Universal Quality Control (UQC) – software for the automatic quality control of hourly observational data of surface air temperature, air humidity, wind speed and wind direction

User guide and scientific description

Table of content

Centre for Climate Change
University Rovira i Virgili
Tortosa 2014

1. Introduction

This software performs various kinds of quality control, and makes error reports in result files, about the found errors. The software can be used only for the quality control of hourly observations of the climatic variables specified in the title. The input data must be carefully prepared before running UQC. After the short introduction, this guide will present the meaning of abbreviations (Sect. 2); the scientific description of the steps of UQC (Sect. 3); the rules of input data preparation (Sect. 4); some peculiarities related to the runs of UQC (Sect. 5); and the explanation of the content of the result files of UQC (Sect 6).

2. Denotations

TT = temperature
RH = relative humidity
VV = wind speed
DD = wind direction
TD = dew point calculated by UQC
TD* = dew point in the input data field
Tn = daily temperature minimum
Tx = daily temperature maximum
NT = normalized temperature
PDF = probability distribution function

3. Scientific description

As a general rule, the program does not stop for the detected errors, either makes any correction in the data. There are two exceptions from these rules: a) The program stops when date error is detected, since the identification of the date is absolutely necessary for the other steps of UQC; b) If absolute outliers (physically impossible values) are detected, these values are replaced with the missing data code for the duration of the running. However, corrected output dataset is not generated. As a consequence, after the manual correction of the detected errors (once they have been accepted by the user), the repeated run of UQC with corrected input is recommended.

Air humidity is controlled based on the observed RH data, but in some steps TD data derived from RH are checked by the program (when both TT and RH are available). In calculating TD, vapour pressure data are derived with the Magnus formula, and then the auxiliary file “td.dat” including the connection between vapour pressure and TD (also based on the Magnus formula) is used. Input data sometimes also contains TD data (TD*), but as the principal purpose of UQC is the check of observational data, TD* is not used except in one examination (at step 13).

Below the specific kinds of quality controls performed by the software are detailed. Many of the parameters of UQC can be manually changed with overwriting the default parameters included in auxiliary file “UQCparam.txt”.

1. Date order check

In the input file, data are expected to be ordered in a way that each line of the file corresponds to one calendar day. The entry of each line shows the year, month and day. If this entry indicates an erroneous date order, the program will stop with an error message indicating the expected date.

2. Search of repetitions of streaks of data for TT, RH, VV, DD

First 28-day section of time series for a specific month and specific hour is taken, and if it contains at least 10 observed values, the search of possible repetition of that streak is performed in all the other time series of the same climatic variable. Always the sections of the first 28 days of months are examined only. This examination aims to explore repeated digitization of entire months with erroneous dates or erroneous station code.

3. Exclusion of physically impossible outliers for TT, RH, VV, DD

Values out of the worldwide absolute extremes (e.g. TT out of (–91°C, +58°C) are obviously physical outliers. The purpose of this step is to remove very bad values affecting the basic statistical characteristics of the sample examined. Users can change the parameterization e.g. with considering stricter limits referring to specific geographical regions, but note that a must fine filtering of outliers is also performed (step 8).

4. Flat lines with at least 10 identical values for TT, RH, VV, DD

The sequence of the hourly values of the same hour is examined. Flat lines indicate the lack of accurately performed observations. The default minimum duration of flat lines is 10 days, the parameter value can be changed by the user.

5. Monthly mean of absolute increments between consecutive values for TT, RH, VV, DD

Climatic normals of mean increments between consecutive values are calculated for the time series of specific months and hours, when at least 50 pairs of temporally adjacent observed values are available. Mean monthly increments below 40% or above 250% of the climatic normal increment indicate permanent observational error, likely instrument error. The threshold ratios can be changed by the user.

This examination is performed in two ways: Once for pairing each hourly value with the observed value of the same hour of the previous day, and secondly, with pairing the values of adjacent observation terms of the same day. In each case, minimum 50 observed value-pairs are needed for deriving the climatic normal, and in the controlled months, the number of value pairs must reach the 70% of the number of days within the calendar month examined.

