**Visual Quality Control for Detecting Systematic Errors**

# Introduction

The Visual Quality Control for Detecting Systematic Errors was created in the project DECADE (Data on climate and Extreme weather for the Central AnDEs), financed by the program r4d (Swiss Programme for Research on Global Issues for Development).

In a first step, the quality control (QC) tests and visualisations are suitable tools for data exploration. In a second step, systematic data quality problems that are not detected by most other QC methods can be identified and classified.

The QC method was developed for daily maximum and minimum temperature (TX and TN, respectively) and precipitation (PRCP), but it is applicable to sub-daily temperature and precipitation observations as well.

The QC approach runs on R, and is therefore widely applicable.

This manual provides brief instructions on the application of the QC method and demonstrates how to interpret the results. Section 2 describes the required data format, Section 3 introduces the settings that can be chosen by the user in order to meet specific aims, and Section 4 illustrates and explains the correct interpretation of the visualisations of the QC method.

Further information on frequent systematic errors can be found here (open access):

Hunziker, S., Gubler, S., Calle, J., Moreno, I., Andrade, M., Velarde, F., Ticona, L., Carrasco, G., Castellón, Y., Oria, C., Croci-Maspoli, M., Konzelmann, T., Rohrer, M., and Brönnimann, S.: Identifying, attributing, and overcoming common data quality issues of manned station observations, International Journal of Climatology, 37, 4131-4145, Doi: 10.1002/joc.5037, 2017.

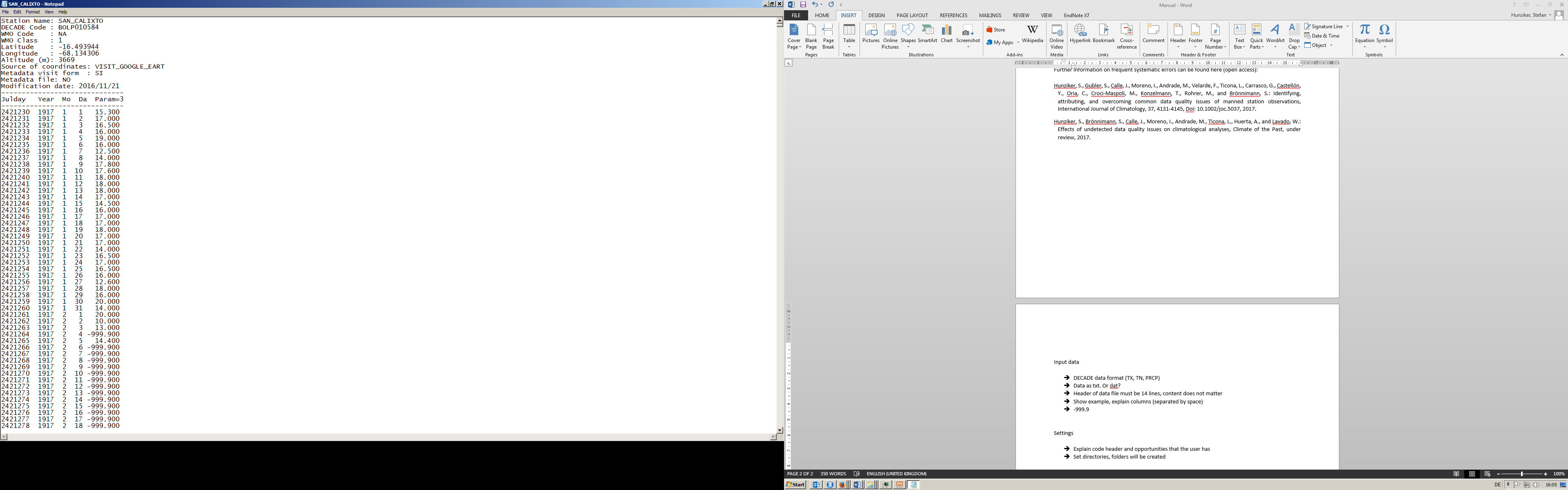
Hunziker, S., Brönnimann, S., Calle, J., Moreno, I., Andrade, M., Ticona, L., Huerta, A., and Lavado, W.: Effects of undetected data quality issues on climatological analyses, Climate of the Past, under review, 2017.

