Homogenization of daily temperature and precipitation data using ProClimDB software - a comparison of methods

J.R. Coll

Outline

The HOME COST Action ES0601 (Advances in Homogenization Methods for climate series and integrated approach) allowed the development of the STSM in Brno (Czech Republic) for two weeks at the CHMI (Czech Hydrometeorological Institute).

1. Introduction

Knowledge of climate variability and change is possible by obtaining long climatic series and quality, but often the records can be altered by non-climatic conditions. There are several methods or techniques to detect and correct inhomogeneities in daily data, such as Percentiles *(Stepanek 2008 and 2009)*, HOM *(Della Marta and Wanner 2006)*, SPLIDHOM *(Mestre et al. submitted)*, Vincent's method *(Vincent et al. 2002)*, Quantile Matching *(Wang 2009 and 2010)*,… that you can run in a software obtaining reliable records and comparable.

For the homogenization of daily data there are specific goals:

- Getting acquainted with daily data correction process using ProClimDB and AnClim software.
- Preparing data for the correction.
- Running correction methods (homogenization).
- Evaluation of the obtained results.

AnClim *(Stepanek 2007)* is handy and friendly software that needs txt files to work. It have a easy structure and it allows to us to calculate basic statistics, running quality control, homogenization using different methods (SNHT *(Alexandersson 1986,1995)*, Bivariate test *(Potter 1981)*, Easterling and Peterson Test,… *(Easterling and Peterson 1995)* and analyse the obtained results (regressions, correlations, filters,…). It can also create plots to show the series or results.

On the other hand, ProClimDB *(Stepanek 2008)* is a powerful database tool. You can choose monthly or daily mode, it depends of the data file. With geography file you can calculate reference series using distances or correlations to find neighbour stations for

adjustment. It's a interesting homogenization software that it calculates all statistic parameters, finding neighbours, running quality control and homogenization using different methods (Percentiles, Empirical, HOM, SPLIDHOM, Vincent's method, Wang,…). In this STSM has been used ProClimDB to detect and adjust daily data series using different methods.

2. Data and methods

For training in this STSM, the Czech Hydrometeorological Institute gave us Czech Republic benchmark dataset composed of daily temperature and precipitation data of whole country *(Stepanek 2009)*. ProClimDB tested 265 daily temperature series approximately for the period 1961-1980 and 676 daily precipitation series (1961-2007). Both series were subjected quality control before homogenization.

Daily precipitation series are inhomogeneous and we chose the bigger break in the middle of "B1NAPA01" series (1985) for adjust.

Daily temperature series are homogeneous, but we created a known break in 1971 in "B1HOLE01" series. We joined two different temperature series creating artificial change (1961-1970, "O3VSET01" series and 1971-1980 "B1HOLE01" series ("O3VSET01" series is the last neighbour selected for ProClimDB of "B1HOLE01" series)). Now we have a new inhomogeneous semi-synthetic series.

