NA 25.5 23.1 18.2
12.5 10.1 10.7 11.2 13.5 NA 18.1 19.5 NA 24.9
```

```
#show the stations file:
write.table(Temp.est,stdout(),row.names=FALSE,col.names=FALSE)
```

```
-2.5059 39.0583 210 "st01" "Station 1"
-2.7028 38.9808 112 "st02" "Station 2"
-2.63 38.8773 111 "st03" "Station 3"
-2.5699 38.9205 112 "st04" "Station 4"
-2.4663 38.9885 125 "st05" "Station 5"
```

It can be seen that each line contains the coordinates X (longitude, °), Y (latitude, °), Z (elevation, m), station code and name, all separated by spaces. Coordinates are expressed in degrees with decimals (not in degrees, minutes, and seconds) and with the appropriate sign to indicate West, East, North or South.

Let us save these input files of monthly temperatures and daily rainfall to perform a couple of examples of homogenization (a working directory must be chosen first in the R session):

```
#monthly temperatures of 5 stations in 1961-2005:
write.table(Temp.est,'Temp_1961-2005.est',row.names=FALSE,col.names=FALSE)
write(Temp.dat,'Temp_1961-2005.dat')
#daily precipitations of 3 stations in 1981-1995:
write.table(SIstations,'Prec_1981-1995.est',row.names=FALSE,col.names=FALSE)
dat <- as.matrix(RR3st[,2:4])
#the series of precipitación are complete, but we can delete some data:
dat[1:300,1] <- dat[c(1000:1200,2000:2015),2] <- dat[5000:5478,3] <- NA
#we will also introduce a few errors:
dat[500:509,1] <- 9.9; dat[600,2] <- -9.9; dat[3000,3] <- 999
#now save the data file:
write(dat,'Prec_1981-1995.dat')
```

To help in preparing your input files in this format, *climatol* provides some useful functions (see the *climatol* standard documentation for more details about its use):

- **db2dat** creates input files directly from a database (accessible through the ODBC protocol).
- **daily2climatol** compiles daily data from individual station files.

- **rclimdex2climatol** converts the data from RClimDex files.
- **sef2climatol** gather the data from SEF[2] files.
- **xls2csv** dumps data from individual *.xls or *.xlsx files into a single CSV file.
- **csv2climatol** reads the data from a single CSV file and generates the *climatol* files.

1.3.2 Homogenization of the daily precipitation

Once the files are generated in the required format, we can address the homogenization of their series. The function that performs this task is called **homogen**, and initially requires only three parameters to be specified: the short name of the variable and the initial and final years of the period covered by the data. Example with the previously saved precipitation files:

```
homogen('Prec', 1981, 1995)
```

The **homogen** function calculates the frequency of the data (subdaily, daily, monthly, bimonthly, quarterly, semiannually or annually[3]) from the amount of data present in the period between the initial and final years. But as the high variability of the daily (or subdaily) data makes it very difficult to detect changes in the mean, the direct homogenization of these data is not recommended, and that is why in this example an error is generated advising to first homogenize your monthly data, whose files are easily obtained using the **dd2m** function.

However, before generating the monthly series it is advisable to check the quality of the daily data, for which we can use the option **onlyQC=TRUE**:

```
homogen('Prec', 1981, 1995, onlyQC=TRUE)
```

The messages output to the console (which have been saved to the file **Prec-QC_1981-1995.txt**) report about the deletion of one negative data and another very abnormal. **homogen** also generates a large collection of diagnostic graphics in the file **Prec-QC_1981-1995.pdf**, whose first three graphics show (figure 3) in this case, for each series:

1. Box plots of the data, indicating the deleted anomalous data with a Red dot.
2. Boxplots of the differences between consecutive data.
3. Lengths of the sequencies of identical data.

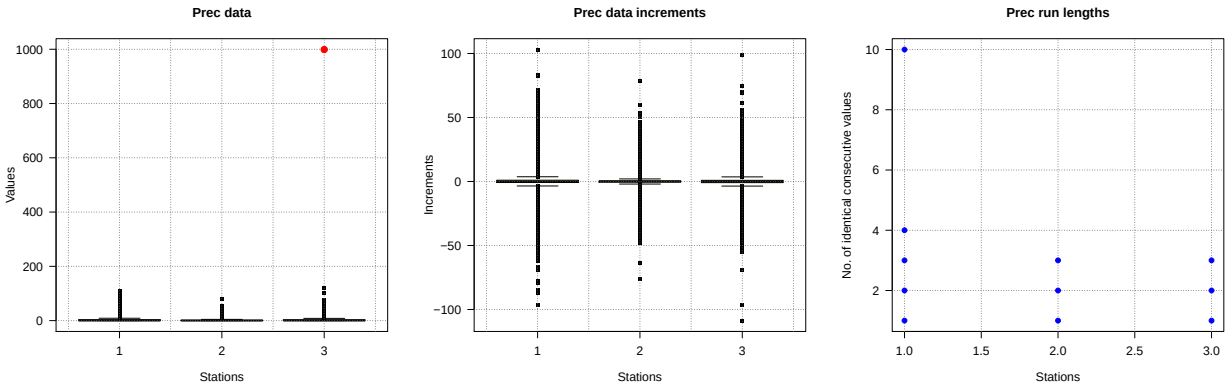


Figure 3: Boxplots and lengths of constant sequencies for the initial quality control.

With daily rainfall, the program automatically excludes zeros, but we see a sequence of 10 days with the same data in series 1 that has not been eliminated because, since there are only three series, it has not been

[2] SEF (*Station Exchange Format*) is the Copernicus Climate Change Service format for Data Rescue projects.
[3] Generally we will treat daily or monthly data.

considered statistically anomalous. But we can force its deletion by lowering the threshold for the number of interquartile distances, which is 1 by default for sequences of identical data, to 0.5 (the first value of `niqd` is the threshold for anomalous data, which by default is 4):