# Input data

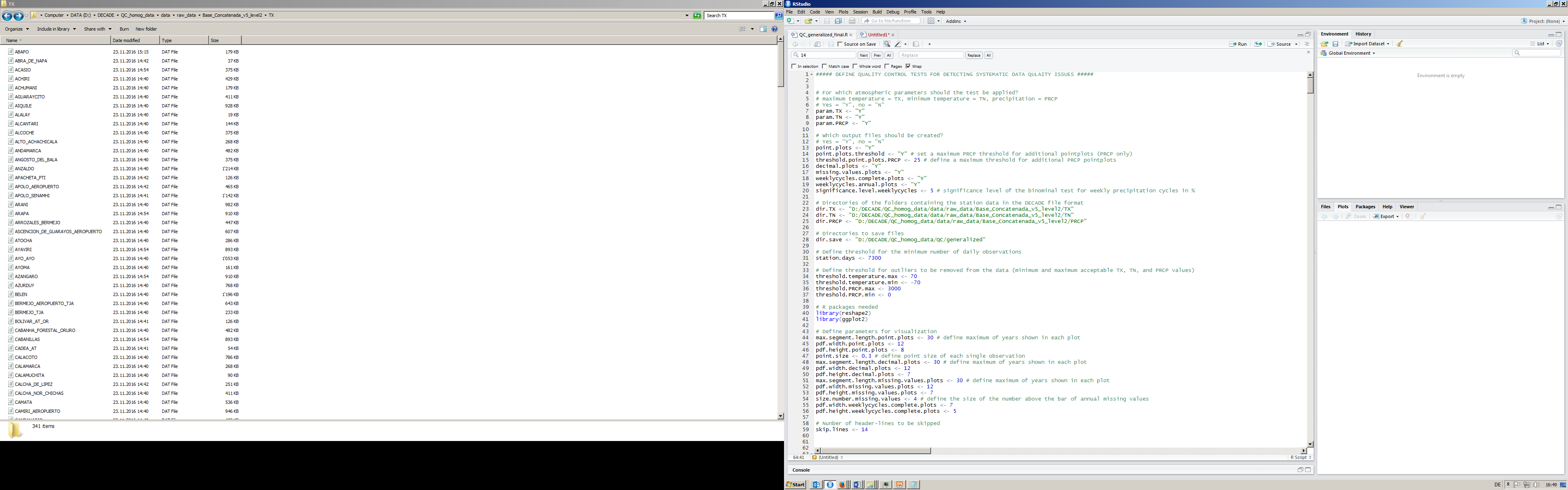
The input data format is based on the DECADE data format. The following file format is required to run the QC:

* One dat-file for each station and parameter (the filename will be shown on the visualisations)
* A header is allowed, but the correct number of header lines must be specified in the settings
* Columns must be separated by a space or spaces
* The files must contain the following four columns (more columns are allowed and will not be used):
  + First column: For instance the Julian day (not used for the QC)
  + Second column: Year
  + Third column: Month
  + Fourth column: Day
  + Fifth column: Parameter (temperature or PRCP)
  + The missing value code is -999.9

The following figure illustrates the required file format:



# Settings



**1**

**2**

**3**

**2**

**4**

**5**

**6**

**7**

**8**

**9**

**11**

**10**

**11**

**10**

**10**

**12**

**11**

**11**

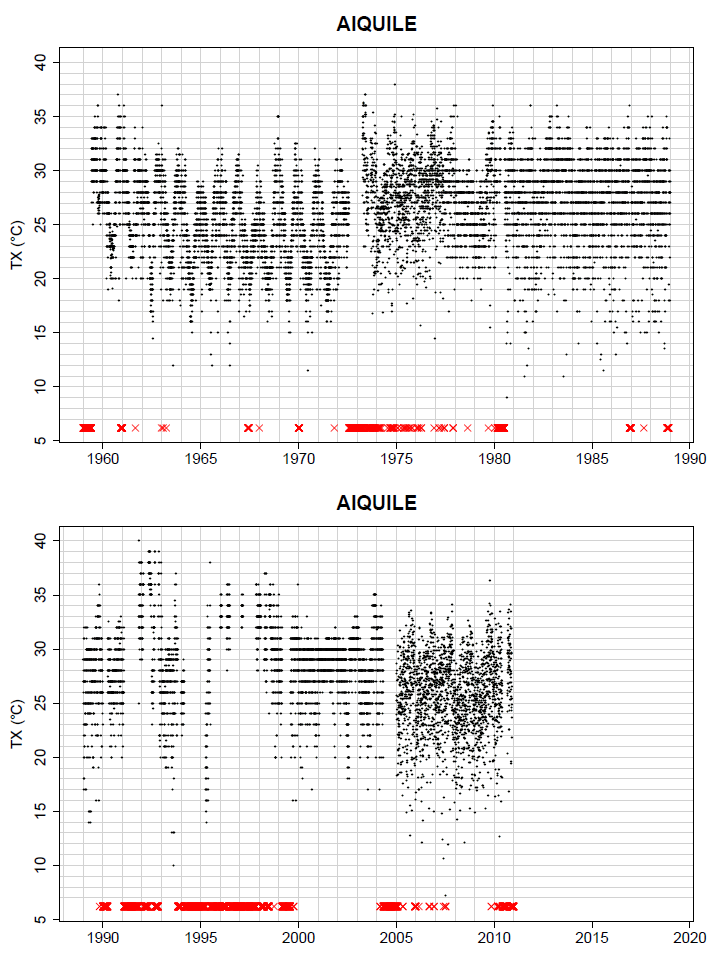
**13**

1. Choose the variables that should be quality controlled.
2. Choose the QC tests you want to run.
3. Define a threshold for daily PRCP events illustrated in this diagram. This plot allows to detect errors (particularly low PRCP gaps) in the record of low precipitation events. It is an addition to the plot containing all observations (point.plots).
4. Choose the significance level for the binominal test of weekly PRCP cycles.
5. Set the directories to the folders containing the data.
6. Choose where the results of the QC should be saved. Folders containing the figures of each test will automatically be created.
7. Stations with very short or incomplete records can be excluded by defining a minimum number of valid observations of each time series.
8. Define thresholds for the exclusions of extreme outliers that disturb the interpretation of the visualisations.
9. R-packages required to run the QC.
10. Define the number of years that should be maximally shown in each visualization (shorter time series segments may allow easier detection of errors). If the station records exceeds the defined threshold, several plots will be shown in one pdf-file.
11. Change the appearance of the visualization by setting the width and height of the pdf-files created.
12. The number of annually missing values is shown above the bars in the missing.values.plots. Adapt the size of these numbers.
13. Specify the number of header rows in the data files. The number must be correct in order to run the QC and to not miss reported observations.