For DD, the difference of values is defined as the shortest arc in the circle of wind rose. (Thus, for instance, the difference between 340° and 20° is 40°.)

6. Check for frequently occurring values for TT

PDF is generated for specific months and hours when the time series contain at least 200 observed values. Seven equidistant intervals are constructed between the threshold values of PDF=0.05 and PDF=0.95. Each observed TT is put into a cluster according to the interval to which it belongs. If one cluster includes more than 50% of the observed values, it indicates erroneous accumulation of the same value.

7. Secondary peak in the frequency distribution of data for TT, TD, VV

Observed values are examined in sections of PDF for which the climatologically expected distribution is monotonous change of frequency of occurrence with increasing observed values. Climatic normal of PDF is calculated for time series of specified months and hours with at least 200 observed values. Histograms based on equidistant intervals of observed values are constructed. For TT and TD, the section between the absolute minimum and PDF=0.25 (between PDF=0.75 and the absolute maximum) is divided into three intervals and if the histogram does not show monotonous increase (decrease) of frequency with increasing observed values, it indicates the erroneous irregularity of the PDF. In examining VV, the section between the median and the absolute maximum is divided into five equidistant intervals.

8. Climatic outliers for TT, TD, VV

Climatic normals of PDF are constructed for specified terms of observation with at least 200 observed values. Large negative anomalies relative to the PDF=0.05 and large positive anomalies relative to the PDF=0.95 are considered to be climatic outliers.

This examination is performed in two ways: a) With creating monthly PDF (for TT and TD only); b) Without seasonal division of data (for TT, TD and VV). The threshold anomalies are additive constant for TT and TD and a multiplicative constant for VV. The threshold anomalies can be changed by the user.

9. Logical check for TT

If Tmin and Tmax data are available, Tn ≤ TT ≤ Tx of a given day are logically expected relations. The examination is performed with the following specifications:

i) When an expected relation is reversed, but the numerical difference is less than 1.0°C, the bias is tolerated and is not included in the report.

ii) When the term of TT observation is between 07h and 24h, Tn ≤ TT is expected either for the Tn of the same day or for the Tn of the following day.

iii) When the term of TT observation is between 01h and 09h, TT ≤ Tx is expected either for the Tx of the same day or for the Tx of the previous day.

iii) When the term of TT observation is between 18h and 24h, TT ≤ Tx is expected either for the Tx of the same day or for the Tx of the following day.

10. Too big sudden change for TT

Climatic normal values of specific months and hours are calculated when at least 50 observed values are available. Normalized temperature values (NT) are derived with the removal of the climatic normal. The time distance between the compared values (denoted with index 0 and index 1) cannot be longer than 12 hours. Sudden big changes are reported as errors when $NT_1 > NT_0 + 18°C$ or $NT_1 < NT_0 – 18°C$. The threshold parameter value can be changed by the user.

11. Too sharp spike for TT

NT values are derived in the same way and with the same conditions as in point 10. Three subsequent NT are considered (denoted with index 0, 1, 2) within a period of no longer than 16 hours. Sharp spikes are reported as errors when

a) $NT_1 > NT_0 + 12°C$ and $NT_1 > NT_2 + 6°C$ or

b) $NT_1 < NT_0 – 12°C$ and $NT_1 < NT_2 – 6°C$

The parameters can be changed by the user.

12. Irregular temporal evolution of climatic variables, joint examination of TT, RH, VV

The time distance between two adjacent observations (denoted with index 0 and index 1) cannot be longer than 12 hours.

i) (warming) TT and RH data must be available both for time 0 and time 1.

If $TT_1 > TT_0 + 10°C$ and $RH_1 > 0.92*RH_0$, then some of the observed values is likely erroneous.

ii) (cooling) TT, RH and VV must be available both for time 0 and time 1.

If $TT_1 < TT_0 – 8°C$ and $100\% – RH_1 > 0.8*(100\% – RH_0)$ and $VV_1 < VV_0 + 3ms^{-1}$ and $VV_1 < 7$ m/sec, then some of the observed values is likely erroneous.

