HOMOGENEITY ANALYSIS OF DAILY TIME SERIES OF ECA&D

Pavel Zahradníček, Petr Štěpánek, Aleš Farda, Petr Skalák

Global Change Research Centre AS CR, Brno, Czech Republic

zahradnicek.p@czechglobe.cz

Introduction

In many scientific disciplines is needed to process long time series of meteorological elements. In recent years considerable attention has been devoted also to analysis of daily data. Prior to any analysis, the need to homogenize data and check their quality arises (Figure 3). Unfortunately, most of the time series of atmospheric data with a resolution of decades to centuries contains inhomogeneities caused by station relocations, exchange of observers, changes in the vicinity of stations (eg. urbanization), changes of instruments, observing practices (eg a new formula for calculating daily average, different observation times, Figure 4), etc (Štěpánek et al 2009). There are several databases of meteorological elements with different data quality. One of them is free available dataset of daily meteorological elements from European Climate Assessment & Dataset. This database was used to create a regular grid points EOBS, which are often used to validate climate models or other climate analysis and there good quality is very necessary. That was reason for testing quality control and homogeneity of this dataset. In this article we present results for temperature.

Data and methods

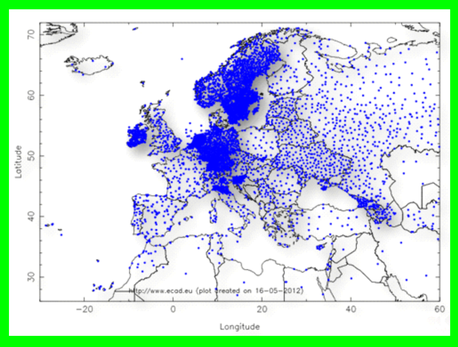

Figure 1. Meteorological stations contains in ECA&D database

ECA&D project was initiated by European Climate Support Network of EUMETNET in year 2002. Coordinated was by the Royal Netherlands Meteorological Institute (KNMI). In ECA&D project 58 institutes from 62 countries participate. ECA&D daily database contains 31058 quality controlled series of 12 climate variables at 6596 meteorological stations in 62 countries, but only half of them are public (Figure 1). In our work we start with the mean daily temperature of 477 meteorological stations from 23 countries (Figure 2). Most of stations were from Norway (112), Spain (91) and Germany (86). Vice versa for same countries is available only few stations (for Czech Republic only 1, for Poland 2) or no stations (Hungary). Altitude of stations are from -4 m (Rotterdam, NL) to 3106 m (Sonnblick, AT).

In this work, data quality control was carried out by combining several methods: (i) by analyzing difference series between candidate and neighbouring stations – i.e. pairwise comparisons (ii) by applying limits derived from interquartile ranges (this can be applied either to individual series, i.e. absolutely or, better, to difference series between candidate and reference series, i.e. relatively) and (iii) by comparing the series values tested with "expected" values – technical series created by means of statistical methods for spatial data (e.g. IDW, kriging). A method for outlier detection that could be automated to the greatest extent was a priority, since millions of values had to be processed for each meteorological element. Such a method was finally found and successfully applied.

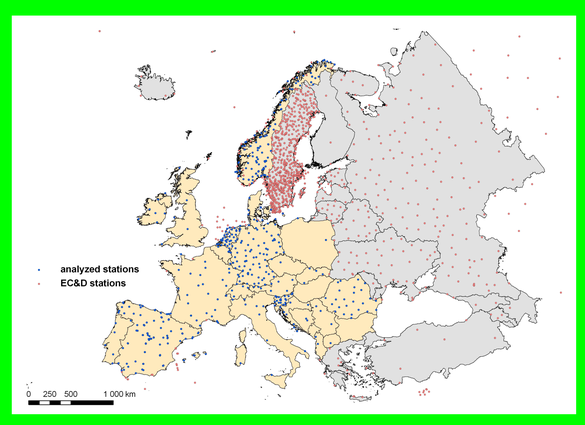

Figure 2. Analysed stations of ECA&D database

Detection of inhomogeneities was performed using monthly means. In the homogenization of the time series, the use of various statistical tests and types of reference series made it possible to increase considerably the number of homogeneity tests results for each series tested and thus to assess homogeneity more reliably. The relative homogeneity tests applied were: Standard Normal Homogeneity Test [SNHT], the Maronna and Yohai bivariate testand the Easterling and Peterson test. Reference series were calculated as weighted average from the five nearest stations (measuring within the same period as candidate series), with statistically significant correlations.