# Output

## Temperature

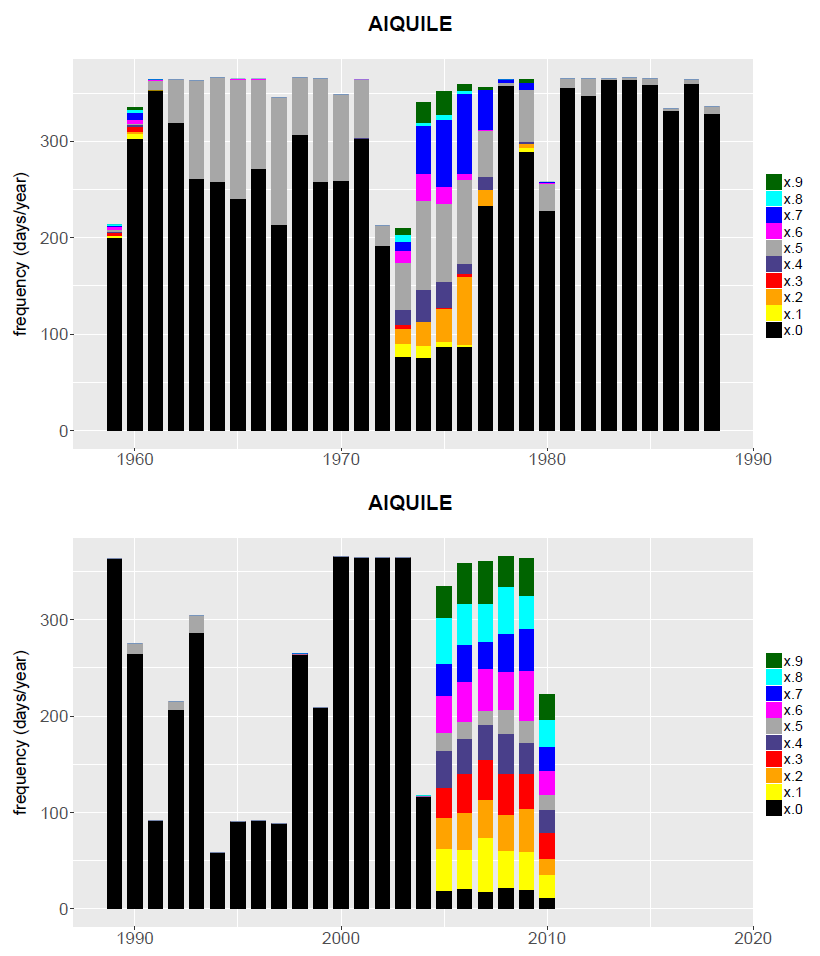
### Point plots



Missing values

Each data point indicates a daily observation

### Decimal plots

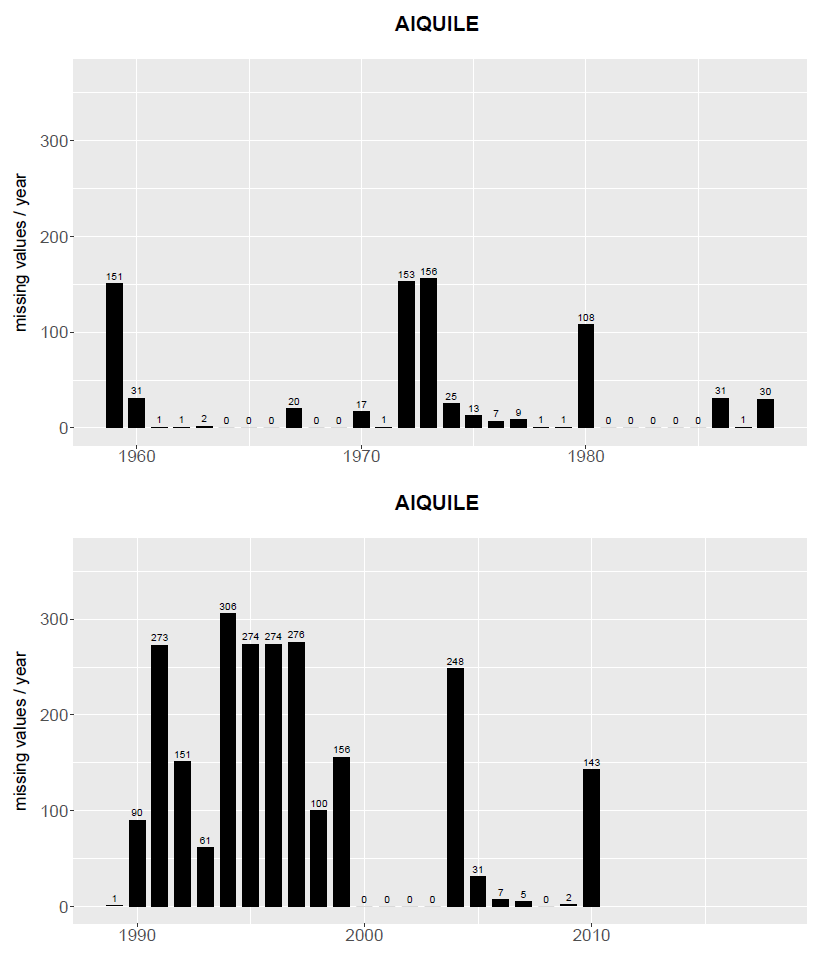


Colours indicate the decimal numbers

Count of the daily occurrence of decimal numbers per year

The occurrence and ratio of decimal numbers show the reported measurement precision, and may indicate the original measurement unit and erroneous rounding practices