Most parameter values can be changed by the user, but the 7 m/sec threshold for VV_1 can not.

13. Check of TD*

For each observation with available TT, RH and TD* and when either TD > –20°C or TD* > – 20ºC, the program controls the differences between TD and TD*. The tolerated maximal difference is 2.0°C for values between –20°C and –10°C of TD / TD*; and 1.0°C when either TD > –10ºC or TD* > –10ºC. The maximal tolerated difference can be changed by the user. On the other hand, TD* is not allowed to be higher than TT, and it is also checked.

4. Preparation of input data

Users must follow strictly the rules presented here, because any deviation from them likely causes an early stop of the program without the fulfilment of quality control, or even when the run ends seemingly normally, a part of the input data may have been left out from the quality control.

Input data must be separated into distinct files a) according to the observing sites, b) according to climatic variables. The input files are text files. Usually the values of one specified variable are included only in one input file, but there are two exceptions: If the input contains TD, its values must be included in the same files as RH data, while DD data must be presented in the files of VV data.

4.1. Kinds and names of input data files

Each file must hold a 14-character long file name. The file names include: a) a 5-character long network identifier (e.g. "Earth"); b) a variable identifier (TT, RH or VV); c) the serial number of station with 3 digits (stations can be ordered arbitrarily, e.g. if the network includes 4 stations, the serial number codes will be 001, 002, 003 and 004); d) fixed four characters at the extension: ".dat". For instance, the humidity data of the 4 stations of "Earth" network will be in the 4 files of
EarthRH001.dat
EarthRH002.dat
EarthRH003.dat
EarthRH004.dat

None of the variables is obligatory to be included, however, if a variable has observed data even only at one site of the network, that variable must be included in the prescribed form of input files at each site of the network. If there is no station data for some variable at a given site, but the inclusion of input file is obligatory, the file must be prepared in a way that it includes a prescribed heading, but no data.

VV data are mostly available together with DD data. If observed DD data is available in any site of the network and any observing term, all VV files and the data of all observing terms must be prepared in a way that DD and VV data of the same observing term are written consecutively, in this order. If one or both of DD and VV are missing, then missing data code (–99.9) must be included. The same structure must even be kept in files without observed data of one of these two variables. However, if there is no DD data in the entire network, VV data stand alone.

If TD* data are available they must be written just after the RH of the same observing hour. If TD* data are available only for some stations or for some observing terms, the structure of RH and TD* data must be kept in all files and for all observing terms (as for DD and VV, see the previous para. See also examples, Sect. 4.4).

4.2. Heading of input data files

Files with observed data include heading of 4 lines. Numbers in the heading must be included always in whole number format and when more than one characteristics are presented in a line, they must be separated with one or more space characters or with TAB.

The first line includes a station identifier. There is no restriction in connection with this identifier, as it is not used by the program. Note, however, that the file cannot start with the content that is expected in the second line only.

The second line shows the dates of the first day and last day with data in the file. It is specific for each file, i.e. there is no need to harmonize the starting and ending dates of files. The file may start and end with any days of the calendar year. Dates are specified in the form of:

dd1 mm1 yy1 dd2 mm2 yy2

where dd1 – day of starting date; mm1 – month of starting date; yy1 – year of starting date; dd2 – day of ending date; mm2 – month of ending date; yy2 – year of ending date.

The third line shows how many observing terms include observed data. For instance, if there are observed data for 06h, 12h, 15h and 18h in a file, "4" must be written into this line. This characteristic is file specific, i.e. the number of terms can be different for files, even for the files of the same station or for that of the same climatic variable.

The fourth line shows the hours of observing term, logically it must include the number of characteristics that is specified in the previous line. The number of observing terms and their order is file specific, there is no restriction around this order. Note however, that one examination of UQC (shown in point 6 of Sect. 3) is effective primarily when the order of the observing terms corresponds to the natural order of the observations in the observing site. There is no effect of data order on the other examinations.