```
> #restore the original data file:
> file.copy('Prec-QC_1981-1995.dat','Prec_1981-1995.dat',overwrite=TRUE)
> #repeat quality control with lower tolerance to sequences of identical data:
> homogen('Prec', 1981, 1995, onlyQC=TRUE, niqd=c(4,0.5))
```

As in the previous case, having found clear errors in the input data, the original files and the quality control results are saved by adding the suffix `-QC` to the name of the variable, and the series free of said errors are saved with the original names of the input files. Now we can proceed to obtain the monthly data files and homogenize their series:

```
# valm=1 aggregates daily data into monthly totals instead of averages:
dd2m('Prec', 1981, 1995, valm=1) #obtain the monthly series 'Prec-m'
# we set 'std=2' (the recommended normalization for precipitation data)
# because it will not be set automatically for monthly series.
# annual='total' makes the last graphics to show running annual totals instead
# of averages:
homogen('Prec-m', 1981, 1995, std=2, annual='total')
```

Both the graphics file `Prec-m_1981-1995.pdf` and the one containing the list of breakpoints `Prec-m_1981-1995_brk.csv` (empty in this case) indicate that no series has been cut, so that all of them can be considered homogeneous. Had jumps in the average been detected, it would be convenient to edit the dates of the breakpoints to adjust them to events in the history of the stations (metadata) justifying the detected changes. Finally, we would adjust the daily series by:

```
homogen('Prec', 1981, 1995, annual='total', metad=TRUE)
```

Again it is advisable to review the results to check if there have been any problems. Special care must be taken with the list of anomalous data that appears in the file `Prec_1981-1995_out.csv`. We see that there are a lot of undeleted suspicious data (marked with a 0 in the 'Deleted' column), and the console messages (saved in `Prec_1981-1995.txt`) indicate that some anomalous data would have been deleted if they had more than one reference available. The histogram of standardized anomalies from the graphics file `Prec_1981-1995.pdf` (figure 4) shows that two of the three eliminated data are quite anomalous with respect to the overall frequency distribution. However, the list in the file `Prec_1981-1995_out.csv` tells that both data correspond to the same day (1993-10-06) in two different series. In reality, the anomalous data occurs at station `p084`, which recorded 121.6 mm, thus causing the data from station `p064` to show a negative anomaly due to having much lower precipitation.

In these cases, the ideal is to check if the suspicious data are correct, eliminating them otherwise, and to repeat the adjustment of the series. However, it must be taken into account that even if a very anomalous data is correct due to a local weather phenomenon, it is better to eliminate it before the homogenization and restore it later in the file of homogenized series, since otherwise that local anomaly will cause unwanted changes in neighboring series. (The `datrestore` function automates the restitution of the outliers for which the user has manually changed the sign to negative in the `Deleted` column of the `*_out.csv` file.)

1.3.3 Homogenization of the monthly temperature

Let's now see another example of homogenization with the monthly temperatures that we had recorded:

```
homogen('Temp', 1961, 2005)
```

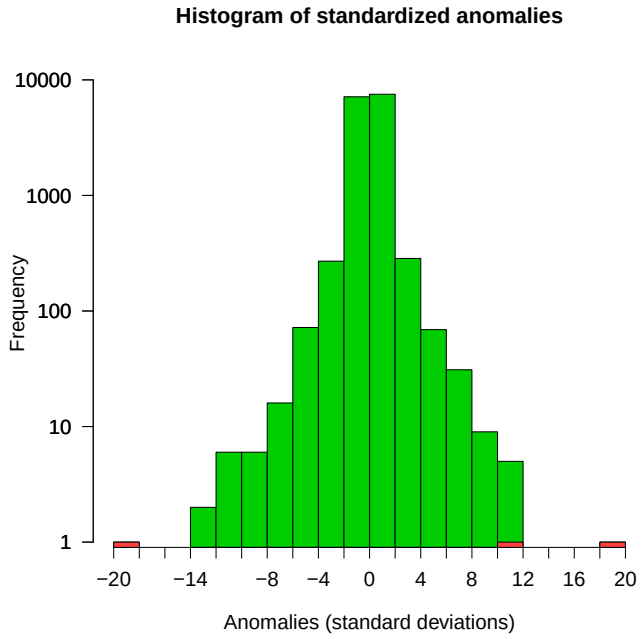


Figure 4: Histogram of spatial anomalies showing in red the rejected data.

This process ends with an error message because the detection and deletion of anomalous data has produced a complete absence of data in some time steps, which prevents the estimation of values for the missing data due to the lack of reference data in those empty terms. The diagnostic graph file `Temp_1961-2005.pdf` shows a very anomalous negative value in series 1 (page 2) and a run of 9 identical monthly values in series 4 (page 4). The messages displayed in the console (and recorded in `Temp_1961-2005.txt`) warn that the anomalous data has been deleted but that excessively long runs of identical data have not been deleted because the interquartile range, which is used to discriminate the excessive runs, is zero. Therefore we must investigate this run of identical data on our own, for which we can read the data and represent series 4 with:

```
d <- read.dat('Temp', 1961, 2005) #list including the 'dat' matrix
plot(d$dat[,4], type='l')
```

The horizontal line observed around term 325 must correspond to the anomalous run of identical data. Let's look at those data:

```
print(d$dat[320:340,4])
```

```
[1] NA NA NA NA NA NA 11 11 11 11 11 11 11 11 11 NA NA NA NA NA NA
```

Indeed, we see that within a period of missing data there is a temperature of 11°C that is repeated 9 months in a row. That has to be an error, so we are going to delete that run, rewrite the data file and repeat its homogenization:

```
d$dat[320:340,4] <- NA
write(d$dat, 'Temp_1961-2005.dat')
homogen('Temp', 1961, 2005)
```

After the first three graphs of the initial quality control of the file `Temp_1961-2005.pdf`, which only show the very anomalous data (automatically deleted), we can see the availability of data, both station by station and in total (figure 5).

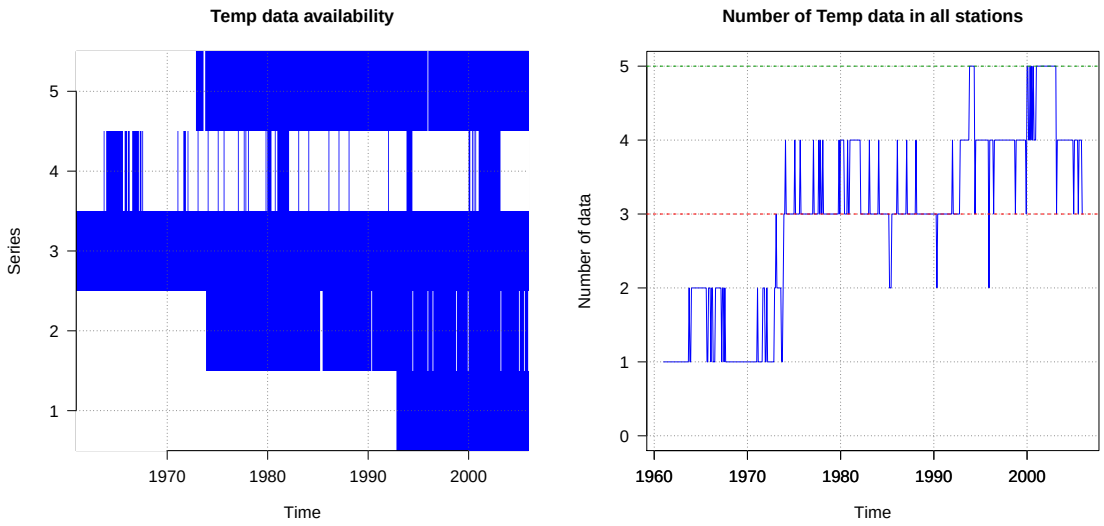


Figure 5: Data availability along the study period.

We see that series 3 is the only complete series, while series 1 is quite short and series 4 is very fragmented, with few data but distributed throughout the study period. (Other methods of homogenization would not work with so many missing data). Figure 5-right shows how many data exists at each time step. The dashed green and red horizontal lines indicate respectively what is the minimum desirable (5 data) and the minimum necessary (3 data) to detect suspicious data, because with only two data in a time step we would not be able to guess which one is wrong in case of discrepancy. The absolute minimum for *climatol* to work is that there must be at least one data at every time step, because the ultimate goal is to fill in all the missing data by spatial interpolation, and this cannot be achieved with no data. When this is the case, the process will stop with an error message, as in the former example, and if this is not due to too many anomalous data having been deleted, new series containing data in the critical periods will need to be added or the study period should be shortened to avoid the problem.

The following graphics focus on the correlations between the series and their classification into groups with similar variability, which are then plotted in a map. Correlations tend to decrease with increasing distance between stations. The higher the correlations, the greater the reliability of homogenization and filling of missing data. In particular, correlations should always be positive, at least within a range of reasonable distances. Otherwise, there will probably be geographic discontinuities that produce climatic differences. (For example, a mountain ridge can produce different precipitation regimes at both of its sides). This can be confirmed with the map of stations, in which groups of similar variability would be located in different areas, in which case it would be necessary to homogenize their series independently[4].

In areas of complex topography and/or low station density, the correlations may be far from optimal. In this situation, the filled in data will be individually affected by major errors, but there statistical parameters are expected to be acceptable.

To avoid processing too large correlation matrices, the number of series used for this cluster analysis is limited to 300 by default, and a random sample of this size will be used when the number of series exceeds that number, but the user can modify it via the `nclust` parameter.

After these initial graphs dedicated to verifying the data, the following pages of the document show plots of (spatial) anomalies standardized for each of the three following stages:

1. Detection in overlapping stepped windows
2. Detection in complete series

[4]The `datsubset` function allows you to get *climatol* files from a subset of stations.

3. Final anomalies of the homogenized series

The graphics of the first two stages show the series of anomalies in which changes in the mean have been detected, marking the breakpoints by a vertical dashed red line and labeling the value of the homogeneity test at its upper end. Final anomalies (in the third stage) are used to check if there have been evident uncorrected changes in the mean, in which case one would have to redo the homogenization setting an inhomogeneity threshold `inht` lower than the default value (25). If, on the contrary, the last cuts in the anomaly series seemed not justified, what we would do is to increase that threshold.

After the series of anomalies we can see the graphics of the reconstructed series and applied corrections. Figure 6 shows an example, with the anomalies of series 1 on the left and the reconstruction of complete series from the two homogeneous fragments on the right. The graphic of anomalies presents two additional lines at the bottom that inform on the minimum distance to the nearby data (in green) and the number of reference data used (in orange), both using the logarithmic scale of the right axis. The graphs of reconstructed series show their moving annual averages[5], unless the series are very short (up to 120 terms), in which case all values will be plotted. The original series are drawn in black[6], and in color the reconstructed ones.

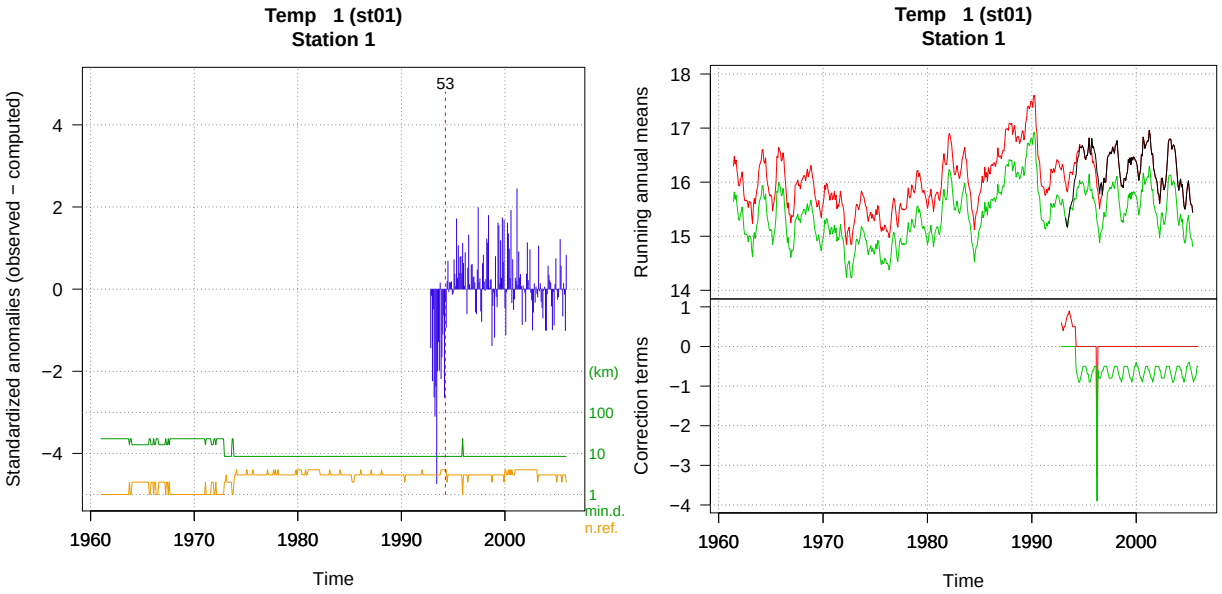


Figure 6: Example of detección of a shif in the mean (left) and reconstruccions of the series from every homogeneous fragment (rigth).