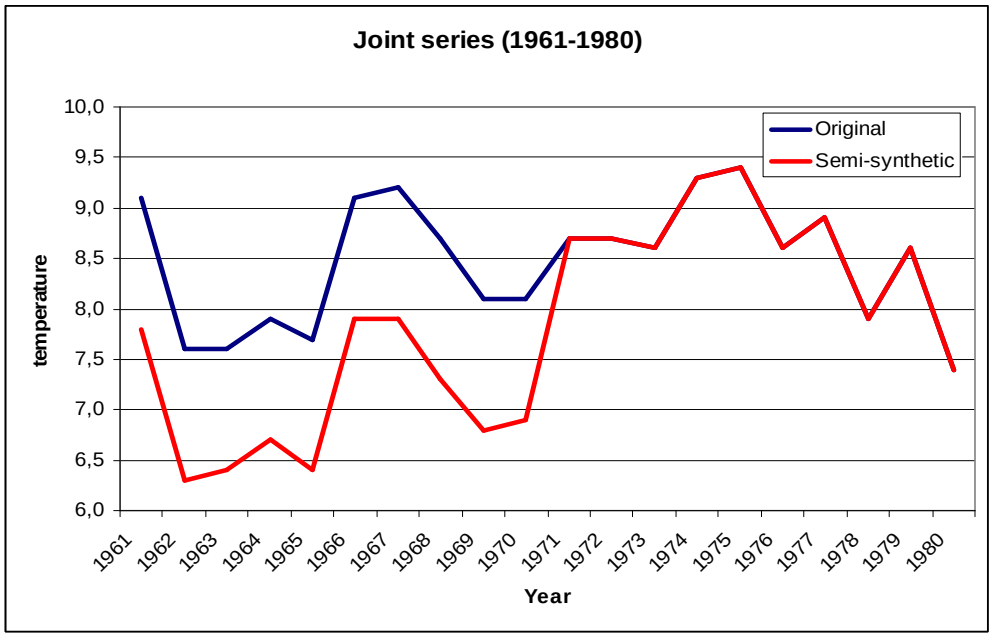

Figure 1. Representation of original and semi-synthetic temperature series (1961-1980)

ProClimDB needs data file to create data info file and, then, the software needs to import geography of the stations to run functions (monthly average,…). After that, you can process data quality control to find errors and outliers comparing to neighbours or using interquartile ranges (daily data). To detect inhomogeneities in data file you must do homonegeity test using some tests (Bivariate test, SNHT,…) and then the software can create reference series from distances or correlations for adjust inhomogeneities.

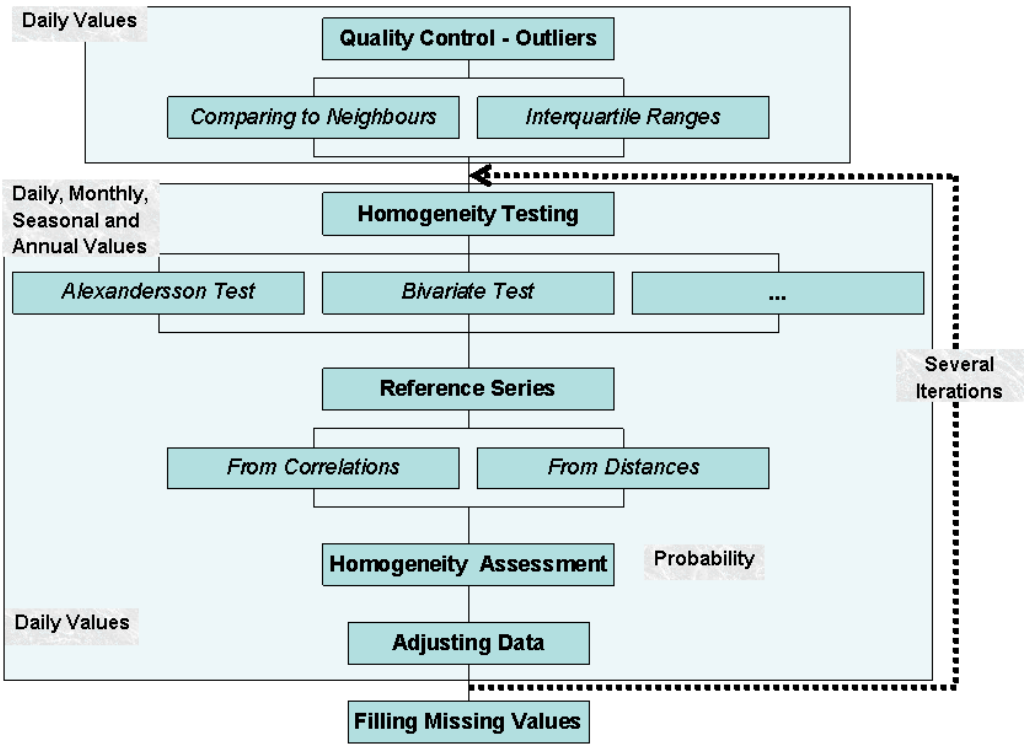

Figure 2. General scheme of data processing (*Stepanek 2008*)

You can choose some homogenization methods in ProClimDB to compare different adjustments, but we selected Percentiles method, Empirical, HOM, SPLIDHOM, Vincent's method and QM by Wang for temperature series. In precipitation series we compared the adjustments of Empirical and Percentiles method, because the other methods work not good in discontinuous variables (e.g. precipitation).

