EXTRAQC QUALITY CONTROL SOFTWARE.

By Enric Aguilar[1] and Marc Prohom[2]

1: Center for Climate Change, C3, Universitat Rovira i Virgili de Tarragona, Spain
2: Climate Change Team, Climate Area, Servei Metorològic de Catalunya, Barcelona.

Introduction.

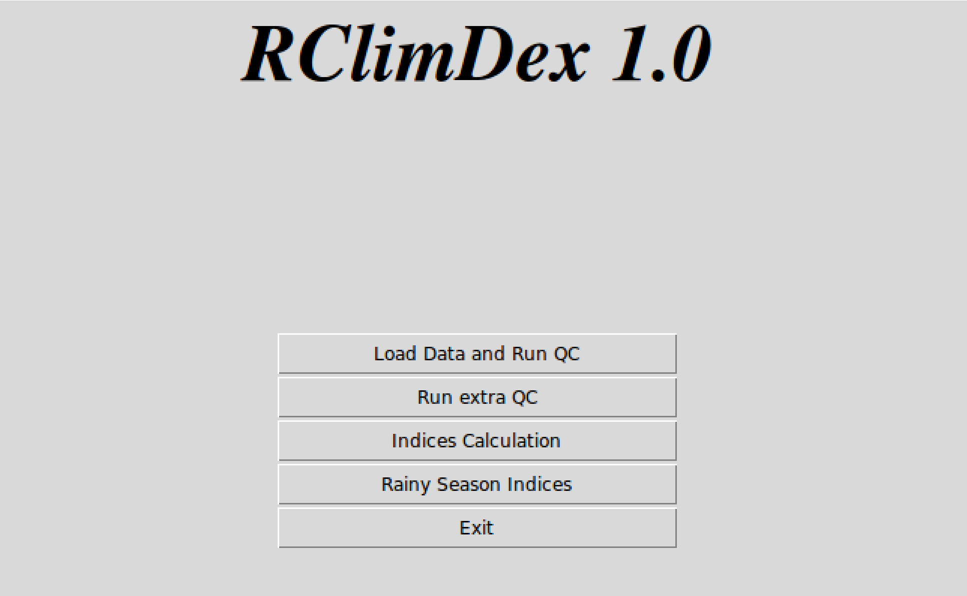
The EXTRAQC routines are a set of R-coded functions for quality control. Although they can be used in a stand alone version, which allows more flexibility, they have been integrated by the authors into the widely used ETCCDI's software R-Climdex, to produced an unofficial version called R-Climdex-Extraqc.
These functions are prepared after the long experience accumulated by the authors in quality controlling temperature and precipitation data with R-Climdex and other software and in training third users to do so. EXTRAQC routines focus mainly on temperature data and include the following tests:

- Duplicate dates control
- Rounding problems evaluation
- Out of range values, based on fixed threshold values
- Outliers, based on Interquartile Range exceedance
- Interdiurnal differences based on fixed threshold values
- Coherence between maximum and minimum temperatures (Tmax > Tmin)
- Consecutive equal values control

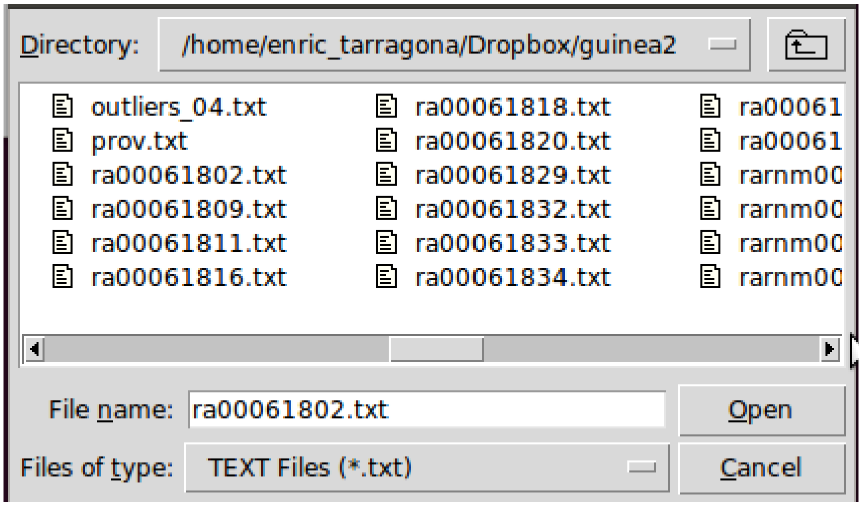
EXTRAQC produces numerical and graphical output, automatically stored in your computer, following R-Climdex conventions. The next section describes how to run EXTRAQC and the expected output.

