Indecis
Sectorial Climate Services

Integrated approach for the **d**evelopment across **E**urope of user oriented **c**limate **i**ndicators for GFCS high-priority **s**ectors: Agriculture, disaster risk reduction, energy, health, water and tourism

Work Package 3

Deliverable 3.1.c

INDECIS Quality Control Software and Manual: MetQc

P. Štěpánek [1]

[1] *Global Change Research Institute of the Czech Academy of Sciences (GCRI), Czech Republic*

This report arises from the Project INDECIS which is part of ERA4CS, an ERA-NET initiated by JPI Climate, and funded by FORMAS (SE), DLR (DE), BMWFW (AT), IFD (DK), MINECO (ES), ANR (FR), with co-funding by the European Union's Horizon 2020 research and innovation programme

TABLE OF CONTENTS

1. Introduction

Prior to any data analysis, data quality control and homogenization have to be undertaken to get rid of erroneous values in time series. Considering quality control, we face the lack of a generally accepted methodology (contrary to homogenization). But without treating outliers properly, homogenization and subsequent analysis may render misleading results. Therefore, we have devoted considerable time to the methodology of detecting outliers, something that could, moreover, be automated to process large datasets of daily (sub-daily) values, and applicable to large regions (like the European dataset).

Comparisons of various approaches to quality control methods were made, tested on a benchmark dataset into which realistic errors were introduced; based on the testing, the final software was prepared. The quality check module is developed as open-source software in R (https://www.r-project.org/), available to the science community. The plan is for the software to be updated and improved regularly.

The quality control must be able to capture the most typical errors that occur in climatic data, but also detect specific cases. The current software is based on years of experience in testing various datasets. Within INDECIS activity, the software was tested on the chosen datasets from various parts of Europe (South Sweden and Slovenia). What is novel in this work is that, on the European level, all basic meteorological elements have been taken into account: cloud cover, wind direction, wind speed, wind gust, relative humidity, sea level pressure, precipitation sum, snow depth, sunshine duration, minimum temperature and maximum temperature.

2. Software description

Original QC functions converted from ProClimDB (http://www.climahom.eu/software-solution/proclimdb) into R software (called MetQC) form the basis of the presented software in order to be better usable by various users and applicable stand-alone, i.e. not necessarily as part of ProClimDB. The QC scripts are written in the R language and are ready for processing among others e.g. ECA&D dataset. A few functions were newly programmed to further help INDECIS users more easily run it.

2.1 Functionality description and motivation behind it

The main script (*MetQC_frame.R*) uses three main steps of the data processing, and is divided into functions accordingly:

1. getting information about data
2. finding neighbours for subsequent data quality control
3. data quality control itself

These steps correspond to the functions, which are wrappers for launching the *launch_QC.r* script, where the calculation itself is performed. These functions (wrappers) are given in the beginning of the *MetQC_frame.R* script (further described below). Please note that it is possible to run the code from the body of the main *MetQC_frame.R* script (at the very bottom). Or it is possible to straight apply functions,

in two ways: functions where all the possible parameters are specified (*Run__Get_info_file, Run__Find_Neighbors, Run__MetQC*), or through theirs wrappers – that will use default settings for each of the meteorological elements (*run_Get_Info_1_1, run_Find_Ngbs_7_2, run_QC_7_4*) – that are simpler to run (by the way, the latter functions are used when applying parallel processing).

Task to be accomplished	Core functions (possibility of changing parameters)	Functions for easy run (default settings)
1. getting information about data	Run__Get_info_file()	run_Get_Info_1_1()
2. finding neighbours for QC	Run__Find_Neighbors()	run_Find_Ngbs_7_2()
3. data quality control itself	Run__MetQC()	run_QC_7_4()

Figure 1. Scheme of launching functions in launch_QC.r script

2.1.1. Getting information about data

The *Run__Get_info_file* function in the main *MetQC_frame.R* script (its wrapper function is *run_Get_Info_1_1*).

The aim of the function is to prepare a list of stations together with geography (coordinates) to check the data file for missing values (including information e.g. about the maximum length of the period with missing values) and other formal problems. It also gives information about the nearest station (distance) and average distance to all other stations within the given region. Such a list is used in the next step for selection of the proper neighbourhood of each of the candidate stations.

Id	Id_orig	Region	Latitude	Longitude	Begin	End	Length	Cnt	Miss_cnt	Miss_max	Prc_miss	Beg_year	End_year	Len_y	Name	Id_near	Dist_min	Dist_avg	Dist_max	Altitude
P0034	P0034	SE	55.9869444444444	12.9136111111111	1.1.1950	31.12.2005	20454	19173	1281	616	6.3	1950	2005	56	gridpoint	P0037	36.5833	348.5469	690.2824	-999
P0037	P0037	SE	56.0083333333333	13.5000000000000	1.1.1950	31.12.2005	20454	15008	5446	3701	26.6	1950	2005	56	gridpoint	P0034	36.5833	336.9189	677.0738	-999
P0053	P0053	SE	56.1477777777778	14.4688888888889	1.1.1950	31.12.2005	20454	18590	1864	371	9.1	1950	2005	56	gridpoint	P0066	12.2596	316.9916	647.1417	-999
P0066	P0066	SE	56.2577777777778	14.4594444444444	1.1.1950	31.12.2005	20454	18200	2254	670	11.0	1950	2005	56	gridpoint	P0053	12.2596	306.3827	635.3122	-999
P0074	P0074	SE	56.3161111111111	12.8727777777778	1.1.1950	31.12.2005	20454	16307	4147	1841	20.3	1950	2005	56	gridpoint	P0034	36.7315	319.2881	656.5080	-999
P0081	P0081	SE	56.3622222222222	14.2525000000000	1.1.1950	31.12.2005	20454	17498	2956	2357	14.5	1950	2005	56	gridpoint	P0099	12.2302	296.9934	626.9164	-999
P0089	P0089	SE	56.3972222222222	15.8330555555556	1.1.1950	31.12.2005	20454	19840	614	421	3.0	1950	2005	56	gridpoint	P0105	27.2928	308.5471	617.3926	-999
P0099	P0099	SE	56.4719444444444	14.2425000000000	1.1.1950	31.12.2005	20454	14943	5511	1881	26.9	1950	2005	56	gridpoint	P0081	12.2302	287.3331	615.2123	-999
P0104	P0104	SE	56.4963888888889	15.2325000000000	1.1.1950	31.12.2005	20454	14837	5617	4032	27.5	1950	2005	56	gridpoint	P0105	12.1775	289.9906	601.1280	-999
P0105	P0105	SE	56.5002777777778	15.4305555555556	1.1.1950	31.12.2005	20454	15625	4829	2257	23.6	1950	2005	56	gridpoint	P0104	12.1775	292.3759	601.9886	-999
P0136	P0136	SE	56.6972222222222	14.4211111111111	1.1.1950	31.12.2005	20454	15148	5306	3361	25.9	1950	2005	56	gridpoint	P0099	27.3651	268.4554	588.2293	-999

Figure 2 Created list of stations, together with information about length of measurement, missing values, coordinates, etc.