Percentiles method is an adaptation of a method for the correction of regional climate model outputs by *Déqué 2007*, itself based on assumptions similar to those implicit in methods described by *Trewin and Trevitt 1996* and *Della-Marta 2006*, which apply variable correction according to individual percentiles (or deciles). Our process is based on comparison of percentiles (empirical distribution) of differences (or ratios) between candidate and reference series before and after a break. Percentiles are estimated from

candidate series and values for differences of candidate and references series are taken from the same time (date). Each month is processed individually, but also taking into account the values of adjacent months before and after it to ensure smoother passage from one month to another. Candidate – reference differences for individual percentiles are then differenced before and after a break and smoothed by low-pass filter to obtain a final adjustment based on a given percentile *(Stepanek 2008)*.

HOM method uses a nonlinear model to estimate the relationship between a candidate station and a highly correlated reference station. The model is built in a homogeneous subperiod before an inhomogeneity and is then used to estimate the observations at the candidate station after the inhomogeneity using observations from the reference series. The differences between the predicted and observed values are binned according to which decile the predicted values fit in the candidate station's observed cumulative distribution function defined using homogeneous daily temperatures before the inhomogeneity. In this way, adjustments for each decile were produced *(Della Marta and Wanner 2006)*.

The SPLIDHOM method relies on a direct non-linear cubic spline regression method. This method should be regarded as an alternative to HOM method developed by Della-Marta and Wanner (2006). We show that is method is able to correct mean of the series, but also high order quantiles and moments of the perturbed series *(Mestre et al. submitted)*.

On the other hand, we used Vincent's Method, which daily adjustments are derived from the monthly adjustments. They are obtained by linear interpolation between midmonth "target" values that are objectively chosen so that the average of daily adjustments over a given month is equal to the monthly adjustment *(Vincent et al. 2002).*

The Quantile Matching (QM) algorithm is similar to that proposed in *Wang et al. (2009)*. The objective of the QM adjustments is to adjust the base series so that the empirical distributions of all segments of the detrended base series match each other. The adjustment value depends on the empirical frequency of the datum to be adjusted (i.e. it varies from one datum to another in the same segment, depending on their corresponding empirical frequencies) *(Wang et al. 2010)*.

All methods run in ProClimDB choosing the same settings: 75 days for smoothing, 1 month for moving window, 5 years before and after the break and not adjust with correlations = or <0. (Fig. 3).

Figure 3. Example of ProClimDB screen applying the settings used for adjust

3. Results

Daily adjustments are showed below using six kinds of homogenization methods in semi-synthetic temperature series (B1HOLE01). The homogeneous series is drawn with black line and the semi-synthetic series (red line) have the known break in 1971.

You can see Percentiles, Empirical, HOM, SPLIDHOM and Vincent adjustments are similar to the original series (Fig. 4). That means daily adjustments homogenized semi-synthetic series. In this case, choosing the settings above and homogenizing a big known break, Quantile Matching (Wang) obtained different results. Might be this result of Wang procedure was so different because it's an absolute method and not use reference series for adjust. But it's not possible to know which's the best homogenization method only viewing the plot, so it's important to show some statistics of adjustments.