Running EXTRAQC

Once in R, after loading the source file *rclimdex_extraqc.r*, press the second button, labeled ***Run extra QC***.



After doing so, a the software will present a pop-up window to navigate through your disk and select the file you are intending to quality control



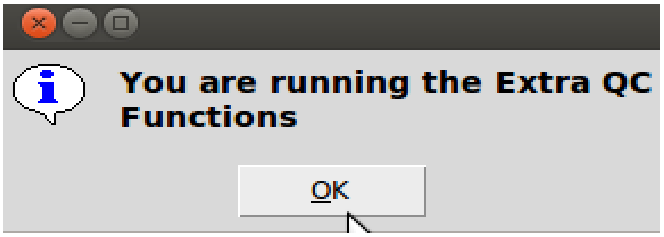
Directory: /home/enric_tarragona/Dropbox/guinea2

outliers_04.txt
prov.txt
ra00061802.txt
ra00061809.txt
ra00061811.txt
ra00061816.txt
ra00061818.txt
ra00061820.txt
ra00061829.txt
ra00061832.txt
ra00061833.txt
ra00061834.txt
ra00061
ra00061
rarnm00
rarnm00
rarnm00
rarnm00

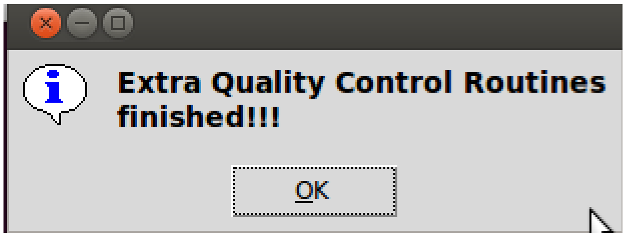
File name: ra00061802.txt

Files of type: TEXT Files (*.txt)

After completing your selection, a message will appear indicating that you are running the extra qc functions (press OK). After quality control is finished, a pop-up window will warn you. Loading and running EXTRAQC needs the same data format than any other R-Climdex functionality and is subjected to the same potential format errors, to be solved identically than with R-Climdex.



You are running the Extra QC Functions



Extra Quality Control Routines finished!!!

Once EXTRAQC has finished for a first time, your working directory will present a new folder named ***extraqc***, where the different tests numerical results an plots will be stored.

Your extraqc folder will contain the following files:

- 3 pdf files, with graphical information on data quality:
  - mystation_boxes.pdf
  - mystation_boxseries.pdf
  - mystation_rounding.pdf