2.1.2. Finding neighbours

The *Run__Find_Neighbours* Function in the main *MetQC_frame.R* script (its wrapper function is *run_Find_Ngbs_7_2*). This function serves for finding proper neighbourhood for each of the candidate stations.

Correct selection of reference stations for testing of data quality is crucial. Neighbouring stations may be selected using either the highest correlations or smallest distances. In the case of temperature, the set of neighbours may differ depending on which approach is taken. For precipitation, the sets of selected neighbours coincide. Correlation coefficients may be applied either to raw series or to series of first differences (see e.g. Peterson, 1998). It is also possible to restrict the neighbour selection to similar

altitudes. It is wise to combine the criteria based on distance and correlation, especially for a complex terrain region. In the case of a mountainous station, a valley station as a neighbouring station has completely different climatic conditions. Such restrictive neighbour selection is important: when performing the data quality control, neighbouring stations are standardized to the base station altitude. With vastly differing climates between two stations, this straightforward linear adjustment is unable to adjust for this difference. Since some test statistics rely on standardized series, the inability to adjust for altitude difference may be confused with suspect data.

In the case of a dense network, the selection may be based uniquely on the nearest stations, since the best correlated stations in the dense networks are, at the same time, those closest (Štěpánek et al., 2011, 2013). Obviously, the more neighbouring stations, the more robust the estimate of the reference is. From our previous experience, we can recommend the following settings and limits for neighbour stations selection: choose at least six of the best correlated (or nearest) neighbours. The motivation to have the threshold at six stations is that the lower and upper quartiles then correspond to the second and fifth station. The software default setting includes neighbours selection in the way that the first 75 percent of the stations found, ordered according to correlations or distances, is taken as it is. The software subsequently compares base station altitude with the neighbour ones, and if the coefficient of interquartile range is above 1.5, it searches for another neighbour (the reason is to have a better sample of altitudes for the linear regression when standardizing values for the same altitude). Note that the correlations should always be statistically significant at the p=0.95 level, but lowering the threshold to p=0.90 or perhaps even less is a sensible choice when short segments of data are used.

Suitable neighbour stations are found by combining weighted criteria on distances, altitude differences of candidate and neighbour stations, and length of common periods (also taking into account the number of missing values). The software chooses the neighbour stations automatically, based on the criteria (function parameters) set, but the user is able to manually manage (edit) the selection made automatically by the software and use the new selection for the following statistical tests (see Figure 3).

Id_1	Id_2	Region	Begin	End	Length	Correl	Distance	Dist_order	Latitude	Longitude	Prc_miss_1	Prc_miss_2	Pen_miss2	Pen_length	Pen_corr	Pen_dist	Pen_alt	K_select	Del_	Mfd_	Ngb_order
P0034		SE	1.1.1950	31.12.2005					55.987	12.914	6.300								F	F	
	P0053	SE	1.1.1950	31.12.2005	20454	0.9255	98.292	4	55.987	12.914	6.300	9.100	0.0910	0	0.106	0.250		1.121	F	F	1
	P0177	SE	1.1.1950	31.12.2005	20454	0.7972	156.320	13	55.987	12.914	6.300	17.300	0.1730	0	0.290	0.390		2.403	F	F	2
	P0104	SE	1.1.1950	31.12.2005	20454	0.7829	154.245	12	55.987	12.914	6.300	27.500	0.2750	0	0.310	0.390		2.605	T	F	
	P0099	SE	1.1.1950	31.12.2005	20454	0.7803	98.372	5	55.987	12.914	6.300	26.900	0.2690	0	0.314	0.250		2.339	F	F	3
	P0066	SE	1.1.1950	31.12.2005	20454	0.7792	100.550	7	55.987	12.914	6.300	11.000	0.1100	0	0.315	0.250		2.185	F	F	4
	P0173	SE	1.1.1950	31.12.2005	20454	0.7490	122.216	10	55.987	12.914	6.300	15.000	0.1500	0	0.359	0.310		2.565	F	F	5
	P0074	SE	1.1.1950	31.12.2005	20454	0.7351	36.731	2	55.987	12.914	6.300	20.300	0.2030	0	0.378	0.090		2.273	F	F	6
			. .	. .															F	F	
P0037		SE	1.1.1950	31.12.2005					56.008	13.500	26.600								F	F	
	P0034	SE	1.1.1950	31.12.2005	20454		36.583	1	56.008	13.500	26.600	6.300	0.0630	0		0.090		0.000	F	F	1
	P0074	SE	1.1.1950	31.12.2005	20454		51.824	2	56.008	13.500	26.600	20.300	0.2030	0		0.130		0.000	F	F	2
	P0081	SE	1.1.1950	31.12.2005	20454		61.036	3	56.008	13.500	26.600	14.500	0.1450	0		0.150		0.000	F	F	3
	P0053	SE	1.1.1950	31.12.2005	20454		62.162	4	56.008	13.500	26.600	9.100	0.0910	0		0.160		0.000	F	F	4
	P0066	SE	1.1.1950	31.12.2005	20454		65.679	5	56.008	13.500	26.600	11.000	0.1100	0		0.160		0.000	F	F	5
	P0099	SE	1.1.1950	31.12.2005	20454		69.091	6	56.008	13.500	26.600	26.900	0.2690	0		0.170		0.000	F	F	6
			. .	. .															F	F	
P0053		SE	1.1.1950	31.12.2005					56.148	14.469	9.100								F	F	
	P0136	SE	1.1.1950	31.12.2005	20454	0.9304	61.237	5	56.148	14.469	9.100	25.900	0.2590	0	0.099	0.150		1.054	F	F	1
	P0099	SE	1.1.1950	31.12.2005	20454	0.9265	38.701	3	56.148	14.469	9.100	26.900	0.2690	0	0.105	0.100		0.994	F	F	2
	P0105	SE	1.1.1950	31.12.2005	20454	0.9020	71.159	7	56.148	14.469	9.100	23.600	0.2360	0	0.140	0.180		1.296	F	F	3
	P0152	SE	1.1.1950	31.12.2005	20454	0.8778	82.066	8	56.148	14.469	9.100	28.200	0.2820	0	0.175	0.210		1.577	F	F	4
	P0089	SE	1.1.1950	31.12.2005	20454	0.8615	88.775	9	56.148	14.469	9.100	3.000	0.0300	0	0.198	0.220		1.460	F	F	5
	P0173	SE	1.1.1950	31.12.2005	20454	0.8464	89.108	10	56.148	14.469	9.100	15.000	0.1500	0	0.219	0.220		1.685	F	F	6

Figure 3 Example of neighbours selection with weights penalizing increasing distance (Pen_dist column), altitude difference (Pen_alt column) and common data availability (Pen_length column), summed together in the K_select column.

2.1.3. Data quality control itself

The *Run__MetQC* function in the main *MetQC_frame.R* script (its wrapper function is *run_QC_7_4*). This function launches quality control based on the list of neighbours for each of the candidate stations.

Finding problems (errors) in the dataset is based on a combination of several statistical methods. In this way, it is possible to detect measurement errors correctly and automatically at the same time, permitting the disclosure of a larger amount of errors and a reduction in the proportion of false data.