Files with entries sometimes must be prepared when there is no observed data available. In this case the heading consists of two lines only. The first line is the same as in files with observed data, but instead of a true starting day, code –99 must be written on the place of dd1 of the second line. This code reports to the program that the file is empty. However, it is necessary to write 5 other whole numbers into this line, because the program searches a line with 6 whole number characteristics when it reads the line of starting and ending dates. Thus the most convenient infilling of this line in empty files is: –99 0 0 0 0 0. Once the program has read the code –99 in the place of dd1, it stops reading that file.

4.3. Data table of input data files

After the heading, the file follows with the table of data. Note that living blank lines is never allowed in input data files. Each line of the data table is for one calendar day. The uppermost line must correspond with the starting date and the last line with the ending date indicated in the heading. The data table includes Arabic numbers only, and the numbers of the same line are separated with one more space characters or with TAB.

Each line starts with presenting the year, month and day of the observations. After the date characteristics the observed data follow, in the order of terms defined by line 4 of the heading. Date characteristics must be shown in whole number format, while the format of the observed data is free, with the only restriction that decimals must be separated by dot, the use of comma is not allowed. Missing data must be marked with –99.9. The physical units of data are fixed: they are °C for TT and TD, % for RH, m/s for VV and degree (°) for DD.

4.4. Example input data files

a) Example of TT input data file

```
Earth Example 000001
16  7  1893  25  2  1972
7
                    07         14        21         06        12         18         24
1893  7  16    17.0       33.0    –99.9    –99.9   –99.9   –99.9    –99.9
1893  7  17    19.5       36.0      26.5    –99.9   –99.9   –99.9    –99.9
1893  7  18    20.0       32.0      21.5    –99.9   –99.9   –99.9    –99.9
………………………………………………………………………..
 ………………………………………………………………………..
1972   2  24   –99.9   –99.9     –99.9     –7.2       0.0      –3.4        –9.7
1972   2  25   –99.9   –99.9     –99.9   –12.3       0.6      –0.8        –1.3
```

In this example the observation terms are 07h, 14h and 21h in the early period and 06h, 12h, 18h and 24h later. Thus the total number of observing terms is 7, and the data table contains 7 columns out of the date characteristics. “17.0”, “33.0”, etc. could be presented also in whole number format (the program accepts that), but the recommended format of TT, TD and VV data is real number with one decimal, because the consequent use of this format reduces errors (e.g. 5 instead of 0.5, 20 instead of 2.0, etc.).

b) Example of VV input data file with DD data

```
Earth Example Townhall District 3524
30  4  2002  6  9  2013
3
7  13  19
2002  4  30     340     2.6     –99.9   4.2      –99.9   1.0
2002  5   1     0       0.0     –99.9   3.4      –99.9  10.1
2002  5   2     270     5.7     –99.9   7.4      –99.9   3.3
…………………………………………………………………………..
 …………………………………………………………………………..
2013  9   5     020     1.8    –99.9    2.1      –99.9   0.0
2013  9   6     160     3.0    –99.9    2.4      –99.9  –99.9
```

In this example the horas of observation (7, 13, 19) are written at the beginning of the line and not above the data columns, but it has no impact on the performance of the program. It seems that wind direction was observed only in the morning, but the format of DD – VV pairs must be kept for each observing term, once it has been applied for one term, moreover this format must be kept for all stations of the network. Wind direction “020” could be simply “20”, and wind speed “0.0” simply “0”, (the program accept all formats) but we recommend to apply unified format for reducing the occurrences of errors.

4.5. Parameter setting

File “UQCparam.txt” contains several parameters of the running. This file has two parts. In the first part, the questions always need answers specific for the individual running. In the second part, the default parameter setting of the quality control is included and although each of them can be changed, runs can be done without any change in them.

In the first part of “UQCparam.txt”, first the name of the network must be specified, then the number of stations within network must be given (it can be between 1 and 100), then the individual climatic variables with available data for the UQC procedure must be specified one by one. Finally, the earliest and the latest years with observed data occurring in the network must be specified. If the dates introduced here do not cover the entire period with data, it leads to serious errors, while if the dates cover an unnecessarily long period, it does not cause computational error, only the time consuming and the place occupation of the program would increase unnecessarily. However, the absolute limit of 300 years is not allowed to be exceeded by the length of the period flagged for the examination by UQC.