Histograms of the residual values of the homogeneity test used (SNHT or Cucconi) are also presented after each phase of the homogenization process. They can help to adjust the `inht` threshold if you want to repeat the process. But when we work with few series (as in our example) it will be difficult to discriminate which value best separates the homogeneous series from the inhomogeneous ones. In this case, it will be convenient to review the anomaly graphics, as previously commented. As a maximum of 4 reference data are used by default in the last phase (instead of the maximum of 10 used in the detection phases), the last series of anomalies may present test values higher than the threshold used, because of its greater variability due to the reduction in the number of references.

The histogram of anomalies that appears near the end of the document has already been previously commented when dealing with the homogenization of daily rainfall `Prec`. Finally, the last page of the graphics file contains a figure that indicates its quality or singularity, in which the stations are located according to their final

[5]Totals if the parameter `annual='total'` is given, which is recommended for precipitation.

[6]But the annual values cannot be calculated when some data is missing, and then they will appear with the color of the reconstructed fragment to which they belong.

Root Mean Standard Errors (RMSE) and homogeneity test values. RMSEs are calculated by comparing the estimated and observed data in each series. A high value may indicate poor quality, but could also be due to the station being located in a peculiar place with a different microclimate. Anyway, the homogeneous series of stations that share the common climate of the region will tend to cluster in the lower left of the graphic.

1.3.4 Other parameters of the `homogen` function

This function has a large number of parameters, as can be seen in its standard documentation. In general, it is not necessary to specify them, their default values being usually appropriate, but depending on the variable studied or the first homogenization results it can be convenient to change their values, especially in the following parameters:

- `dz.max` sets the thresholds for rejecting anomalous data or warning about suspicious data. Example: `dz.max=c(7,9)` will remove the data whose anomaly is greater than 9 standard deviations, and will list as suspect those with anomalies between 7 and 9 standard deviations. If you want to set different thresholds in the left tail of the distribution of anomalies, values can be given to the parameter `dz.min`.
- `inht` is the inhomogeneity threshold, that is, the value of the homogeneity test above which the series will be split. The review of the test histograms and anomaly graphics may suggest varying the default value, which is 25. If you want to force direct homogenization of daily series, this value will have to be increased by an order of magnitude (using, for example, '`inht=250, force=TRUE`').
- `std` is the type of normalization applied to the data. If the variable is detected as being strongly biased and bounded by zero, `std=2` will be used (the data will be divided by its mean value). As this is the recommended normalization for precipitation and wind speed, although it will be assigned automatically in the daily series, the same will not happen with the monthly values, for which it will be advisable to specify `std=2`. The default normalization is `std=3` (subtract the mean and divide by the standard deviation), valid for other variables such as temperature, relative humidity, atmospheric pressure, etc. A third type of normalization that the user can specify is `std=1`, which only centers the data by subtracting its mean value.
- `vmin` and `vmax` serve to limit the possible values that the data can take. `vmin=0` is automatically set when normalization is `std=2`, but for example for relative humidity it would be convenient to specify '`vmin=0, vmax=100`'.
- `nref` is the maximum number of nearby data to use to estimate those of the problem series. Up to 10 will be used by default (if they exist in each time step) in the first two stages, and up to 4 in the last one, but sometimes it may be convenient to change these values. For example, in the homogenization of daily rainfall, using 4 reference data will smooth the estimated data, thereby increasing the number of days of rain and decreasing the maximum values. That can be avoided by setting `nref=1`, although a very high value of the nearest series can produce one too high in the problem series if its average precipitation is much higher than that of the nearest series.
- `wd` specifies the distance at which the weight of the neighboring data is halved. By default, `wd=c(0,0,100)` is set, so that no weighting will be assigned to the data in the first two stages of detection of shifts in the mean and spatial anomalies, while in the third stage (filling in of all missing data) the data will lose weight with distance, as shown in figure 7, so that at 100 km they will weigh half.

1.3.5 Files of results

The only `homogen` output files not commented previously are the R binary files containing the homogenization results. They have the same base name as the others, but with an `rda` extension. The function's standard documentation gives details about their contents. The user can load the results of one of the above examples into the R memory space for its manipulation by means of the order:

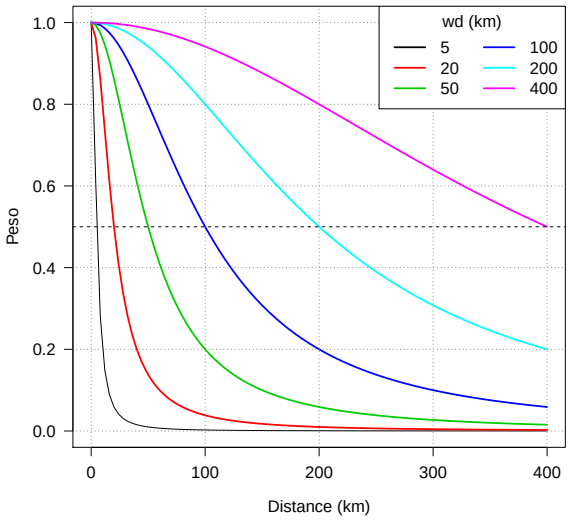


Figure 7: Variation of weights for different values of wd.