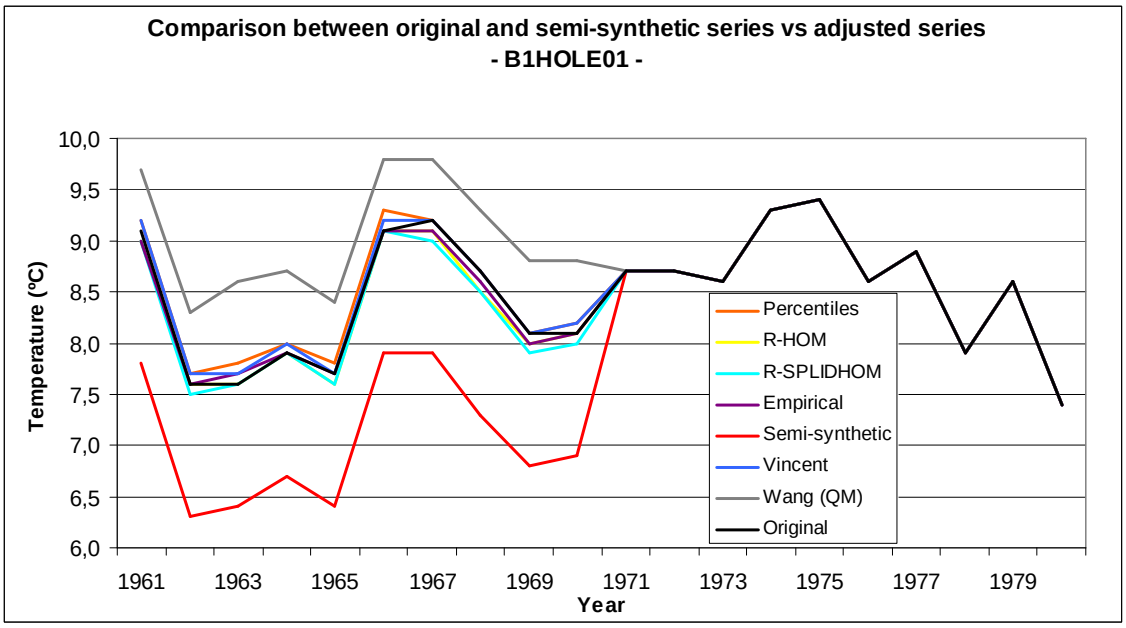

Figure 4. Temperature series adjustments using different homogeneity methods

In ProClimDB exists a function which calculates the validation and differences of both series (original and adjusted). Results of four statistic parameters for whole period (1961-1980) are showed in the table below.

B1HOLE01	Statistic parameters			
	AVG anom.	STD anom.	Correlation	RMSE
Percentiles	0,047	0,076	0,987	0,629
Empirical	0,052	0,059	0,987	0,624
R-HOM	0,04	0,069	0,987	0,645
SPLIDHOM	0,021	0,102	0,987	0,647
Vincent	0,672	-0,049	0,982	0,952
Wang (QM)	0,985	-0,242	0,966	1,451

Table 1. Statistic parameters of different methods in temperature series

This results demonstrates little differences in first four methods (Percentiles, Empirical, HOM and SPLIDHOM), but you can see the differences in Vincent and Wang methods. It exists lower correlation and bigger difference in RMSE and average/STD anomalies in both methods.

In monthly data you can see similar results, but note Wang's method obtained worse results (AVG and RMSE) in winter months and Vincent's method, in opposite, the worse results were in summer months (Fig.5,6,7,8).