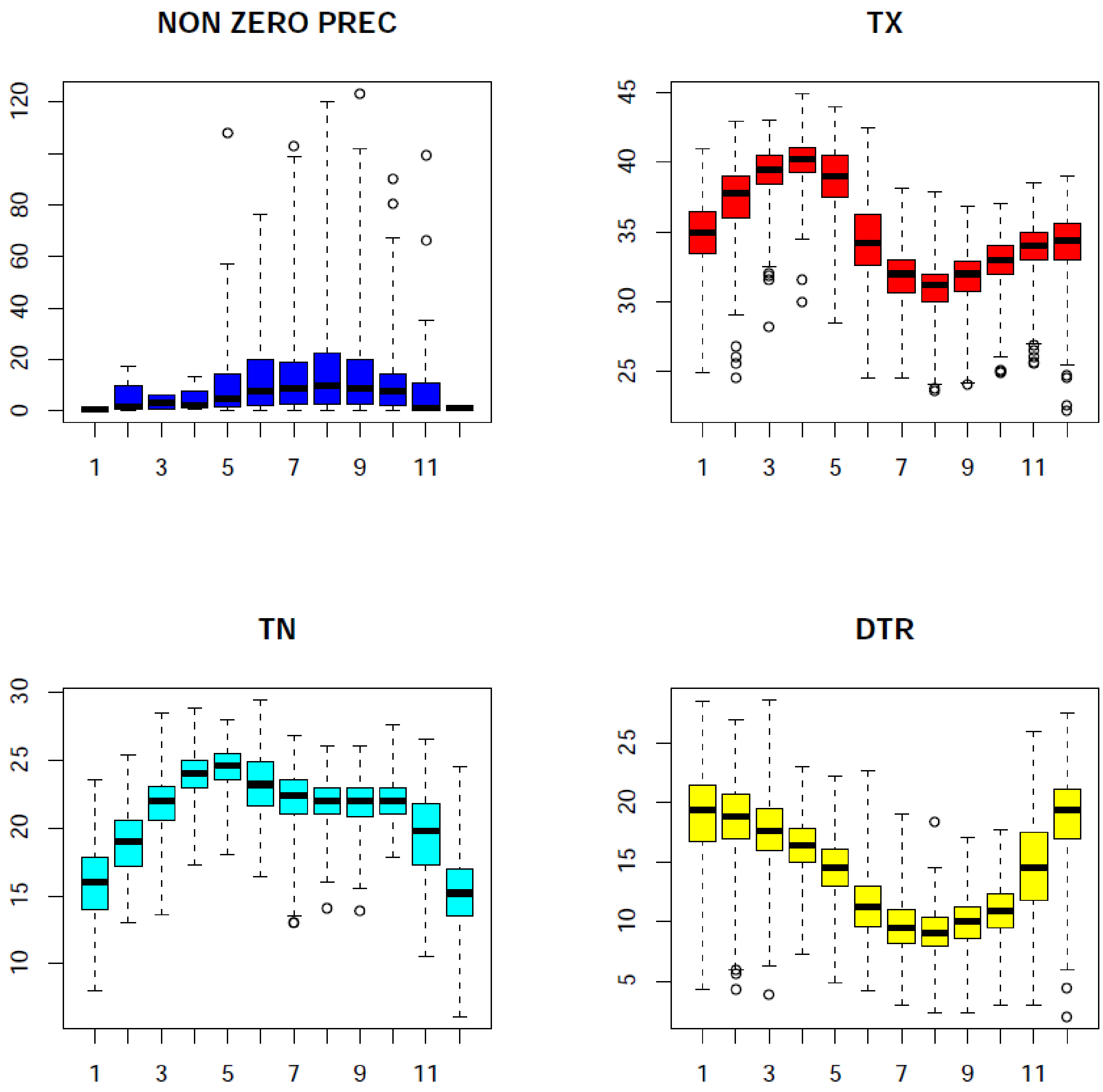
- 8 text files with numerical information on data quality
  - mystation_duplicates.txt
  - mystation_outliers.txt
  - mystation_tmaxmin.txt
  - mystation_tx_flatline.txt
  - mystation_tn_flatline.txt
  - mystation_toolarge.txt
  - mystation_tx_jumps.txt
  - mystation_tn_jumps.txt

These files offer useful information to flagg data quality control problems.

QUALITY CONTROL FILES CONTENTS:

mystations_boxes.pdf:

As can be seen in the example below, the mystation_boxes.pdf file, produces boxplots of temperature and precipitation data flagging as outliers (round circles) all those values falling outside a range with p25 – 3 interquartilic ranges (lower bound) and p75 + 3 interquartilic ranges (upper bound) for temperature. The same ranges, but using 5 IQR are used for precipitation.