```
load('Temp_1961-2005.rda')
```

However, *climatol* provides post-processing functions to facilitate obtaining products from the homogenized series, so that in most cases it is not necessary to use the `*.rda` files directly, as we will see below.

1.4 Obtaining products from the homogenized data

Although the user can load the results of the homogenization as noted above, *climatol* provides the post-processing functions `dahstat` and `dahgrid` for easily obtaining frequently used products from the homogenized series.

1.4.1 Statistical summaries and homogenized series

The homogenized series can be dumped into two CSV text files using the `dahstat` function by specifying the `stat='series'` parameter. In the previous examples `Temp` were monthly temperatures and `Prec` were daily precipitations, from which the monthly values were obtained and saved as variable `Prec-m`. So we can get the homogenized series (adjusted from the last homogeneous fragment backwards) through:

```
dahstat('Temp', 1961, 2005, stat='series')   #monthly temperatures
dahstat('Prec', 1981, 1995, stat='series')   #daily precipitations
```

Each of these commands generates two CSV files. Those named `*_series.csv` contain the homogenized series, while the `*_flags.csv` files have flags indicating whether the data is observed (0), filled in (1, originally absent) or corrected (2, due to inhomogeneities or excessive anomaly). With a similar order applied to 'Prec-m', the homogenization series of monthly rainfall could be obtained, but it is better to calculate these series from the daily ones, since the absence of daily data at the time of calculating the monthly aggregates will cause some differences.

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

```
dahstat('Prec',1981,1995) #monthly averages (the default statistic)
dahstat('Prec',1981,1995,stat='tnd') #trends and p-values
dahstat('Prec',1981,1995,stat='q',prob=.2) #first quintile
```

This function includes parameters to choose a subset of series, either giving a list with the desired codes, as for example with `cod=c('p064','p084')`, or by specifying that we want the series reconstructed from the longest subperiod (`long=TRUE`). We can also ask for statistics for all series (reconstructed from all homogeneous fragments) using the parameter `all=TRUE`.

1.4.2 Series of homogenized grids

The other postprocessing function, `dahgrid`, generates grids calculated from the homogenized series (without using filled in data). But before calling this function, the user must define the desired limits and resolution of the grid, as in this example that uses the results of the `Temp` homogenization:

```
grd <- expand.grid(x=seq(-2.7,-2.5,.025),y=seq(38.8,39.1,.025)) #desired grid
#this command requires the sp package to be installed:
sp::coordinates(grd) <- ~x+y #convert the grid to a spatial object
```

The R function `expand.grid` is used to define the sequences of X and Y coordinates, and then the function `coordinates` (from the `sp` package) is applied to convert the grid, saved under the name `grd` (you could have used anyname), into a spatial class object.

Homogenized grids can now be generated (in NetCDF format) with:

```
dahgrid('Temp', 1961, 2005, grid=grd) #monthly grids
```

These grids have been built with dimensionless normalized values. You can get new grids with the original units (°C in the example) through external tools, such as the *Climate Data Operators (CDO)*, taking advantage of the fact that `dahgrid` has also saved grids with the averages (`*_m.nc`) and standard deviations (`*_s.nc`). Thus, if the `CDOs` are installed on your system, we can call them from R with:

```
command <- paste('cdo add -mul Temp_1961-2005.nc Temp_1961-2005_s.nc',
  'Temp_1961-2005_m.nc Temp-u_1961-2005.nc')
system(command)
```

But the new grids contained in `Temp-u_1961-2005.nc` (we could have given the output file any name, respecting the `nc` extension) will only be based on geometric interpolations. Therefore, if there are mountains devoid of data in our domain, expected climatic variations will not be reflected in the grids. To get a better representation of the climate of the studied area, better grids of means `Temp_1961-2005_m.nc` and standard deviations `Temp_1961-2005_s.nc` should be obtained by geostatistical methods before using them to obtain the grids of values with their original units.

1.5 Frequently asked questions about the `homogen` function

The examples shown above discuss the most common applications of the `climatol` homogenization functions. However, doubts may arise regarding how to proceed when dealing with other climatic variables or temporal resolutions. This section is dedicated to solving the most frequent doubts.

1.5.1 How to save the results of different tests

If you run `homogen` with different parameters to explore which give better results, you can avoid overwriting the previous outputs by renaming them with the help of the `outrename` function. For example, the following command will rename all `Temp_1961-2005*` output files to `Temp-old_1961-2005*`:

```
outrename('Temp', 1961, 2005, 'old')
```

1.5.2 How to change the cutoff level in clustering analysis

In the clustering analysis that *climatol* performs in its initial check of the data, the number of clusters is determined automatically. Looking at the dendrogram (in the first graphs of the PDF output document) a different cutoff level can be chosen by setting the `cutlev` parameter.

1.5.3 How to use reanalysis series as references

When the series are very fragmented and some time steps of our study period do not have data in any of them, or when we try to homogenize an isolated series, a solution is to use series from reanalysis products to serve as references that provide data in those critical gaps.

Although the appearance of new observation systems (such as satellites) introduces inhomogeneities in the amount of data available for assimilation by the models, we can consider that the reanalysis products are generally more homogeneous than the observed series. To use these products as references, the series of one or more grid points located in the study domain should be added to the `*.dat` data file, and the coordinates of those points added to the `*.est` station file. Their codes should start with an asterisk (example: `*R43`) so that quality and homogeneity checks will skip those more reliable series.

1.5.4 What homogenized series should I use?

Most homogenization methods return the series adjusted from the last homogeneous subperiod, but *climatol* generates complete reconstructions from each subperiod (unless it is too short for such a reconstruction to be reliable). Therefore, the user may wonder which one to use in his climate study. The answer depends on the objective of the investigation. To obtain normal values with which to calculate anomalies from new incoming data for climate monitoring, the series adjusted from the last homogeneous sub-period (the default option of `dahstat`) should be used. But if the goal is to make maps, all series should be considered (by adding `all=TRUE` to the `dahstat` parameters), then choosing those that best fit the spatial variability of the map scale, and ignoring those that appear as are affected by local microclimates.

1.5.5 The process is taking too long

This can happen if we are processing many, very long series with many shifts in the average. Example: 400 daily thermometric series from 1951-2020. The direct homogenization of these long daily series can take days, especially if the `inht` and `dz.max` parameters have not been given high enough values, because then the series would be suffering a high number of splits and rejecting too much data which will then have to be filled in.

If the recommended procedure of first homogenizing the monthly series obtained with the `dd2m` function is followed, the processing time will be much shorter. But in any case, it is worth considering whether it is necessary to homogenize all the series at the same time, because it is probably better to divide our data into smaller sets, grouping the series according to climatically more homogeneous subregions. (The `datsubset` function can be used to generate the files for a selected group of stations, which can be based on the clustering analysis generated by `homogen`, adjusted by the user based in his knowledge of climate and physiography).

1.6 References

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.

Cucconi O (1968): Un nuovo test non parametrico per il confronto tra due gruppi campionari. *Giornale degli Economisti*, 27:225-248.

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

WMO (2020): Guidelines on Homogenization. WMO Nº 1245, 54 pp., Geneva, Switzerland, ISBN 978-92-63-11245-3.

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.

2 Other functions

In addition to the homogenization functions explained so far, *climatol* also provides some utilities and graphic products that will be briefly shown below (see the standard documentation for all the details about its use).

2.1 Utilities

- **`fix.sunshine`** is used to prune any excess of sunshine hours that may have occurred when adjusting the daily series (see example in the help of the function).
- **`QCthresholds`** allows obtaining, for each daily (or subdaily) series, monthly quantiles of extreme values, of increments between consecutive values and of sequences of identical values. These quantiles can be used to implement quality control alerts in Climate Data Management Systems. Example for the `Prec` series previously homogenized (a minimum value is set to avoid counting long sequences of consecutive zeros):

```
QCthresholds('Prec_1981-1995.rda',minval=0.1)
```

```
=========== thr1: Monthly quantiles of the data
--------- Station p064
         1    2    3    4    5    6    7     8    9    10   11   12
0      0.0  0.0  0.0  0.0  0.0  0.0  0.0   0.0  0.0   0.0  0.0  0.0
0.001  0.0  0.0  0.0  0.0  0.0  0.0  0.0   0.0  0.0   0.0  0.0  0.0
0.01   0.0  0.0  0.0  0.0  0.0  0.0  0.0   0.0  0.0   0.0  0.0  0.0
0.99  35.1 40.9 44.5 33.1 31.6 31.0 31.9  49.5 46.7  60.2 63.9 42.9
0.999 54.1 50.8 56.5 54.2 69.4 44.6 48.3  91.2 71.7 120.2 91.9 56.1
1     54.3 53.8 57.1 62.7 92.8 49.0 54.2 110.1 73.1 140.0 93.8 57.1
--------- Station p084
```

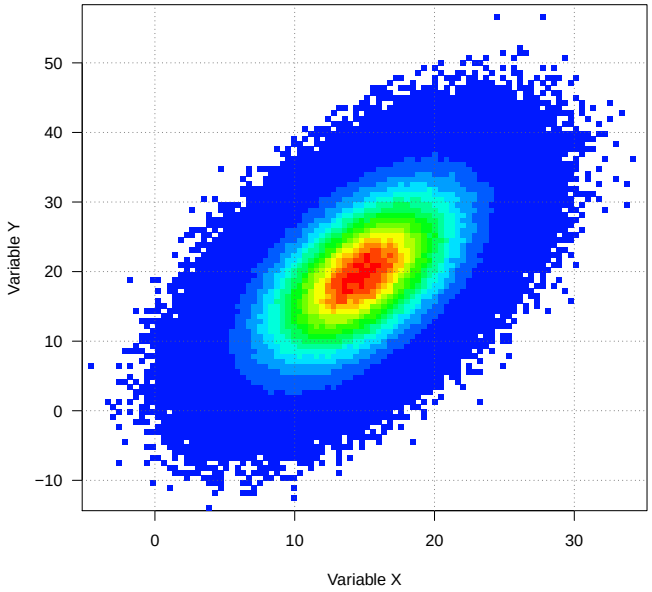
```
         1    2    3    4    5    6    7    8     9    10   11   12
0      0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0   0.0   0.0  0.0  0.0
0.001  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0   0.0   0.0  0.0  0.0
0.01   0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0   0.0   0.0  0.0  0.0
0.99  24.1 31.3 29.0 27.9 30.0 29.7 30.9 36.6  38.0  51.5 49.2 30.1
0.999 39.8 42.2 40.7 46.3 50.1 46.5 40.4 67.5  74.6  87.8 67.2 39.8
1     42.2 48.2 44.2 52.8 51.1 51.6 42.0 76.7 100.4 116.5 71.3 41.1
--------- Station p082
         1    2    3    4    5    6    7    8    9   10   11   12
0      0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
0.001  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
0.01   0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
0.99  17.9 24.2 21.4 23.1 21.0 26.4 28.2 33.2 29.3 38.2 31.8 19.7
0.999 24.3 32.7 29.9 31.5 34.4 34.0 47.1 48.0 61.4 63.4 48.0 24.5
1     26.9 37.6 30.4 33.8 36.8 34.0 47.9 52.0 78.7 78.0 51.2 26.3