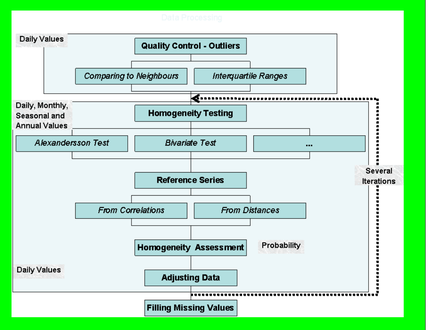

Figure 3. Scheme of usual data processing during data quality control and homogenization (preparing data for timeseries analysis)

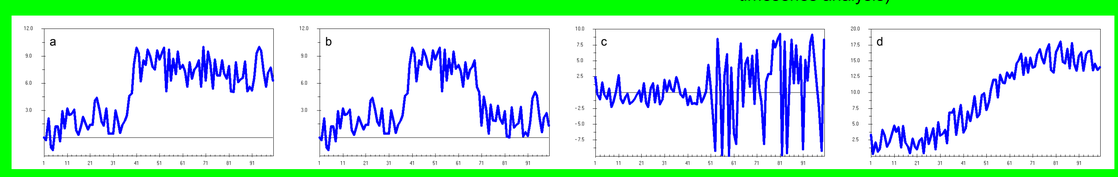

Figure 4. Type of inhomogeneity in the climatological time series a) one shift in mean b) two shifts in mean c) inhomogeneity in variance d) inhomegeneity in trend

Data were corrected on a daily scale. We created our own correction method (called DAP – Distribution Adjusting by Percentiles), an adaptation of a method for the correction of regional climate model outputs by Deque (2007). Our process is based on comparison of percentiles (empirical distribution) of differences (or ratios) between candidate and reference series before and after a break. The above-mentioned steps (homogeneity testing, evaluation and correction of inhomogeneities detected) were performed in several iterations. At each iteration, more precise results were obtained. For data processing, the software package consisting of AnClim and ProClimDB was created by Petr Štěpánek (www.climahom.eu) . For more information about quality control and homogenization please check Štěpánek 2009 and 2011.

Results

Spatial correlation

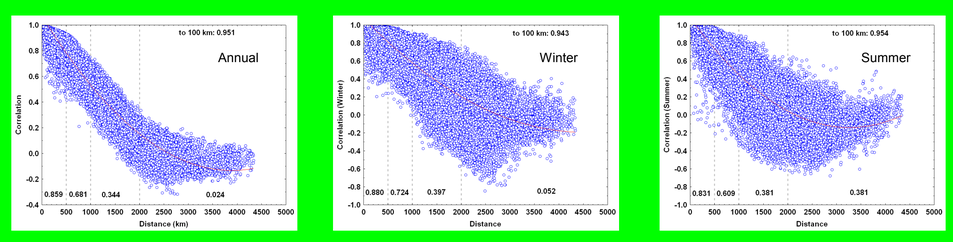

Figure 5. Correlation coeficient of the monthly mean temperature of ECA&D database in the dependence on the distance for annual, winter and summer seasons

For the best selection of neighbours and other time series analysis is is necessary to know the spatial connection of the meteorological element. The correlation coefficient was calculated reciprocaly for each station in the monthly scale and for selected station Wurzburg in the daily scale (Figure 6). Correlation coefficient decrease with distance. Spatial correlation is low for station with distance higher than 100 km (Figure 5).

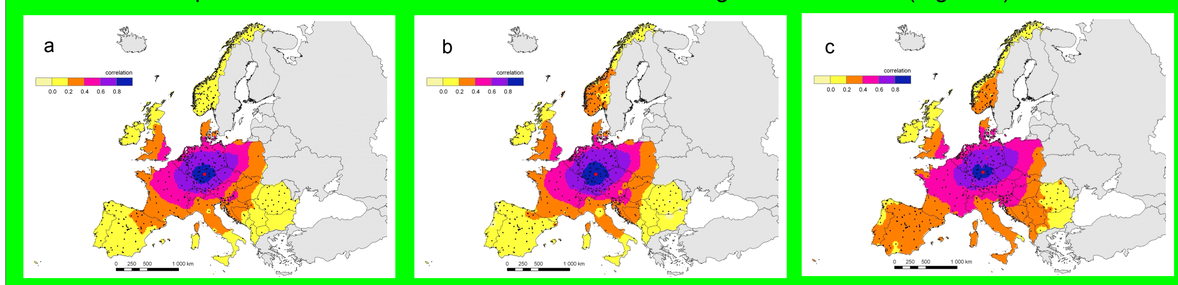

Figure 6. Correlation coeficient of the daily mean temperature of ECA&D database calculated for Wurzburg station for a) mean correlation coefficient calculated from all months b) April and c) October