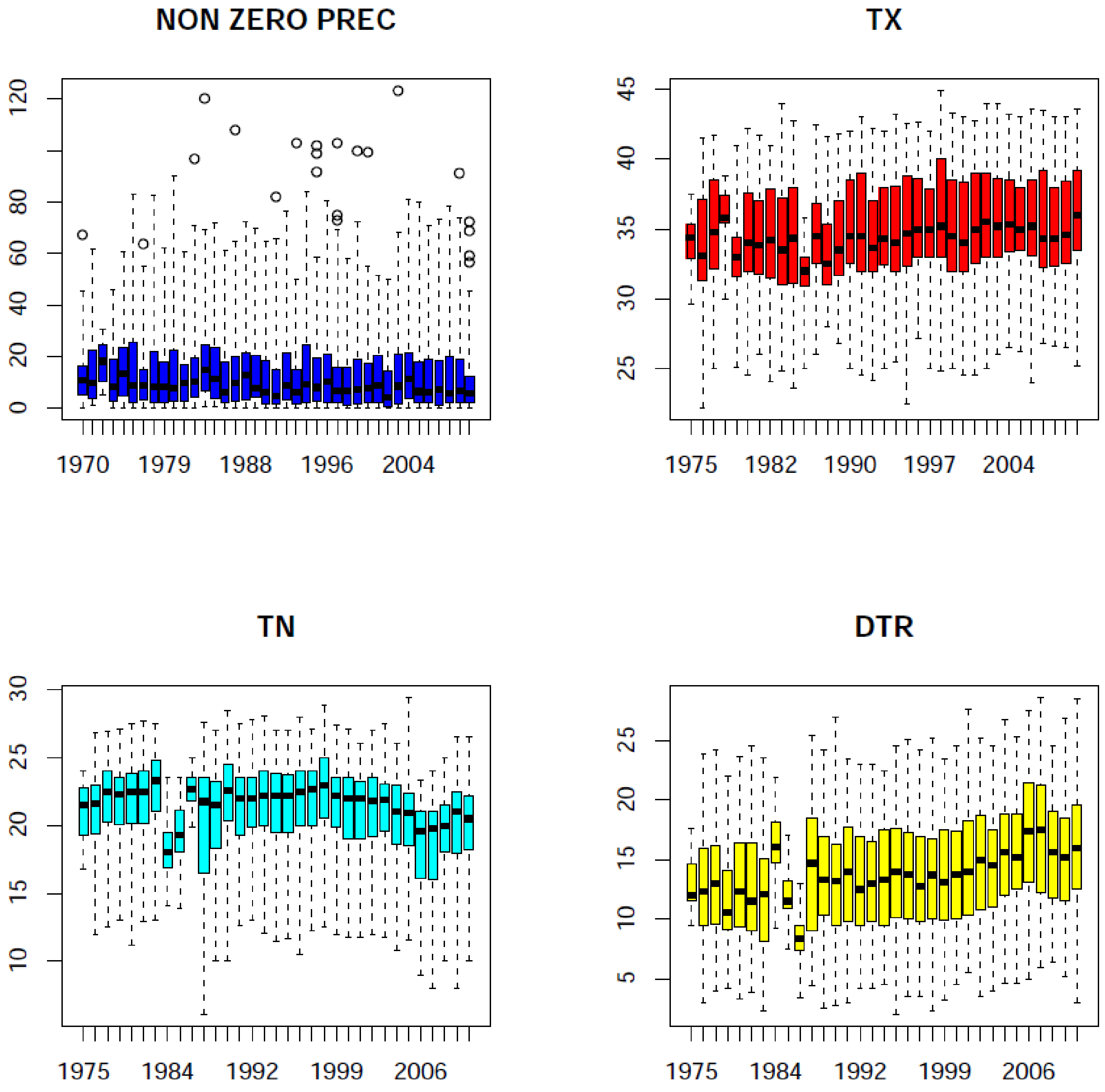


The values identified by this graphical quality control, are sent to a simple text file, the ***mystation_outliers.txt***. The file lists the outliers grouped under the element that produced the inclusion of the record in the file and specifying the margin (upper bound or lower bound) that is surpassed. So, under ***pc up*** appear those values that represent a precipitation outlier; under ***tx up*** those that represent a maximum temperature higher than p75+3IQR; under ***tx low*** outliers that represent an observation lower than p25-3IQR. The explanation given for tx, applies to tn and dtr. **The advantage of this approach is that the detection of this percentile based outliers is not affected by the presence of larger outliers, so ONE RUN OF THE PROCESS IS ENOUGH!**
Each record listed includes: ***year, month, day, precip, tx, tn and dtr***. ***NA*** stands for ***not available***.

```
pc up
1987 5 9 108.1 32.5 NA NA
1993 7 24 102.7 32.1 22.1 10
2003 9 21 123.2 33 20.2 12.8
1976 10 24 80.2 26.1 19.6 6.5
1980 10 10 90.1 33 21.5 11.5
1976 11 1 66.2 32.8 20.5 12.3
2000 11 11 99.3 32.5 21.4 11.1
tx up
tx low
1980 2 5 2.1 25.5 21.2 4.3
1980 2 6 0.9 24.5 18.6 5.9
1980 2 7 0.1 26.8 18.5 8.3
1992 2 13 0 26 20 6
1977 3 27 0 31.8 25.5 6.3
1981 3 21 5.9 28.2 24.4 3.8
1981 3 22 0.4 31.6 20.9 10.7
1982 3 11 0 32 19.7 12.3
1976 4 18 0 31.6 24.3 7.3
1987 4 5 0 30 NA NA
1985 8 5 1.6 23.8 NA NA
1985 8 25 1.6 23.6 NA NA
2006 9 23 0.9 24 19 5
tn up
tn low
1989 7 16 7.5 31.5 13 18.5
1989 7 21 62 31 13 18
1979 8 6 3.6 32.5 14.1 18.4
1985 9 12 0 31 13.9 17.1
tr up
1979 8 6 3.6 32.5 14.1 18.4
tr low
1980 2 5 2.1 25.5 21.2 4.3
1980 2 6 0.9 24.5 18.6 5.9
2005 2 14 0 29 23.4 5.6
1981 3 21 5.9 28.2 24.4 3.8
1976 12 19 0 22.2 17.8 4.4
```