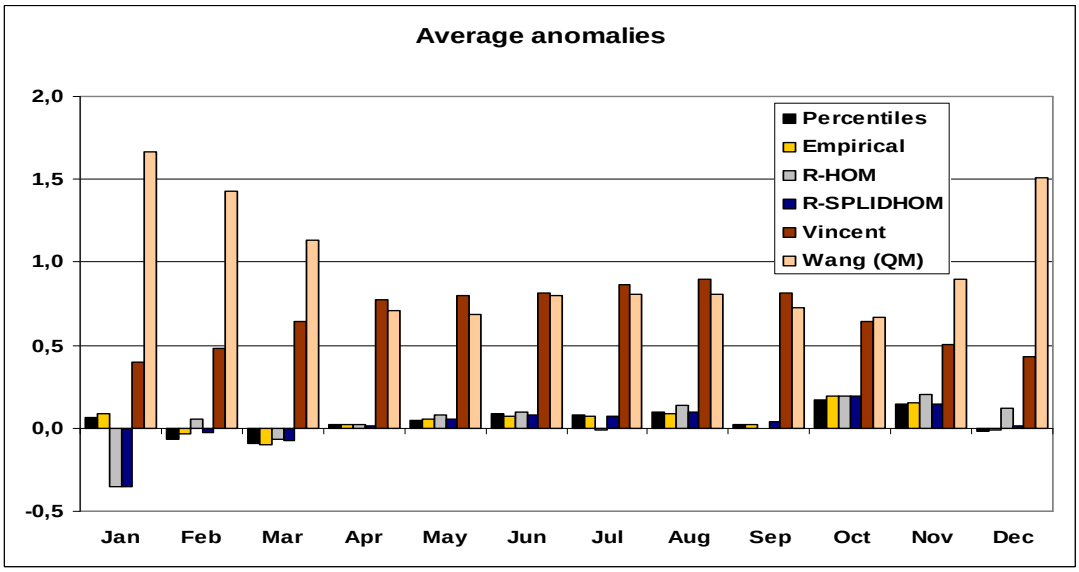

Figure 5. Monthly average temperature anomalies for different adjustments (1961-1980)

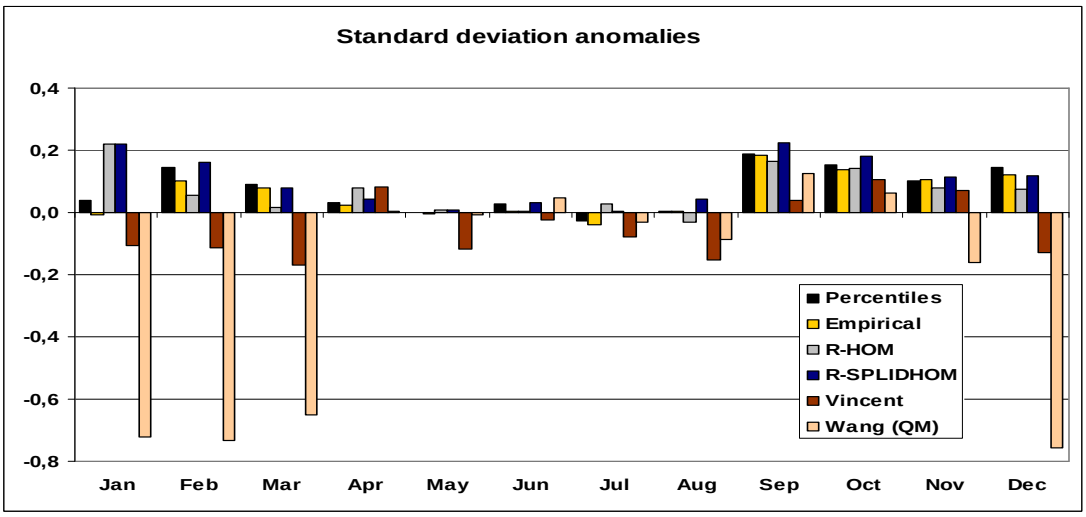

Figure 6. Monthly standard deviation anomalies for different adjustments (1961-1980)