### Missing values plots



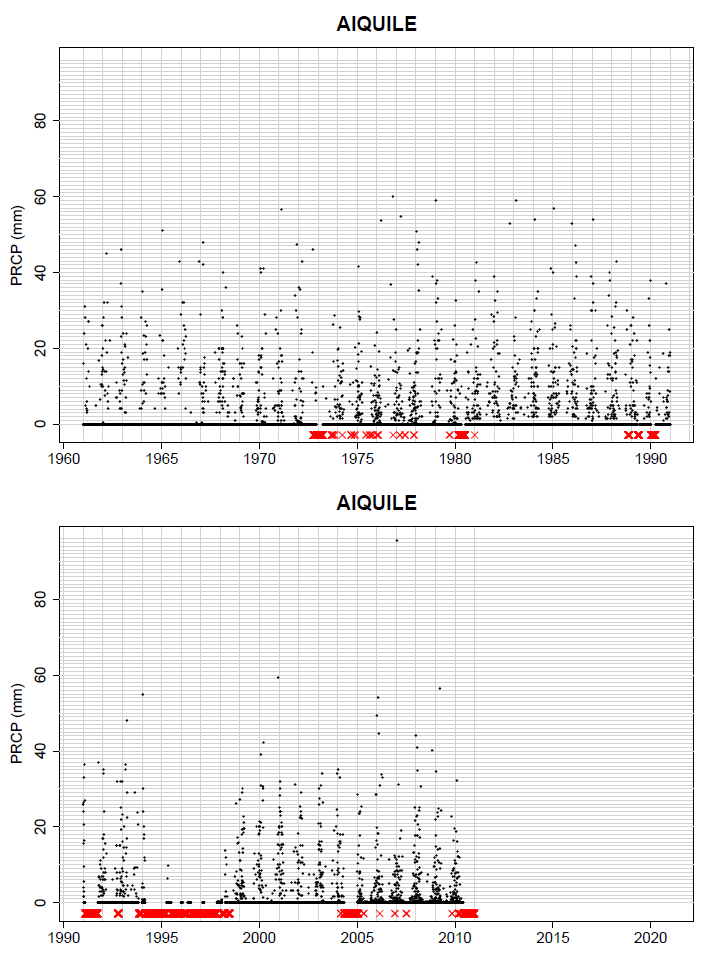
No bars or numbers indicate that the time period is out of the temporal range of the station record