Quality control

From more than 10 million values were 1201 detected as suspicious. It is only 0.011% of all tested values. Very similar results were detected as from the quality control testing in the Czech Republic (Štěpánek et al 2013). In monthly and daily data, outliers are mostly masked, therefore it is better to test subdaily data. The most common size of detected error, thus the difference between the test and the calculated values was between -6 to -8°C, or between 4 to 6°C (Figure 7a). Many of these outliers were zero values (0°C; Figure 7b). This is very common database error, because the zero values is instead of a missing values flag. In the cases, when neighbours stations temperature are near the 0°C, this errors are very bad detectable. Vice versa there were detected 10 values higher than 50°C (for example 190.5°C), which is a physical nonsense in European condition. ECA&D performed own basic quality control of data and each value have flag with mark 0 for valid data and 1 for suspect data (Figure 9). Many of these detected physical nonsense and zero values have right flag (1), but we find same zeros values without suspect flag. We recommend to users work with these flags and make the additional own quality control.

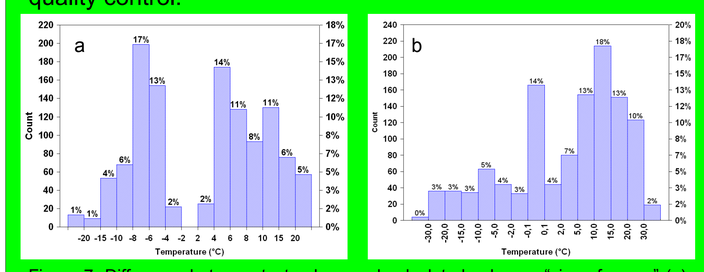

Figure 7. Difference between test values and calculated values – "size of errors" (a) and scale of detected temperature outliers (b)

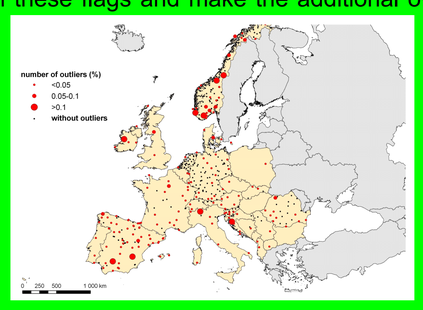

Figure 8. Percentage of detected outliers for each analysed ECA&D station

1679,	4985,	19910803,	240,	0
1679,	4985,	19910804,	240,	0
1679,	4985,	19910805,	258,	0
1679,	4985,	19910806,	264,	0
1679,	4985,	19910807,	279,	0
1679,	4985,	19910808,	283,	0
1679,	4985,	19910809,	275,	0
1679,	4985,	19910810,	1906,	1
1679,	4985,	19910811,	292,	0
1679,	4985,	19910812,	287,	0
1679,	4985,	19910813,	277,	0
1679,	4985,	19910814,	290,	0
1679,	4985,	19910815,	266,	0
1679,	4985,	19910816,	241,	0

337,	1187,	19650522,	205,	0
337,	1187,	19650523,	200,	0
337,	1187,	19650524,	0,	1
337,	1187,	19650525,	216,	0
337,	1187,	19650526,	234,	0
337,	1187,	19650527,	0,	1
337,	1187,	19650528,	229,	0
337,	1187,	19650529,	0,	1
337,	1187,	19650530,	195,	0
337,	1187,	19650531,	193,	0
337,	1187,	19650601,	222,	0
337,	1187,	19650602,	210,	0
337,	1187,	19650603,	200,	0
337,	1187,	19650604,	236,	0
337,	1187,	19650605,	223,	0
337,	1187,	19650606,	182,	0
337,	1187,	19650607,	170,	0
337,	1187,	19650608,	0,	1

Figure 9. Example of detected errors in ECA&D mean temperature database, nonphysical value in the left (190.6°C) and many nonsense zeros values in the right