=========== thr2: Quantiles of the first differences
     0.99 0.999     1
p064 46.1  84.0 129.4
p084 37.8  66.9 103.6
p082 28.8  49.2  78.5

=========== thr3: Quantiles of run lengths of constant values >= 0.1
     0.99 0.999 1
p064    2     3 4
p084    2     3 3
p082    2     3 3

Thresholds thr1,thr2,thr3 saved into QCthresholds.Rdat
(Rename this file to avoid overwriting it in the next run.)
```

With `load('QCthresholds.Rdat')` we would have these results in the R session memory and we could write them in the appropriate format to import them into the Climate Data Management System to implement alerts of suspicious values.

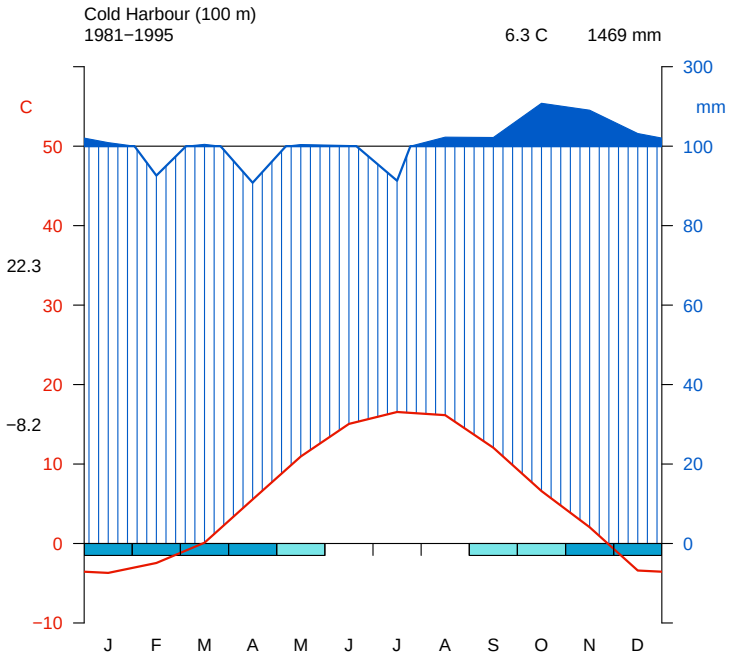
2.2 Graphic products