Exact number of missing values for each year of the station record

Number of missing values per year

## Precipitation

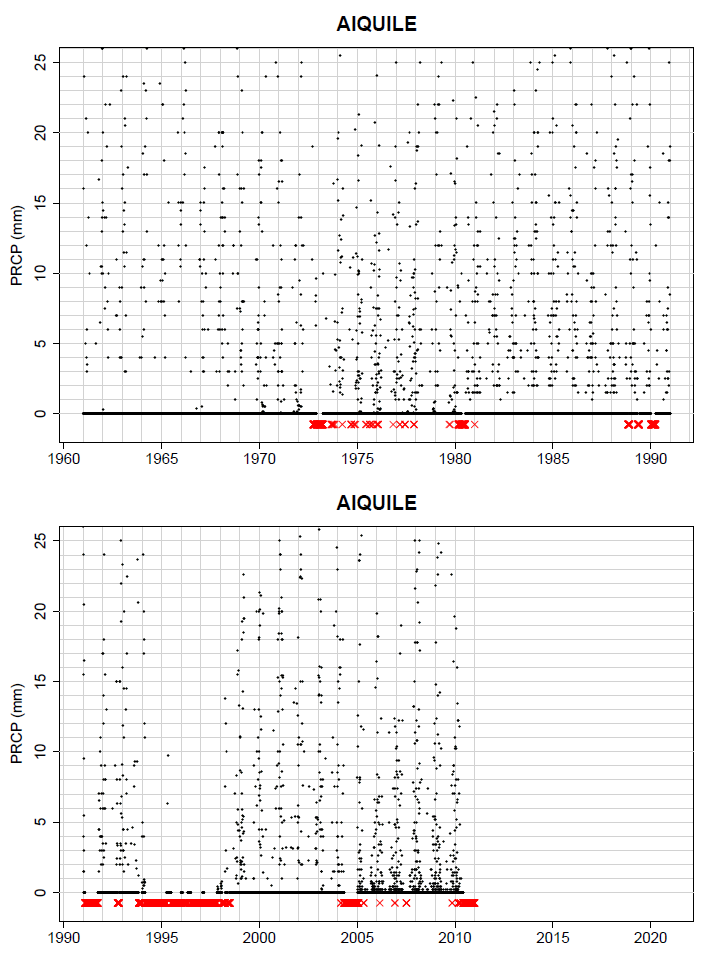
### Point plots



Each data point indicates a daily observation

Missing values

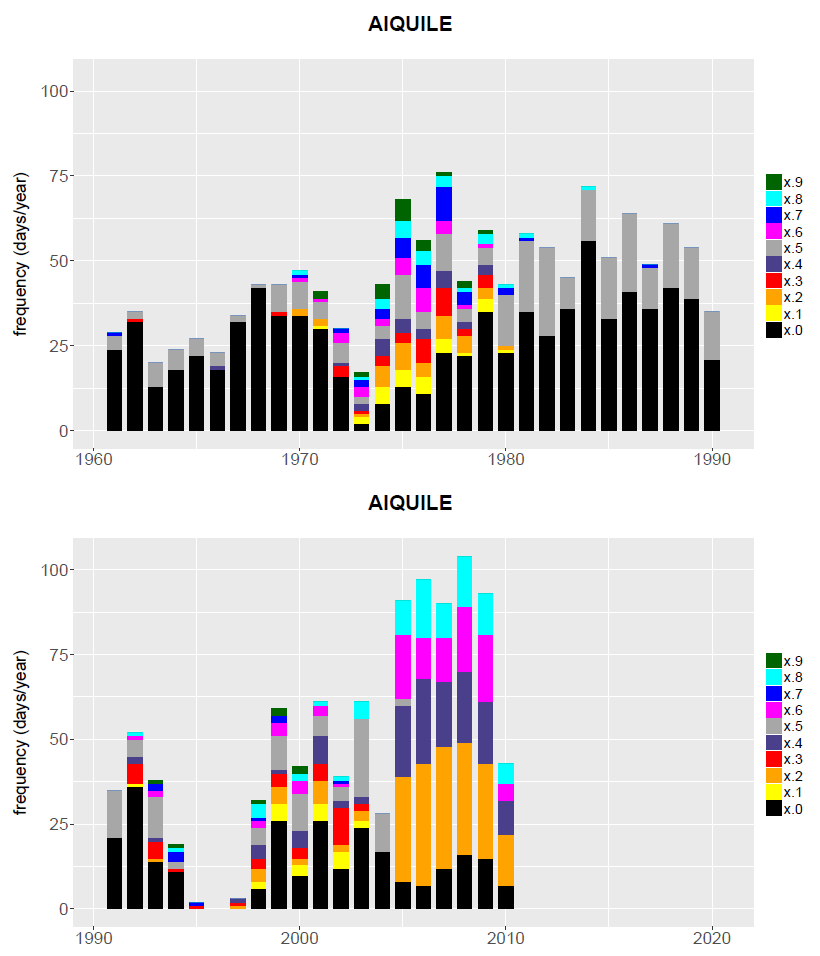
### Point plots threshold



Setting a threshold allows the detection of data quality issues occurring when there is little PRCP such as low PRCP gaps

The same as for the Point plots, but only observations up to a threshold defined by the user are shown (in this example 25 mm)