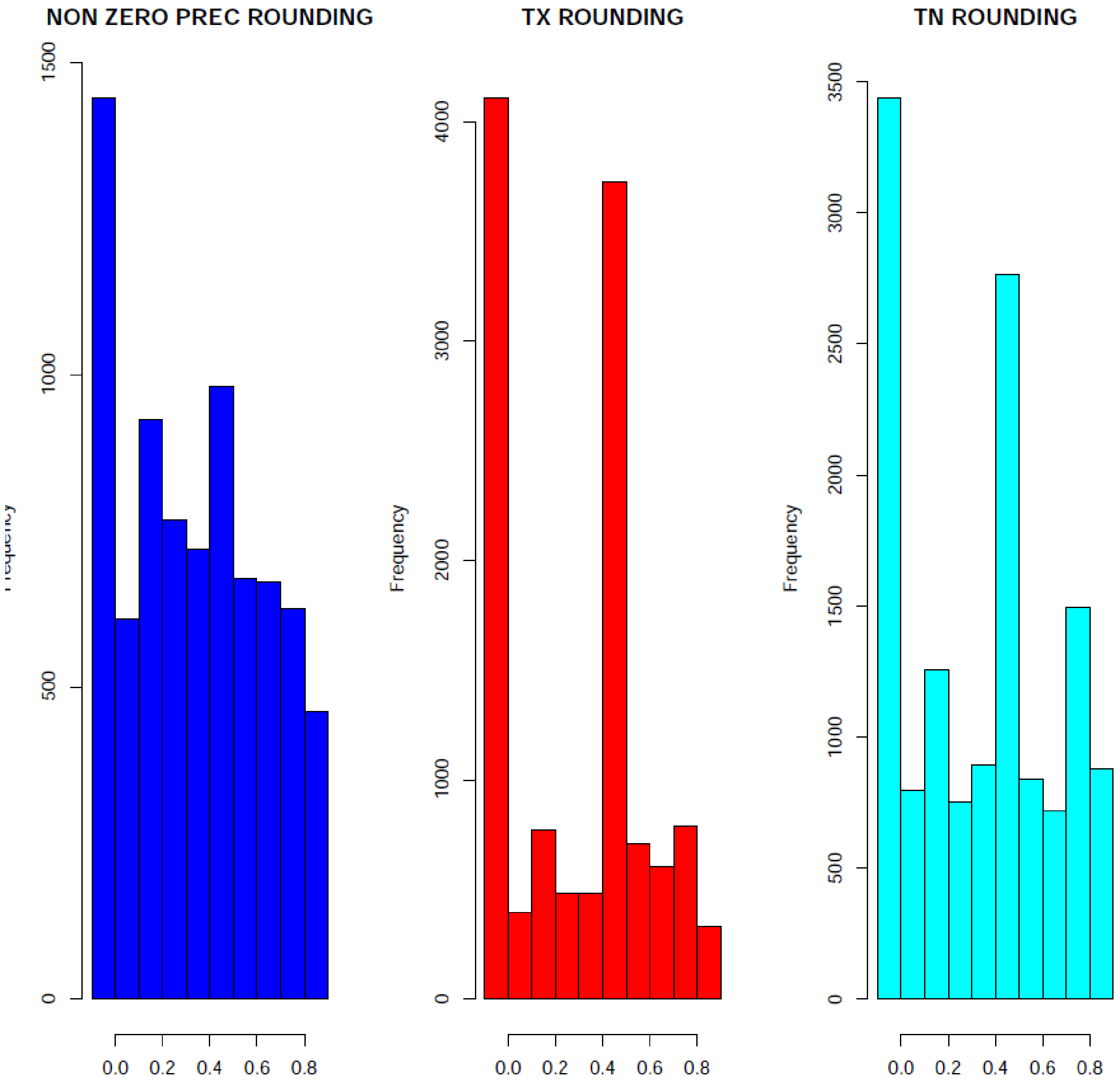
1995 12 15 2.2 22.5 20.5 2

The graphic file ***boxseries.pdf*** (which does not have a numerical counterpart) produces annual boxplots. This file is useful to have a panoramic view of the series and be alerted of parts of the series which can be problematic (see values around 1984 in the example)



The third graphical quality controlis the file ***mystation_rounding.pdf***. It looks at rounding problems by plotting the values after the decimal point. It shows how frequently each of the 10 possible values (.0 to

.9) appears. It is expected that .0 and .5 will be more frequent (although there is no statistical reason for this!), but if the rest of the values are not represented, one might think of discarding the series or using an statistical approach to restitute the 1/10th of degree resolution .



The ***mystation_tn_flatline.txt*** and ***mystation_tx_flatline.txt*** report on ocurrences of 4 or more equal consecutive values in, respectively, tmax and tmin. The text file outputs a line for each sequenece of 4 or more consecutive equal values, with 5 fields: year, month, day, value and length of sequence. In the example below all sequences are 4 values long. The date specified belongs to the end of the sequence.

```
1988 12 24 13 4
1989 8 10 23.5 4
1989 8 29 23 4
2009 9 9 21 4
```

Looking at the data, the first sequence identied by the QC test is shown below

```
1988    12      21      0.00    33.0    13.0
1988    12      22      0.00    32.5    13.0