The methods can be divided into three groups (Štěpánek *et al.*, 2011a, 2013):

(i) analysis of series of differences (ratios) between candidate and neighbouring stations (i.e. pair-wise comparisons). Probabilities of the differences (CDF) are estimated and evaluated.

(ii) estimation of the coefficient of interquartile range for base station value compared to all the neighbours, for a given time stamp;

(iii) comparison of tested values with "expected" (theoretical) values. This part of the testing reflects spatial information (compared to the time aspect of the previous two parts). The expected values are calculated from the neighbouring stations, standardized to the altitude of the station being tested. In this 'standardization' step, the observed value, such as temperature, of the neighbouring stations is adjusted to the elevation of the base station. The adjustment of observed value of the base and neighbour stations for altitude is estimated for each time (e.g. day) individually. In the case that the regression model is not able to describe the dependence well, the original values of the neighbouring station are used. The probability (CDF) of the base station difference to the median of standardized (to altitude) neighbours is calculated. From these standardized (to altitude) values, an expected value is estimated as a weighted mean. The weights are reciprocal values of distances, with a given power (IDW; $1/(Distance^{power})$). In the case of wind speed, the power of distance is two to reflect spatial variability of a given meteorological element; for precipitation, the power is set to three.

Various statistical tests, based both on time and spatial aspects, performed, are:

- Estimation of coefficient of interquartile range (Wilks, 1995) for base station value compared to all the neighbours, for a given time stamp. Default settings: 3.0

- Differences between neighbours and base stations are calculated. In the case of precipitation, the values may be transformed, e.g. by logarithms, but there are also other options of transformation available, the user may choose the one most appropriate. These differences (ratios) are standardized and their probabilities (CDF) are estimated. Default settings: the 75% of neighbours exceeds p=0.95.

- The first differences are calculated for the base and all the neighbour series and the coefficient of interquartile range and probability of differences (ratios) are evaluated in the same way as described above.

- Neighbour values are standardized to base station altitude, applying linear regression techniques in which the dependence on altitude of the values of the base and neighbour stations is estimated. In this 'standardization' step, the observed value, such as temperature, of neighbouring stations is adjusted to the elevation of the base station. The adjustment of the observed value of the base and neighbour stations for altitude is estimated for each time (e.g. day) individually. In the case that the regression model is not able to describe the dependence well, the original values of the neighbouring station are used. The probability (CDF) of the base station difference to the median of standardized (to altitude) neighbours. Default settings: the probability exceeds p=0.95.

- From these standardized (to altitude) values, an expected value is estimated as a weighted mean. The weights are reciprocal values of distances, with a given power. In the case of temperatures, the power of distance is one; for wind speed, two; for precipitation, three; to reflect the spatial variability of a given meteorological element. The difference between the original and the expected value is used as another criterion for attribution of an error.

- Checks of the same values in a row are applied as well. Two versions are used: the same values of a base station (i.e. repeating values), and the same values of differences between a base and a neighbour station.

The final list of errors is created by a combination (common detection) of the previously mentioned statistical tests, which reflects various aspects of the found problem (both time and spatial ones). The more tests coincide in the error detection, the more probable it is that the found error is really a flawed value. In this way, it is possible to run the QC method fully automatically. Interaction and supervision by the user remains possible due to the many auxiliary outputs that are created during the process.

2.1.4. How to run the R script

Before running the R scripts, the user should make the following settings in the main *MetQC_frame.R* script (applicable for the INDECIS benchmark dataset; in other cases, please adjust the code accordingly):

- *metqcpath*: variable for setting of working path (usually where the R script is located, but the directory can be set in a different location than are the executive scripts with parameter settings).
- *datapath_in* and *datapath_out*: sets directory with input data files and output files.
- setting filenames: the software creates, by its default, filenames according to the naming convention of INDECIS benchmark dataset: *cc_fF-csv.tgz*, where "cc" stands for a region, with a variable to be set: *ss_region* (for example *se* of *si*), "*fF*" – stands for a name of dataset, with a variable to be set: *ss_bench_datase* (e.g. fC, fB, fQM, ...), and "csv" in the filename means that *csv* files are expected. In case your data are in ECAD format, please apply function *ecad2csv* for conversion from ECAD format to INDECIS CSV files. If other dataset is to be processed, please

adjust the script code by specifying filenames straight inside functions (*run_Get_Info_1_1*, *run_Find_Ngbs_7_2*, *run_QC_7_4*).

- *file_geography* variable: specification of a file with geography (coordinates – best in decimals and possibly also altitudes of processed stations).

- *elem_list*: set elements to process (the codes are described in section 2.2 and correspond to the naming convention of ECAD dataset). By means of variable *ii_seq* – it is possible to select smaller subset of variables to be processed

- the variable *k_ECAD_mode* set to TRUE enforces (overwrites) the proper parameter settings in such a way that the calculation is suitable for running larger datasets (like ECAD, whole Europe); some auxiliary (and not necessary) outputs are then suppressed to get results more quickly.

- the r *k_run_parallel* parameter: the user can choose between running the calculations serially (normally) or in parallel – e.g. 9 meteorological elements may be divided into three groups by three elements (depending on the number of cores in the computer). The number of cores to be used is set in the *Run_function_parallel* function, *n_processes* variable (set to 3 by default).

Then, after the variables settings, the user can run proper functions (applying default settings):

```
run_Get_Info_1_1(ii_seq,elem_list) ;
run_Find_Ngbs_7_2(ii_seq,elem_list) ;
run_QC_7_4(ii_seq,elem_list),
```

or in case of parallel running:

```
Run_function_parralel( elem_list, "run_Get_Info_1_1") ;
Run_function_parralel( elem_list, "run_Find_Ngbs_7_2") ;
Run_function_parralel( elem_list, "run_QC_7_4")
```

Default settings are prepared for each of the meteorological elements individually. In case the user need to change the default parameters settings, it is possible via changing parameters in *Run__Get_info_file, Run__Find_Neighbors, Run__MetQC* functions.
These functions then create proper .txt files that are later used by *launch_QC.r* – where the main calculation is performed. It is also possible to run straight *launch_QC.r* with the .txt parameters files (as it is performed e.g. through ProClimDB, where such .txt files are created by ProClimDB). In such a case, the main settings are given in this parameters file: *R_input_parameters.txt*, in which the first line shows to a file with date file settings (e.g. *Temp/R_input_files_7_4_0.txt*), and the second line shows to a file with parameter settings (e.g. *Temp/R_input_controls_7_4_0.txt*). Examples of settings for various meteorological elements are given in folder: *Templates/* (you will find .txt files for a given function and a given meteorological element there - distinguished in the name of the file, e.g. *Templates/R_input_controls_7_4_0__RR.txt*). The settings in these .txt files correspond to the parameters set in the above-mentioned three main functions. These parameters are described, in detail, later (see in Annex 1).