### Decimal plots

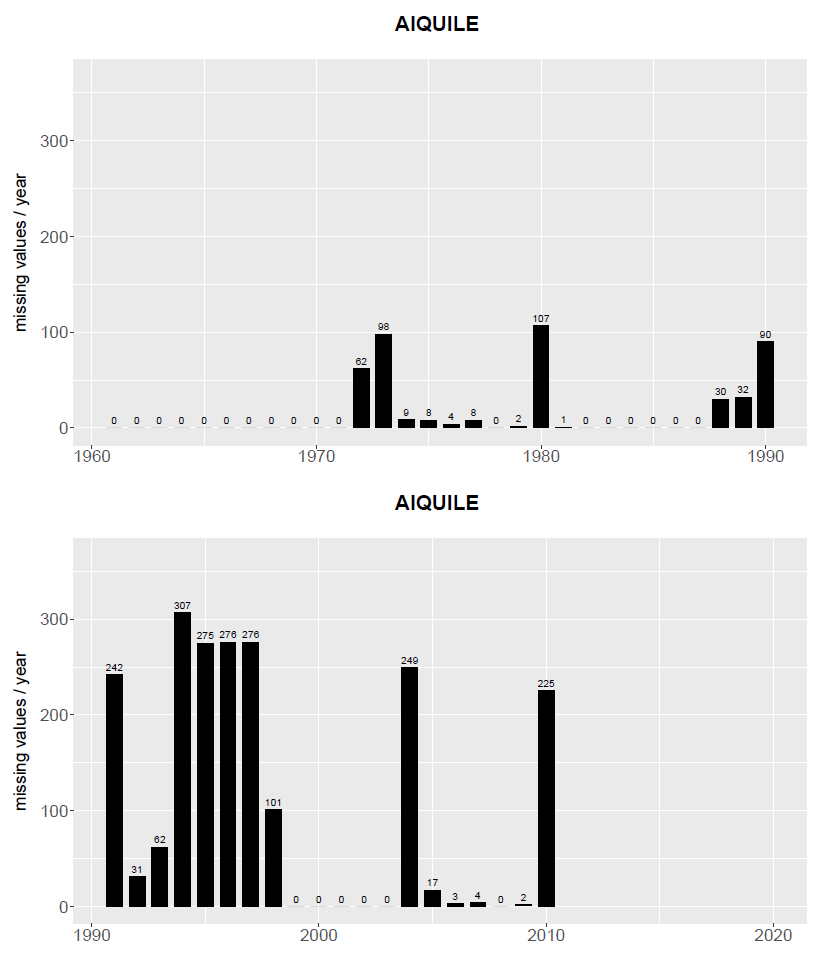


Colours indicate the decimal numbers

Count of the daily occurrence of decimal numbers per year

The occurrence and ratio of decimal numbers show the reported measurement precision, and may indicate the original measurement unit and erroneous rounding practices