1988    12      23      0.00    31.0    13.0
1988    12      24      0.00    32.0    13.0
```

The file ***mystation_duplicates.txt*** includes all dates wich appear more than once in a datafile. Each record contains 3 fields: year, month, day and refers to the duplicated date. In the listing below, one can see that days between 1958/08/21 and 1951/09/02 are duplicated.

```
1951 8 21
1951 8 22
1951 8 23
1951 8 24
1951 8 25
1951 8 26
1951 8 27
1951 8 28
1951 8 29
1951 8 30
1951 8 31
1951 9 1
1951 9 2
```

The file ***mystation_toolarge.txt*** reports precipitation values exceeding 200 mm (this and any other threshold can be easily reconfigured before excecution) and temperature values exceeding 50 ºC. Fields included are year, month, day, precip, tmax, tmin. In the example below, a temperature of 99.9 does not pass this QC test.

```
1951 8 23 0 99.9 21.7
```

The files ***mystation_tx_jumps.txt*** and ***mystation_tn_jumps.txt*** will list those records where the temperature difference with the previous day is greater or equal than 20 ºC. The outputed fields are year, month, day and difference with the following day. In the example below, the 99.9 temperature detected by the previous QC, appears here twice, as produces two excessive interdiurnal differences.

```
1951 8 22 69.4
1951 8 23 72.3
```

Finally, the ***mystation_tmaxmin.txt*** file, includes all those cases where maximum temperature is lower than minimum temperature. Fields included are year, month, day, precip. tmax, tmin.

```
1951 8 27 0.7 21.1 29.1
1951 8 28 27.8 18.7 27.5
```