2.2. Requirements for input data

The software is able to run one region in one step, and all the meteorological elements within it. The used codes for meteorological elements are:

CC - Cloud Cover,
DD - Wind Direction,
FG - Wind Speed,
FX - Wind Gust,
HU - Humidity,
PP - Sea Level Pressure,
RR - Precipitation Sum,
SD - Snow Depth,
SS - Sunshine duration,
TN - Minimum Temperature and
TX - Maximum Temperature.

An info file and a data file are needed for each region. The first one (that is created from the geography file) contains a list of stations with basic metadata: station id, latitude, longitude, and optionally, also elevation (if available), start and stop of the segment, and possibly any other information. If information about periods of measurements is not available, it is later complemented (the *Run__Get_info_file* function), together with information about the number of missing values (this is then used for selection of neighbour reference stations).

The software accepts various file formats and data structures. The most typical file format will probably be a .csv file. In case your data are in the ECAD format, the .CSV files may be created by a special *ecad2csv* function from Jose A. Guijarro (being part of the software).

The software is aimed especially for daily, but monthly or sub-daily data are accepted as well.

2.2.1. Input data

The data file is a (database) table where stations of a given region are gathered. The data structure can be varied. From having all data in one column, through to months or days in columns (Figure 5), to a structure in which you can have the elements in columns (shown on the figure 4 below, such input then allows very quick preliminary quality control for physical limits and control for consistency of the data, for example, that the maximum temperature is higher than the minimum), to a structure in which you have the stations (grid points) in a column (.csv files as prepared in the INDECIS project, see Figure 6). The input file must contain the *ID* of the station (of the character data type). Date is solved either as one column (*Date*) with a Date data type (but different countries use different conventions), or better, as three separate columns: *Year*, *Month* and *Day* of the numeric type. You can also input a *Time* column – for cases of processing sub-daily data. In the case of using a *Time* column, the daily averages or sums can be marked by an AVG/SUM code or by the same time for all cases. The time is of the character type. *Value* columns may be given either with or without decimal places (multiplied by ten).

Given below are examples of some of the accepted data structures. Note that the code for a missing value can be -9999, -999 or NA.

Id	Year	Month	Day	Time	Cc	Dd	Fg	Fx	Hu	Pp	Rr	Sd	Ss	Tg	Tn	Tx
40	2001	1	1	12:00	7	-9999	23	90	78	10113	5	1	14	-15	-96	8
40	2001	1	2	12:00	5	-9999	35	80	90	10041	48	1	0	31	1	51
40	2001	1	3	12:00	5	-9999	25	70	88	10091	12	0	5	34	8	76
40	2001	1	4	12:00	7	-9999	26	80	91	10075	9	0	0	29	-13	43
40	2001	1	5	12:00	8	-9999	41	120	90	10015	145	0	0	49	19	65
40	2001	1	6	12:00	8	-9999	37	110	82	10024	16	0	0	69	52	107
40	2001	1	7	12:00	8	-9999	24	50	93	10136	0	0	0	47	37	57
40	2001	1	8	12:00	8	-9999	13	60	90	10146	0	0	0	33	12	47
40	2001	1	9	12:00	7	-9999	12	30	94	10175	1	0	0	8	-20	33
40	2001	1	10	12:00	8	-9999	11	50	98	10163	95	0	0	5	-31	21
40	2001	1	11	12:00	8	-9999	19	50	87	10207	0	0	2	20	13	42
40	2001	1	12	12:00	1	-9999	16	70	78	10279	0	0	78	-4	-39	48
40	2001	1	13	12:00	1	-9999	21	70	72	10345	0	0	78	-30	-55	10
40	2001	1	14	12:00	1	-9999	26	90	65	10345	0	0	79	-25	-67	18
40	2001	1	15	12:00	0	-9999	23	90	52	10291	0	0	80	-18	-50	33
40	2001	1	16	12:00	2	-9999	12	50	74	10267	0	0	68	-54	-103	5

Figure 4 Example of input data where meteorological elements are given in columns.

	Eg_gh_id	Eg_el_abbr	Year	Month	Time	Val01	Val02	Val03	Val04	Val05	Val06	Val07	Val08	Val09	Val10	Val11	Val12	Val13	Val14	Val15	Val16	Val17	Val18	Val19	Val20	Val21	Val22	Val23	Val24	Val25	Val26	Val27	Val28	Val29	Val30	Val31	Validation
▸	B1BRBY01	TMA	1961	1	21:00	1.8	1.4	5.8	8.0	5.0	4.5	3.6	2.4	2.4	2.4	3.3	2.7	1.7	0.2	7.4	3.0	-0.3	-0.8	-1.0	1.0	0.8	-1.2	1.6	1.0	-0.8	1.3	2.8	-1.2	-4.2	0.0	2.8	B
	B1BRBY01	TMA	1961	2	21:00	2.8	4.5	4.4	2.2	1.8	1.6	1.0	4.2	1.4	2.8	4.0	3.0	7.2	9.5	5.5	3.2	7.0	5.2	1.8	6.0	6.7	5.5	5.2	4.1	3.2	1.0	7.1	7.5	-999.0	-999.0	-999.0	B
	B1BRBY01	TMA	1961	3	21:00	8.4	7.2	8.5	7.3	8.2	12.0	14.2	14.8	16.6	20.5	16.6	14.5	11.6	9.0	14.2	16.6	16.0	19.2	15.4	7.2	4.0	4.2	4.6	10.9	11.8	12.2	9.4	4.0	4.1	7.3	9.0	B
	B1BRBY01	TMA	1961	4	21:00	11.2	12.5	13.4	16.2	20.6	22.5	24.5	22.5	13.6	9.8	17.6	18.1	20.8	23.1	23.8	19.0	19.2	20.4	18.2	18.0	19.4	17.1	17.8	21.0	19.6	19.2	17.0	16.1	19.0	21.0	-999.0	B
	B1BRBY01	TMA	1961	5	21:00	17.8	18.5	20.5	20.0	22.9	23.5	24.0	16.5	13.1	11.4	13.0	11.8	14.8	15.0	13.9	14.4	11.2	14.5	17.2	19.0	19.6	18.2	19.5	20.8	23.2	24.2	23.8	15.6	21.5	12.6	21.4	B
	B1BRBY01	TMA	1961	6	21:00	24.8	25.0	19.0	17.6	24.6	22.4	24.2	24.2	25.6	23.1	20.0	21.2	22.1	18.0	20.0	22.1	24.2	26.4	29.3	22.6	25.2	27.1	24.8	26.5	30.4	29.0	30.0	23.6	22.7	26.5	-999.0	B
	B1BRBY01	TMA	1961	7	21:00	28.8	30.5	31.0	30.1	19.6	18.2	19.2	22.0	18.0	20.6	23.4	27.0	28.0	22.2	23.8	23.0	21.0	20.9	22.0	18.5	19.6	20.5	22.6	20.5	19.0	24.0	26.5	26.8	19.8	18.0	18.1	B
	B1BRBY01	TMA	1961	8	21:00	22.6	25.6	21.8	22.9	26.6	30.0	30.2	29.9	33.0	32.7	29.9	26.7	18.2	21.3	20.6	18.6	20.5	17.0	18.0	18.6	21.6	24.6	18.5	19.0	21.6	25.0	28.0	26.5	27.4	26.4	27.6	B
	B1BRBY01	TMA	1961	9	21:00	28.5	28.4	27.8	28.5	29.1	22.4	24.0	19.0	17.0	13.3	16.0	20.6	22.1	26.5	21.0	27.5	28.5	29.6	31.5	27.1	25.0	26.2	23.2	22.9	22.8	23.2	16.5	23.7	26.5	26.8	-999.0	B
	B1BRBY01	TMA	1961	10	21:00	21.6	22.0	21.6	21.5	20.7	21.5	19.0	20.1	18.6	22.4	21.0	20.6	16.0	13.5	12.0	17.8	18.0	14.2	14.0	14.0	14.1	19.0	14.5	13.7	14.6	14.2	16.5	18.2	14.0	11.0	9.7	B
	B1BRBY01	TMA	1961	11	21:00	13.1	14.2	11.3	6.2	2.2	4.4	4.6	6.1	7.5	12.0	9.8	13.6	13.2	11.5	6.6	4.8	4.6	6.0	6.0	3.0	3.2	4.6	5.2	5.0	5.2	6.2	11.8	9.8	8.1	7.2	-999.0	B
	B1BRBY01	TMA	1961	12	21:00	9.9	10.9	9.5	10.8	12.6	9.9	1.4	1.1	0.7	0.6	0.8	3.8	4.5	-1.6	-2.4	-6.1	-7.0	-1.5	-1.1	-0.2	1.6	-0.8	-0.8	-9.5	-10.6	-6.6	-2.0	0.0	2.8	4.0	5.5	B
	B1BRBY01	TMA	1962	1	21:00	6.0	2.0	-0.2	1.5	-2.4	-4.0	2.5	3.0	0.2	1.0	4.8	4.8	3.2	5.0	3.5	3.3	1.1	1.8	1.6	7.5	5.2	4.8	4.7	2.0	3.0	4.6	4.0	2.8	-2.6	-1.6	-1.4	B
	B1BRBY01	TMA	1962	2	21:00	-3.4	0.2	4.5	-0.2	1.2	3.6	2.8	4.5	2.7	5.5	2.0	4.5	7.8	1.2	1.0	4.0	4.0	1.8	4.1	4.0	3.6	-0.5	-2.5	-1.0	-0.6	2.0	3.0	5.2	-999.0	-999.0	-999.0	B