5. Running Universal Quality Control

The operation of UQC is fully automatic, although its results are recommended to be checked manually.

For running UQC, the executive program “UQC.exe”, the auxiliary file “td.dat” and “UQCparam.txt” with an adequate parameter setting must be put into the working directory together with the carefully prepared input data files. Then with a simple click on UQC.exe, the running will start. During the running, the program creates temporary files. The place occupation of such files and the time consumption of the program largely depend on the size of the input dataset. The usual size of place occupation is between 50MB and 500MB, and the time consumption is usually less than 10 minutes. However, in case of very big datasets (large number of stations and long periods with observational data) the place occupation might be larger than 1GB and the time consumption may reach several hours. Note that the control of streak repetition (point 2 of Sect 3) is the only examination in which data of different observing sites are examined together. Thus creating networks of large number of series is reasoned only when the occurrence of this error type (change of stations in the digitization) has realistic possibility.

For any case, the program writes information into the file titled “Phase of calculation” from time to time during its operation. With controlling the content of this file, one can check the operation of the program (see if it runs or has been frozen).

6. Kinds and content of results files

The result files are separated according to the principal kinds of errors, but not according to the controlled variable. In the reports, variables are always indicated with their two-letter abbreviations.

The program can generate ten result files. They are:
Repetiti_streaks.dat
Outliers.dat
Flat_lines.dat
Irregular_monthlymeanchanges.dat
Irregular_PDF.dat
TnTx_inconsistency.dat
Big_jumps.dat
Sharp_spikes.dat
Intervar_inconsistency.dat
TD_inconsistency.dat

The first four files of the list are always generated, while the other output items are conditional, their existence depends on if the input includes the target climatic variable(s). If one or more of the generated files are empty, it means that the referred type of quality control was done, but no incident of the indicated error type has been detected.

The result files include the list of the detected errors with some characteristics for each occurrence. Each line includes one detected error, in each of the result files. The name of the network and the date of the detected error are always shown. As the name of the network is indicated, the results are allowed to be gathered in the same files from different networks. In one specific run only one network can be quality controlled, but UQC never overwrites result files, but if it finds them in the directory of the running, the new results will be added to the existing lists.

The indicated characteristics of the detected errors in result files (from left to right) are as follows:

Repetiti_streaks.dat: network, climatic variable; first occurrence of streak with: serial number of station, year, month, hour; repeated occurrence of streak with: serial number of station, year, month, hour; number of observed values in the streak

Outliers.dat: name of the input data file (including the names of network and climatic variable), year month, day, hour, indication of absolute or climatic outlier with "abs" or "clm", detected outlier value

Flat_lines.dat: length of the period, name of the input file, date of the end of the streak (with year, month, day, hour)

Irregular_monthlymeanchanges.dat: name of input data file, year, month, hour, difference between the timings of the compared observed values (hours), detected mean change, climatic mean change

Irregular_PDF.dat: name of the input data file, month, hour; kind of the detected error with "peak" when more than 50% of the values are accumulated in one cluster, "Lq." for the irregularities of the lowest quartile, and "Uq." For those of the upper quartile (for VV the denotation is always "Uq."); number of occurrences in each cluster, thresholds of the intervals defining the clusters

TnTx_inconsistency.dat: name of the input data file, year, month, day, hour

Big_jumps.dat: name of input data file, year, month, day, hour at the end of the change, duration of change (hours), size of change

Sharp_spikes.dat: name of input data file, year, month, day, hour of spike, duration of the detected change (hours, from the observation before the spike until the observation after the spike); detected values before the spike, at the spike and after the spike

Intervar_inconsistency.dat: network, serial number of station, year, month, day, hour at the end of the changes, duration of the detected changes; a) for warming events: TT and RH before the changes, TT and RH after the changes; for cooling: TT, RH and VV before the changes, TT, RH and VV after the changes

TD_inconsistency.dat: name of the input data file, year, month, day, hour, TT, RH, TD, TD*.