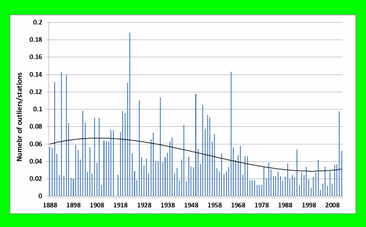

Figure 10. Number of outliers per station in ECA&D mean temperature database for each year

Homogenization

The homogenization process detected 429 breaks at 179 stations (Figure 11). It means 37.5% of tested values were inhomogeneous. In comparison with the results of the same process with the data from the Czech Republic is this percentage lower. Mean temperature series are from 55% inhomogeneous at the area of the Czech Republic. Time series can have more than one break. One break were detected in 41.3% of inhomogeneous series, two breaks in 21.8%, three breaks in 20.7% and more than 3 breaks were detected in 16.2% of inhomogeneous stations. For Milan stations were detected most breaks of all tested series, namely 13. Temperature series for Milan station is for period 1763-2008 and in the average is one break for each 20 years.

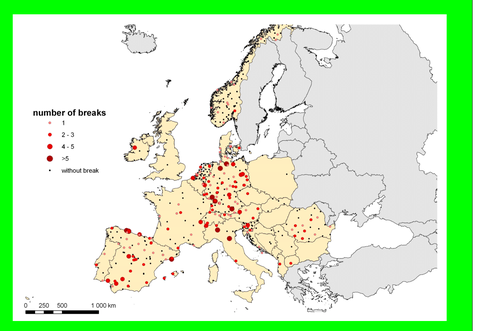

Figure 11. Number of detected breaks for each analysed ECA&D station

Size of correction was uniform on both side in most months (Figure 12b). In June and October correction were in the negative scale prevailed. Annual cycle of absolute correction values are shown in Figure 12c. Highest absolute correction was in summer and in January. The absolute average size of breaks was between 0.3-0.4°C. It is similar results as in the case of homogenization in the Czech Republic.

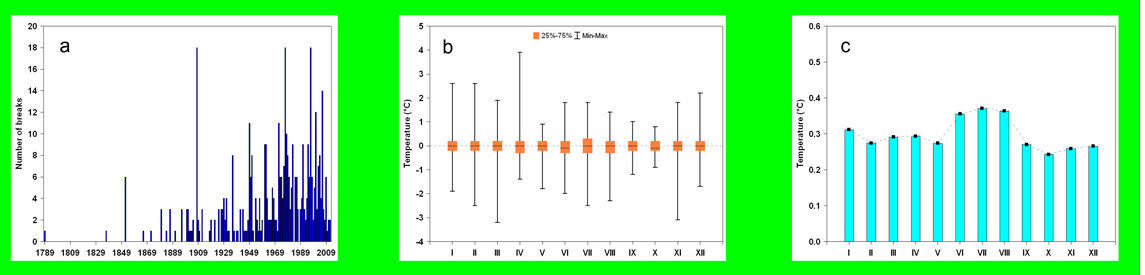

Figure 12. Number of inhomogeneity fo each year (a) and box plot of correction factor (b) and absolute values of correction (c) for mean temperature of ECA&D database.

Conclusions

There are several databases with climatological elements covering different periods and areas. Time series of these databases are of varying quality. As described in this work is necessary to data quality control and homogenization. ECA&D is one of most used databases and the stations were used for recalculating to the regular grid network with 25 km resolution, which is called EOBS. Many climatological studies work with EOBS database, mainly for verification climatological models. We create own approach for quality control and homogenization of daily data. These methods were used for testing quality condition of mean temperature of stations bellows to ECA&D database. We found 0.011% of suspicious values. Many errors were incorrect marking of missing values, where instead flag were zero values. Furthermore we detected few values with mean temperature higher than 50°C (maximum 190.6°C).

We detected 37.5% inhomogeneous mean temperature series. For more than 50% of this series were detected two and more break. Maximum was detected 13 breaks at the station Milan (1763-2008). The average size of break was between 0.3-0.4°C. The correction of daily mean temperature was done by our own method (DAP).

In the future we start with control of other meteorological elements and applying next methods of control and homogenization and comparing results.

Acknowledgment: The work was supported by the project InterDrought (no. CZ.1.07/2.3.00/20.0248) and by the project Adaptation (No. CZ. 1.07/2.4.00/31.0056).

ClimaHom.eu

Tools for processing and homogenization of large climatological datasets