In this section, only examples of the functions that produce graphics useful in climatology will be shown. The standard *climatol* documentation shows the details of each of them.

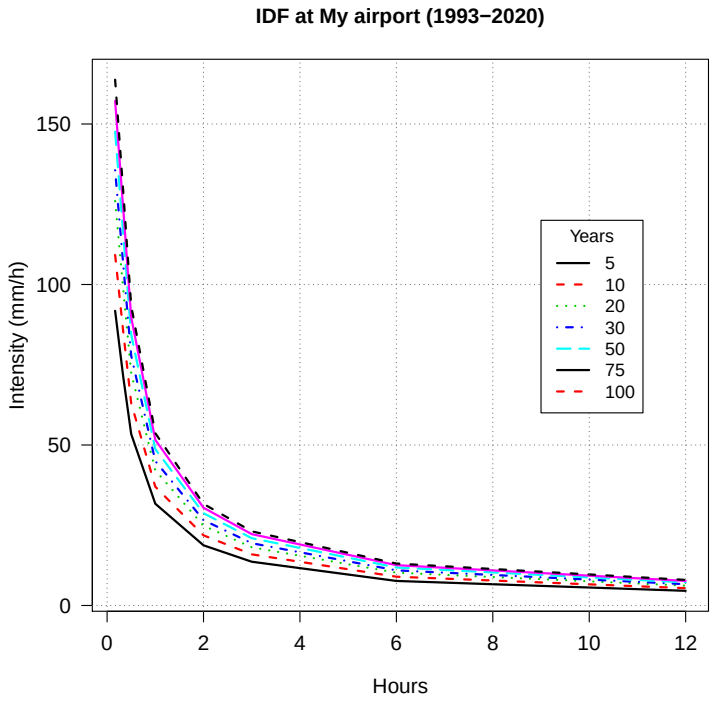
2.2.1 dens2Dplot: Two-dimensional scaterplot



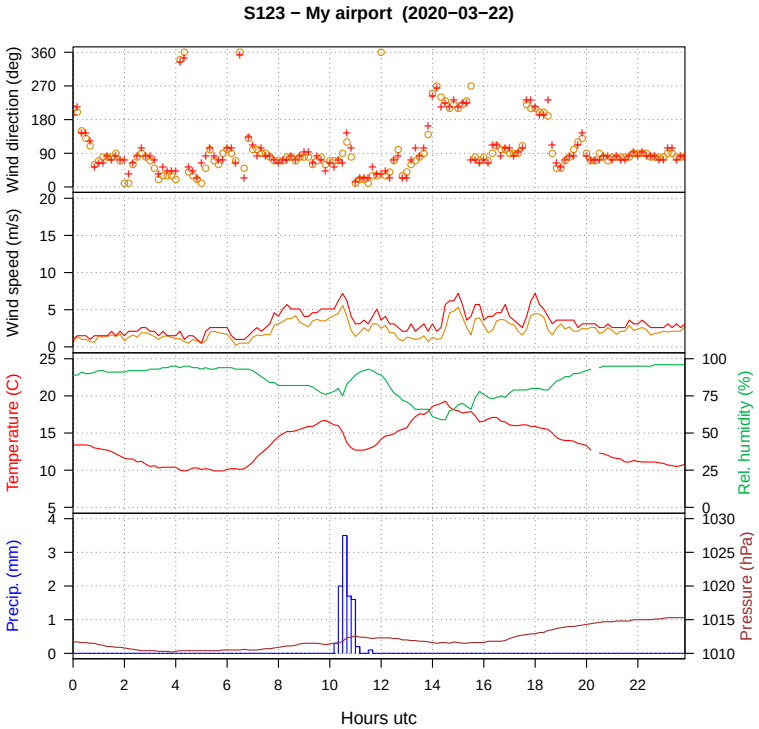
2.2.2 diagwl: Walter & Lieth diagram



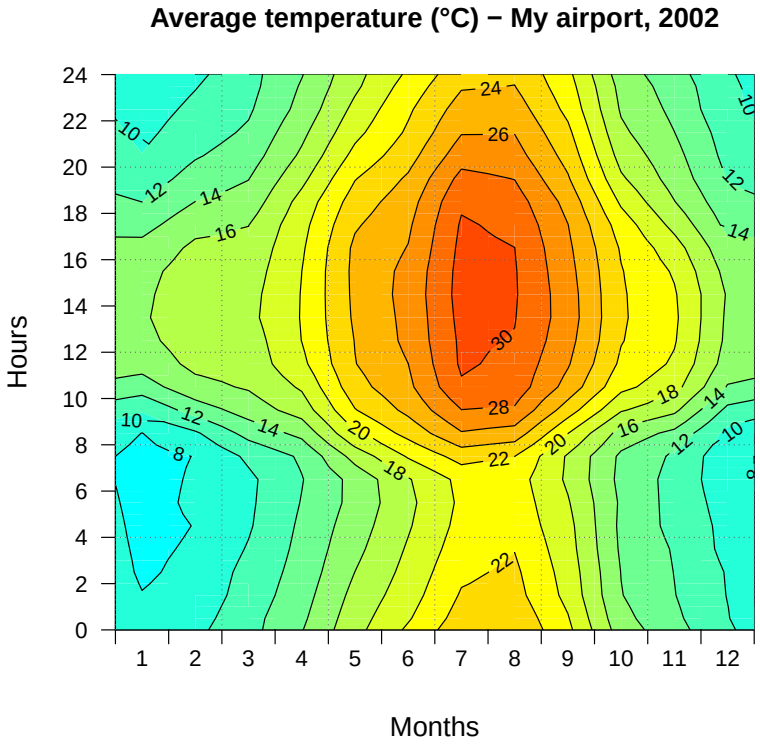
2.2.3 IDFcurves: Intensity-Duration-Frequency diagram from subdaily precipitation data



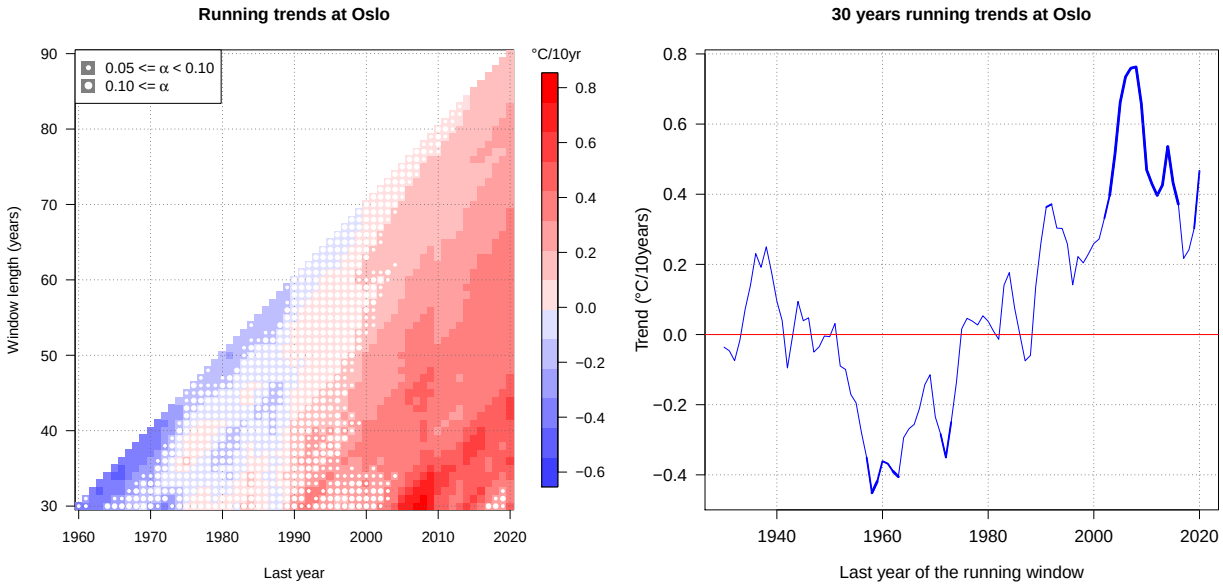
2.2.4 meteogram: Meteogram of 1 day from Automatic Weather Station data



2.2.5 MHisopleths: Isopleths in a Months-Hours diagram



2.2.6 runtnd: Diagrams of running trends



2.2.7 windrose: Wind rose from wind direction and speed data