Figure 5 Example of input data where days of the month are given in columns.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
	Dates	p0034	p0037	p0053	p0066	p0074	p0081	p0089	p0099	p0104	p0105	p0136	p0143	p0152	p0167	p0173	p0177	p0181	p0198	p0222	p0240	p0256	p0269	p0277	p0281	p0282	p0288	p0298	p0321	p0337	p0344	p0357	p0359	p0360	p0361	p03
	1950-01-01	-46	NA	NA	NA	NA	-41	NA	NA	-43	NA	-44	NA	NA	-45	NA	NA	NA	NA	NA	-68	NA	NA	NA	NA	NA	NA	NA	NA	-76	-76	NA	NA	NA	-76	-78
	1950-01-02	-15	NA	NA	-20	NA	-22	NA	NA	-25	NA	-26	NA	NA	-14	NA	NA	NA	NA	NA	-34	NA	NA	NA	NA	NA	NA	NA	NA	-40	-46	NA	NA	NA	-53	-44
	1950-01-03	-4	NA	NA	-6	NA	-9	NA	NA	-13	NA	-17	NA	NA	-5	NA	NA	NA	NA	NA	-24	NA	NA	NA	NA	NA	NA	NA	NA	-24	-32	NA	NA	NA	-22	-32
	1950-01-04	18	NA	NA	17	NA	13	NA	NA	13	NA	11	NA	NA	11	NA	NA	NA	NA	NA	7	NA	NA	NA	NA	NA	NA	NA	NA	-5	4	NA	NA	NA	-14	3
	1950-01-05	18	NA	NA	14	NA	12	NA	NA	13	NA	8	NA	NA	11	NA	NA	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA	-7	0	NA	NA	NA	2	0
	1950-01-06	15	NA	NA	7	NA	6	NA	NA	3	NA	1	NA	NA	8	NA	NA	NA	NA	NA	-2	NA	NA	NA	NA	NA	NA	NA	NA	-3	-12	NA	NA	NA	-20	-13
	1950-01-07	8	NA	NA	5	NA	3	NA	NA	1	NA	-4	NA	NA	-22	NA	NA	NA	NA	NA	-43	NA	NA	NA	NA	NA	NA	NA	NA	-16	-21	NA	NA	NA	-48	-21
	1950-01-08	-24	NA	NA	-20	NA	-26	NA	NA	-40	NA	-50	NA	NA	-14	NA	NA	NA	NA	NA	-46	NA	NA	NA	NA	NA	NA	NA	NA	-44	-49	NA	NA	NA	-27	-47
	1950-01-09	4	NA	NA	-4	NA	-2	NA	NA	-16	NA	-7	NA	NA	18	NA	NA	NA	NA	NA	-2	NA	NA	NA	NA	NA	NA	NA	NA	5	-3	NA	NA	NA	8	0
	1950-01-10	9	NA	NA	-58	NA	-41	NA	NA	-69	NA	-63	NA	NA	12	NA	NA	NA	NA	NA	-102	NA	NA	NA	NA	NA	NA	NA	NA	-52	-118	NA	NA	NA	-24	-12
	1950-01-11	10	NA	NA	-66	NA	-53	NA	NA	-76	NA	-75	NA	NA	6	NA	NA	NA	NA	NA	-108	NA	NA	NA	NA	NA	NA	NA	NA	-64	-141	NA	NA	NA	-37	-14
	1950-01-12	-20	NA	NA	-34	NA	-33	NA	NA	-42	NA	-38	NA	NA	-3	NA	NA	NA	NA	NA	-52	NA	NA	NA	NA	NA	NA	NA	NA	-29	-61	NA	NA	NA	-40	-52
	1950-01-13	-26	NA	NA	-37	NA	-38	NA	NA	-36	NA	-46	NA	NA	-4	NA	NA	NA	NA	NA	-36	NA	NA	NA	NA	NA	NA	NA	NA	-26	-39	NA	NA	NA	-20	-37
	1950-01-14	-44	NA	NA	-50	NA	-49	NA	NA	-66	NA	-126	NA	NA	-46	NA	NA	NA	NA	NA	-155	NA	NA	NA	NA	NA	NA	NA	NA	-137	-126	NA	NA	NA	-139	-11
	1950-01-15	-54	NA	NA	-72	NA	-83	NA	NA	-157	NA	-149	NA	NA	-68	NA	NA	NA	NA	NA	-193	NA	NA	NA	NA	NA	NA	NA	NA	-194	-174	NA	NA	NA	-172	-16

Figure 6 Example of input data where stations are given in columns.