### Missing values plots

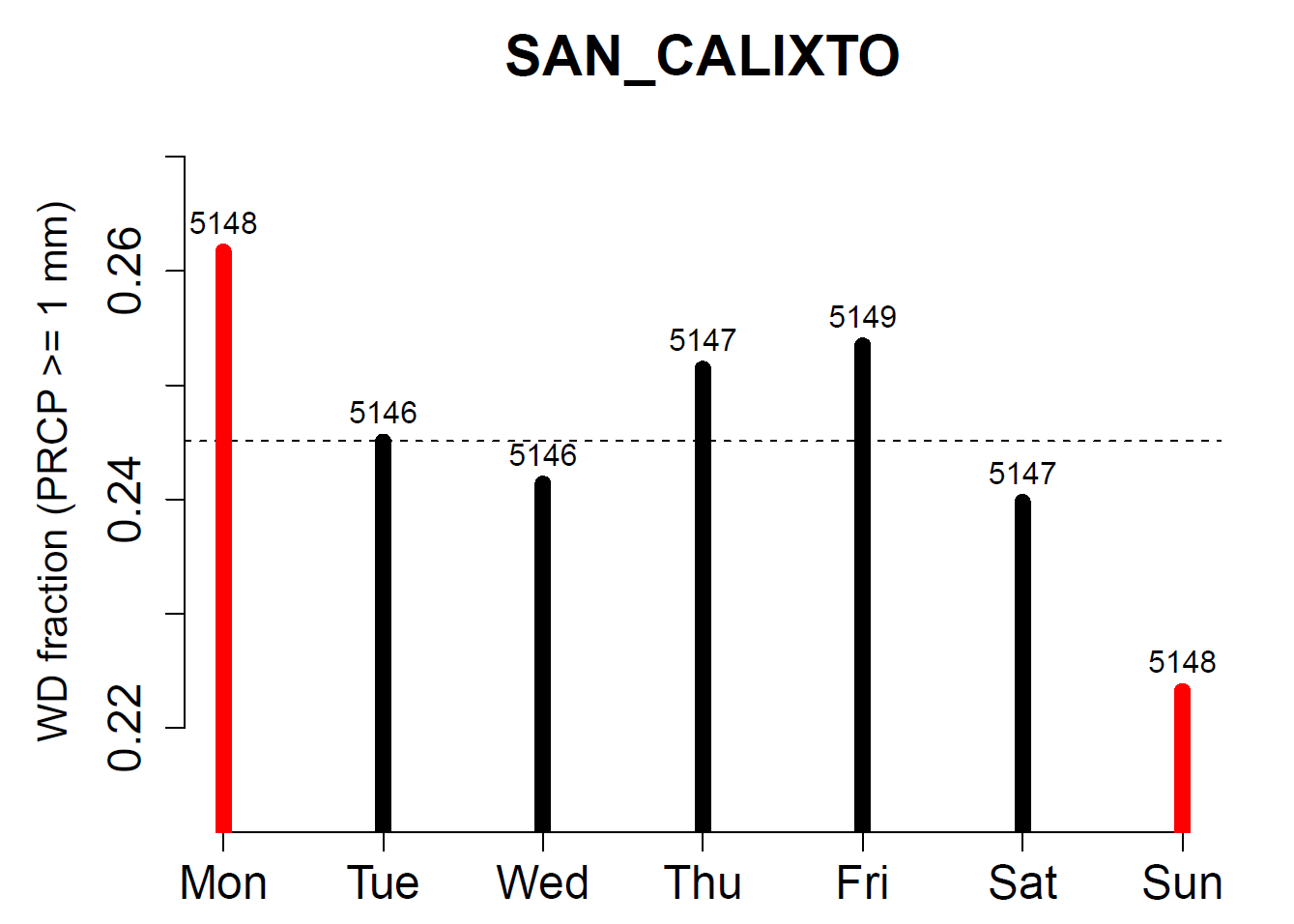


Exact number of missing values for each year of the station record

No bars or numbers indicate that the time period is out of the temporal range of the station record

Number of missing values per year

### Weekly PRCP cycles



Black: WD fraction does not differ significantly from the expected value (binominal test)

Red: WD fraction differs significantly from the expected value (binominal test)

Average WD fraction on all days of the station record

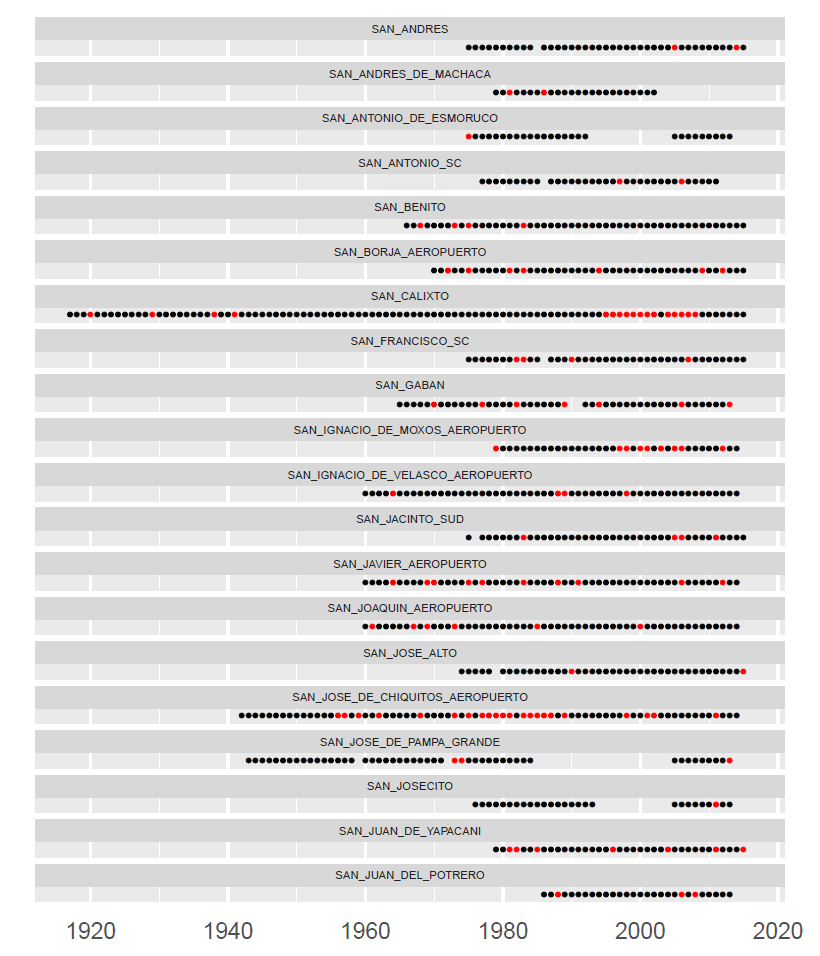
Total number of observations on each day of the week (here Friday)

Fraction of WD of each day of the week (here Tuesday)

Fraction of Wet Days (WD) in the station record (PRCP ≥1 mm)

Days of the week

### Weekly PRCP cycles on annual time scale



Consecutive detections are a clear indication for significant weekly PRCP cycles (years without detections in between are most likely affected as well)

Single detections do not necessarily mean the occurrence of a weekly PRCP cycle

Black: annual Wet Day fraction does not differ significantly from the expected value (binominal test)

Red: annual Wet Day fraction differs significantly from the expected value (binominal test)

Station names