2.1.3. Data info file

Information about stations (metadata) is necessary for the quality control process. The nearest neighbouring stations are selected based on coordinates. Altitude is used during the standardization and evaluation of errors. Further information may be used as well, such as the region of a station, to further

divide stations into meaningful sub-groups. The data info file is created from a geography file (*file_geography* variable) and data file when running the *Run__Get_info_file* function.,

The data info file must contain the *ID* of the station as a character type (same as in the data file). The next important columns are the geographical coordinates. The *Latitude* and *Longitude* columns of the numeric data type, best given in decimals (but degrees may be converted in the scripts as well, in which case, input is of the string data type). *Altitude* is again of the numeric data type. For the right selection of the neighbouring stations, the *Begin* and *End* columns may give information about the period of measurements – but such columns can be created later by the scripts (*Run__Get_info_file* function) if such information is missing. The type of the *Begin* and *End* columns may be Numeric (if you input years) of Date. Other columns may contain further information, like the name of the station, code for region or number of the data for each station in the dataset (*Length*).

Id	Station	Altitude	Latitude	Longitude	Lat_old	Lon_old	Element_	Begin	End	Length	Region	Region_n
105696	LJUBLJANA BEZIGRAD	299	46.065556	14.516945	+46:03:56	+014:31:01	TN8	01.01.1900	30.04.2017	42853	30	Tanja Cegnar
105935	KREDARICA	2514	46.379167	13.853889	+46:22:45	+013:51:14	TN8	01.01.1955	30.04.2017	22765	30	Tanja Cegnar
114849	BABNO POLJE	756	45.600000	14.500000	+45:36:00	+014:30:00	TN8	01.01.1961	31.03.2017	20543	30	Tanja Cegnar
114859	CRNOMELJ-DOBLICE	157	45.600000	15.200000	+45:36:00	+015:12:00	TN8	01.01.1989	31.03.2017	10316	30	Tanja Cegnar
114869	IVANJKOVCI	225	46.500000	16.200000	+46:30:00	+016:12:00	TN8	01.01.2009	31.03.2017	3011	30	Tanja Cegnar
114879	LENDAVA	190	46.600000	16.500000	+46:36:00	+016:30:00	TN8	01.01.1963	31.03.2017	19813	30	Tanja Cegnar
114889	MARIBOR-TABOR	275	46.500000	15.600000	+46:30:00	+015:36:00	TN8	01.01.2005	31.03.2017	4472	30	Tanja Cegnar
114899	PLANINA POD GOLICO	970	46.500000	14.100000	+46:30:00	+014:06:00	TN8	01.03.1961	14.07.2013	19128	30	Tanja Cegnar
114909	PREDDVOR	485	46.300000	14.400000	+46:18:00	+014:24:00	TN8	01.01.1992	10.11.2011	7253	30	Tanja Cegnar
114919	STARSE	240	46.500000	15.800000	+46:30:00	+015:48:00	TN8	01.01.1961	31.03.2017	20543	30	Tanja Cegnar
114929	VOGEL	1535	46.300000	13.800000	+46:18:00	+013:48:00	TN8	01.01.1983	31.03.2017	12508	30	Tanja Cegnar

Figure 7 Data with information about stations necessary for the quality control process.

3. Data quality control results and outputs

To help the user in the decision about acceptance of found errors, the software may give results in various forms. The main one are various tables. Other possibilities are graphs (plots of time series) or maps (giving spatial context). Maps are optional and are part of the ProClimDB software from which these R scripts come (the R code for them can be provided later, to complement the current R scripts).

3.1. Output file with errors and suspicious values

As for tables, there are many ways of outputs (depending on the settings of the scripts, if it is set to produce them or not). One of such tables *(*__errors.csv* with list of errors, or *__errors2.csv* with list of both errors and suspicious values, e.g. *Data/QC_output_PP_se_fQM__errors2.csv*) gives a detailed report on the found errors. These files are useful for a user when a deeper analysis of the detected errors is needed. Figure 8 shows such a table, with the detected list of errors. It contains information about the distance, correlation and altitude of the tested and reference stations. Furthermore, the user can see the

test value compared to the expected value and also the values of the neighbouring stations which were used for the testing and calculation of the expected value. The other columns, not shown on this example, give the results of the individual auiliary statistics and the criteria for determining the reliability of the input data.

	Date			Test station	Calculate value	Difference		Reference stations					
ID	YEAR	MONTH	DAY	ST_BASE	EXPECT_VAL	DIFFS	REMARK	ST_1	ST_2	ST_3	ST_4	ST_5	ST_6
							Distances	64	110	116	128	142	136
128522	TEST STATION			120.0			Altitudes,limit	110	112	169	110	450	112
122702							st_1, Correl	0.8					
100149							st_2, Correl		0.8				
132674	REFERENCE STATIONS						st_3, Correl	Correlation coef.		0.8			
123506							st_4, Correl				0.8		
126122							st_5, Correl					0.7	
100183							st_6, Correl						0.7
128522	1950	4	10	19.9	9.4	-10.5		9.6	9.0	8.0	8.4	5.3	10.0
128522	1950	11	1	12.4	4.7	-7.7		5.3	4.3	3.5	4.6	4.2	4.6
128522	1951	12	24	9.4	0.6	-8.8		-0.4	0.5	0.0	2.3	2.0	-0.8
128522	1953	11	24	-0.6	6.8	7.4		7.5	7.2	6.2	7.9	5.1	6.0
128522	1959	11	26	10.5	3.6	-6.9		2.8	4.5	2.9	4.1	5.6	2.7

Figure 8 Example of data quality control output detailed report (maximum temperature) showing detected suspicious values with expected values, difference, values of the reference stations, correlation coefficient between candidate (test) and reference station, distance and altitude difference.

3.2. Output file with flagged values

Another form of table output (output files with the names **__ecad.csv*, e.g. *Data/QC_output_PP_se_fQM__ecad.csv*) gives all the original values in one column, together with the expected values (estimation of the tested value based solely on neighbour stations, this can be used e.g. for supplementing missing values), difference of the original and the expected value, and this information is accompanied by a quality flag. Thus, the user is able to quickly filter the most problematic values and decide how to handle these marked values. The flags were chosen in this way to support common user needs. The flags reflect the probability of errors and other common problems. The four categories divide the values into errors, suspicious values, repeated values and duplicated values. Suspicious value means the data was evaluated as an error with a probability of 40–70%. A flag for an error means a 70–100% probability of the wrong value. Repeated values means the same values within the tested series. For example, in the case of the temperature, three consecutive values are already suspicious, but in the case of precipitation, it may be 60 values (two months). Duplicated stations mean that the same data were found in the candidate and reference stations. A missing value is marked by a flag too.

The codes used for the flags are:

0…valid
1…erroneous value (70/100% probability of being an error)
2… suspicious value (40/70% probability of being an error)
4...repeated value (the same values of the test series)

5…duplicated value (the same values found in neighbour station)
9…missing value

The columns given in the output are: *Value* - original value, *Expect_val* - expected value (calculated from neighbours), *diffs* - differences or ratios of Expected value and original value, *QC_flag* - used flags –see explanation above, *QC_prob* - probability of being an error.

Id	Year	Month	Day	Value	Expect_val	Diffs	Qc_flag	Qc_prob
P0034	1950	1	11	37	12.000	25.000	0	
P0034	1950	1	12	36	19.000	17.000	0	
P0034	1950	1	13	34	12.000	22.000	0	
P0034	1950	1	14	40	33.000	7.000	0	
P0034	1950	1	15	2	-16.000	18.000	0	
P0034	1950	1	16	44	25.000	19.000	0	
P0034	1950	1	17	20	9.000	11.000	0	
P0034	1950	1	18	44	36.000	8.000	0	
P0034	1950	1	19	36	26.000	10.000	0	
P0034	1950	1	20	-32	-76.000	44.000	2	50.000
P0034	1950	1	21	-5	-3.000	-2.000	0	16.700
P0034	1950	1	22	12	9.000	3.000	0	
P0034	1950	1	23	21	18.000	3.000	0	
P0034	1950	1	24	18	13.000	5.000	0	
P0034	1950	1	25	12	10.000	2.000	0	
P0034	1950	1	26	12	13.000	-1.000	0	
P0034	1950	1	27	11	7.000	4.000	0	
P0034	1950	1	28	18	11.000	7.000	0	
P0034	1950	1	29	8	1.000	7.000	0	
P0034	1950	1	30	-31	-35.000	4.000	0	
P0034	1950	1	31	-26	-32.000	6.000	0	
P0034	1950	2	1	-22	-28.000	6.000	0	
P0034	1950	2	2	8	6.000	2.000	0	
P0034	1950	2	3	8	7.000	1.000	0	
P0034	1950	2	4	0	2.000	-2.000	0	
P0034	1950	2	5	-14	-19.000	5.000	0	
P0034	1950	2	6	-33	-10.000	-23.000	2	50.000
P0034	1950	2	7	-23	-18.000	-5.000	0	
P0034	1950	2	8	-40	-33.000	-7.000	0	
P0034	1950	2	9	-39	-35.000	-4.000	0	
P0034	1950	2	10	-59	-26.000	-33.000	2	50.000
P0034	1950	2	11	-55	-30.000	-25.000	0	33.300
P0034	1950	2	12	-22	5.000	-27.000	0	33.300
P0034	1950	2	13	-18	13.000	-31.000	0	33.300
P0034	1950	2	14	-18	-7.000	-11.000	0	

Figure 9 Example of data quality control output showing detected problematic values marked with flags, accompanied with expected values and differences.

3.3. Graphical outputs

The tabular outputs can be automatically used for further processing. In operational mode, the suspicious values and errors may be discarded from further processing. Besides these tabular outputs, the software can also produce graphical outputs that then help a user to better understand what is behind the error (suspicious value) detection. The example in Figure 10 shows a comparison of a tested time series with its neighbours, either absolutely, or relatively, together with the values of the evaluation statistics.

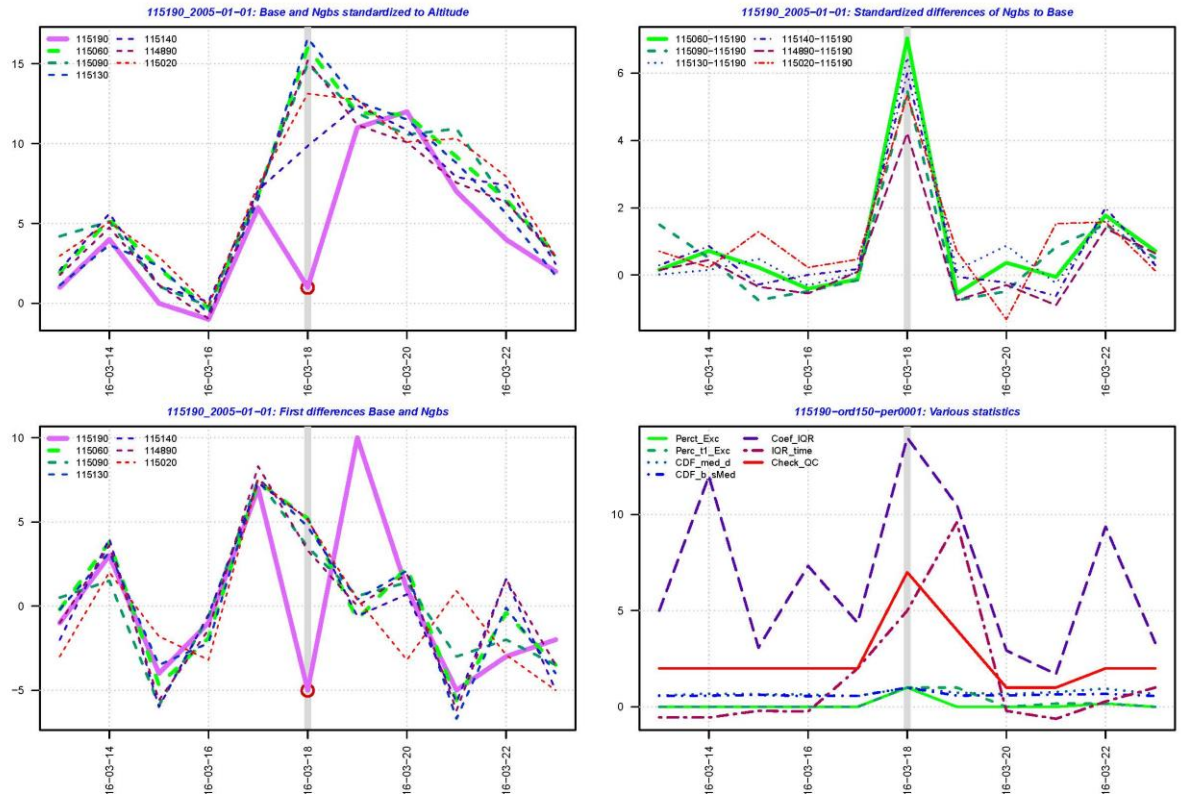

Figure 10 Example of outlier detection. Top left: tested and neighbour stations – the values of neighbour stations are converted to the tested station altitude. Bottom left: first differences series for tested and neighbour stations. Top right: standardized differences between tested and neighbour series. Bottom right: various statistics used for evaluation of outliers.

4. Conclusions and results

The R script that is responsible for data quality control is based on the original source code (functions) of the ProClimDB software and includes the features and smart functionality of other methods. It was re-programmed into R software, thus, is open and may be further improved by the broad scientific community.

An efficient and effective automated method for data quality control has been developed, programmed (open source) and validated on selected regions throughout Europe. It utilizes a combination of several statistical tests. Details on the quality control process will be available through an article (scientific basis). No single approach has been found adequate and sufficient for automation of the data quality control process; only their combination leads to satisfying results in the discovery of real outliers and suppression of false alarms, given the requirements on efficient and automated computing. The MetQC software gives

full control regarding the quality control process, and through a variety of tabular or graphical outputs, gives the user vital information for further investigation of the problems found.

The most important comments and recommendations on the QC process may be recapitulated as follows:

- Quality control must not be a “black box”, the user has to have full control and should be informed in detail about the process.
- In the averaging (generally, in any type of aggregation), errors are masked to a certain extent (in annual, seasonal, monthly, but even in daily values). It is recommended to test unprocessed, directly measured data (e.g. observed hourly data).
- The selection of reference (neighbour) stations to reflect similar geographic conditions is crucial; this may be based on correlations, distances or other metrics.
- For the automated method to give acceptable results, it has to combine several statistical approaches. In other words, various methods have to coincide in their error indication. Only in this way are false alarms suppressed.
- Automated methods of QC are necessary for large datasets, but the user still needs to have full control of the process to be able to intervene and make their own (expert) decision. For this, the tool should give the user appropriate outputs to support their decision making.
- Graphical output is beneficial – it helps a user control the process.
- More complicated meteorological elements, such as precipitation and wind, should be tested (validated) on a dense station network. The spatial relationship is weak otherwise, and leads to failures in error detection.
- The QC method discussed in this report was validated for European data and the proper default settings have been found. The method is made general in the sense that the user can tune the settings of the individual statistical approaches – e.g. criteria used, and make it more appropriate for their region, if needed.

References

Peterson, T.C. (1998): Homogeneity adjustments of in situ atmospheric climate data: a review. Int. J. Climatol. 18, 1493-1517.

Štěpánek, P., Zahradníček, P., Brázdil, R., Tolasz, R. (2011): Metodologie kontroly a homogenizace časových řad v klimatologii (Methodology of quality control and homogenisation of time series in climatology on the example of the Czech Republic in 1961–2009). ČHMÚ, 118 pp. ISBN 978-80-86690-97-1

Štěpánek, P., Zahradníček, P., Farda, A. (2013): Experience with data quality control and homogenization of daily records of various meteorological elements in the Czech Republic in the period 1961–2010. Időjárás, 117, 1, 123–141.

Štěpánek, P. (2012): ProClimDB – software for processing climatological datasets. CHMI, regional office Brno. http://www.climahom.eu/ProcData.html

Annex 1 – description of parameters of the main R functions

Annex 1.1 Function Run_Get_info_file

MetQC_path: directory where scripts with QC to run are located
launch_script: a script to launch processing itself
a_files: txt file with parameter settings for input and output files - to be used in the main QC script (launch_QC.r)
a_controls: txt file with parameter settings of the function - to be used in the main QC script (launch_QC.r)
templates are saved here: a_controls_template
this is a new file created based on the a_controls_template, with changed parameter settings - input into this function, usually saved into Temp folder

input_data_file: input data file
geography_file=: with list of stations and their coordinates
info_file: output info_file (list of stations, with coordinates joined from geography, period of measurement, and number of missing values)
data_file: output data file, with filled gaps - replaced with missing value code (missval)

missval: code for missing values (in data file)
k_fill_gaps - do you want to fill gaps? Put 1, otherwise 0
n_fill_max_years - maximum number of years with missing values to be filled with NA, otherwise the series is split into more parts

Annex 1.2 Function Run_Find_Neighbours

MetQC_path: directory where scripts with QC to run are located
launch_script: a script to launch processing itself
a_files: txt file with parameter settings for input and output files - to be used in the main QC script (launch_QC.r)
a_controls: txt file with parameter settings of the function - to be used in the main QC script (launch_QC.r)
templates are saved here: a_controls_template
this is a new file created based on the a_controls_template, with changed parameter settings - input into this function, usually saved into Temp folder

input_data_file: input data file (not needed, can be void)
info_file: input info_file (list of stations, with coordinates joined from geography, period of measurement, and number of missing values)
ngbs_info: output file with list of neighbours for each base station

missval: code for missing values (in data file)
n_number_of_stations: number of neighbours to find
limit_distance: limit distance (weight=0)

limit_diffs_altitude: limit altitude difference (weight=0)
k_clever_altitude_sel: put 1 to use a larger sample of altitudes for the linear regression
n_period_length: put 1 to cut the periods into segments with the given length
n_period_overlap: in case of period segments, length of overlap
k_common_period: put 1 for finding only neighbours with the common period
k_the_same_region: put 1 for finding only neighbours with the same region

Annex 1.1 Function Run_MetQC

MetQC_path: directory where scripts with QC to run are located
launch_script: a script to launch processing itself
a_files: txt file with parameter settings for input and output files - to be used in the main QC script (launch_QC.r)
a_controls: txt file with parameter settings of the function - to be used in the main QC script (launch_QC.r)
	# templates are saved here: a_controls_template
	# this is a new file created based on the a_controls_template, with changed parameter settings - input into this function, usually saved into Temp folder
a_controls_input: txt file with parameter settings, for a particular element, can be the same as a_controls_template (saved in Template folder),
	# if specified: it replaces individual parameter settings given in the input of this function <<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<

data_file: input data file
info_file: input info_file (list of stations, with coordinates joined from geography, period of measurement, and number of missing values)
ngbs_info: input file with list of neighbours for each of the base stations
ngbs_file: output file with neighbours info and quality control (main output)
ngbs_stats: output file with neighbours and base station statistics

parameter settings:
missval: code for missing values (in data file)
k_add_standardized_vals_cols: 1 for adding columns with standardized values into the output,
k_transformation: transformation of values
	# if k_transformation==1:
	# ways of tranformation, to be specified in the n_transformation variable (and passed into file_parameters):
		# n_transformation==100: "differences (Y-X), NO transformation"
		# n_transformation== 0: "ratios (Y/X), NO transformation"
		# n_transformation== -1: "ratios X/Y, NO transformation"

		# n_transformation==-10: "Equitable ratios, NO transformation (2nd/1st or -1st/2nd, the greater value, negative for 2nd/1st, original one (1st or -2nd) if the other is"

		# n_transformation==1: "Differences - values transformed with log(x)"
		# n_transformation==2: "Differences - values transformed with sqrt(x)"

n_transformation==3: "Differences - values transformed with log(x+1)"
n_transformation==4: "Differences - values transformed with sqrt(x+1)"
n_transformation==5: "Differences-values transformed with sqrt(x)+sqrt(x+1)
"

k_AVG_standardization: 1 for standardization to average
k_STD_standardization: 1 for standardization to standard deviation
k_Standardize_to_ALTitude: 1 for standardization to altitude
k_Regr_for_indiv_cases: 1 for performing regression for each time step individually
k_1_station_apply_monhly_AVG: 1 for applying monthly AVG (+STD) in case there is only one station available (thus regression cannot be calculated),
moreover, controls values, during individual regression, if neighbours' altitudes are similar, but base station altitude is very different ... then the regression does not work well - and is not performed
n_Regression_correction: correction by means of determination coefficient, 1 - full correction, 0 - original linear regression
k_Outliers_check: put 1 for checking standardized neighbours values (comparing with original values of all neighbours) if they are not outliers
n_Outliers_check: CDF values for the outliers check
k_Add_IQR_coef_value: put 1 for adding other characteristics - CDF_b_sMed - probability (CDF) of base station compared to median from standardized (to altitude) neighbours
k_Add_Expected_value: put 1 for calculating expected value,
n_Power_for_weights: expected values are calculated as weighting average with distances, distance is used with the specified power