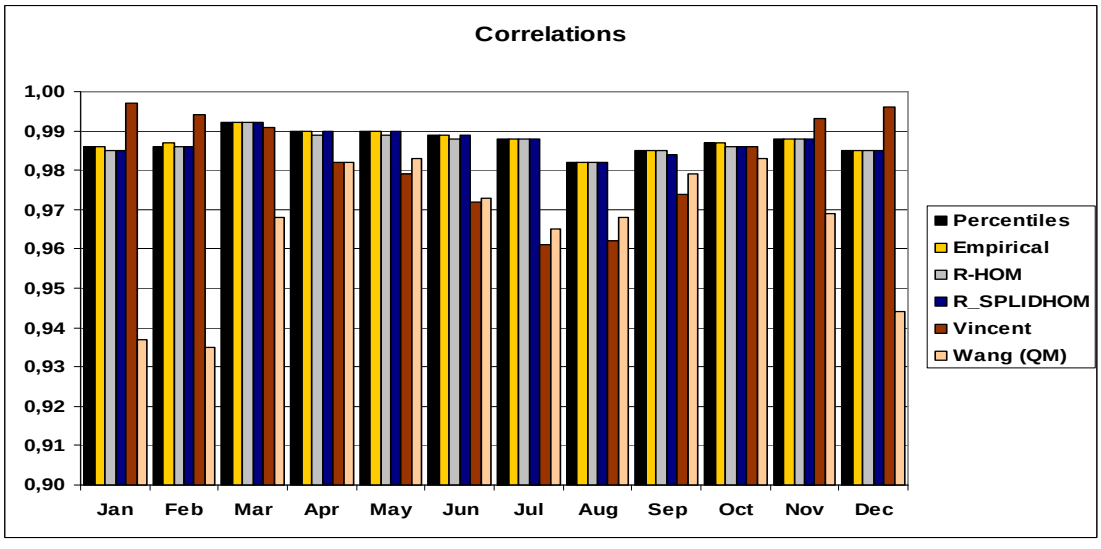

Figure 7. Monthly correlations for different adjustments (1961-1980)

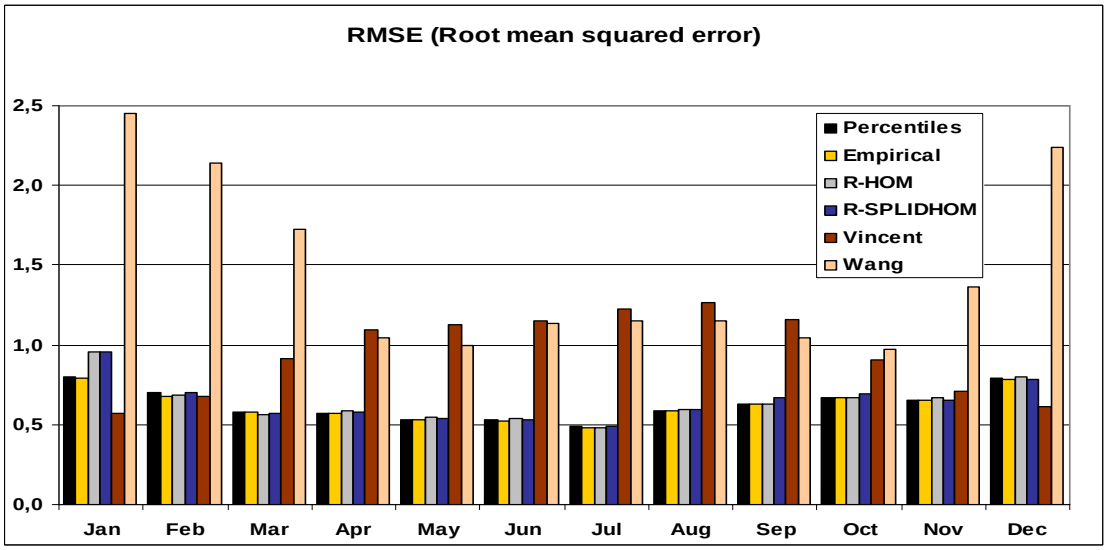

Figure 8. Monthly RMSE for different adjustments (1961-1980)

On the other hand, we run Percentiles and Empirical method in ProClimDB for daily precipitation series (B1NAPA01 – 1961-2007). Remember the original series is inhomogeneous, so you can see different adjustments from break in 1985.

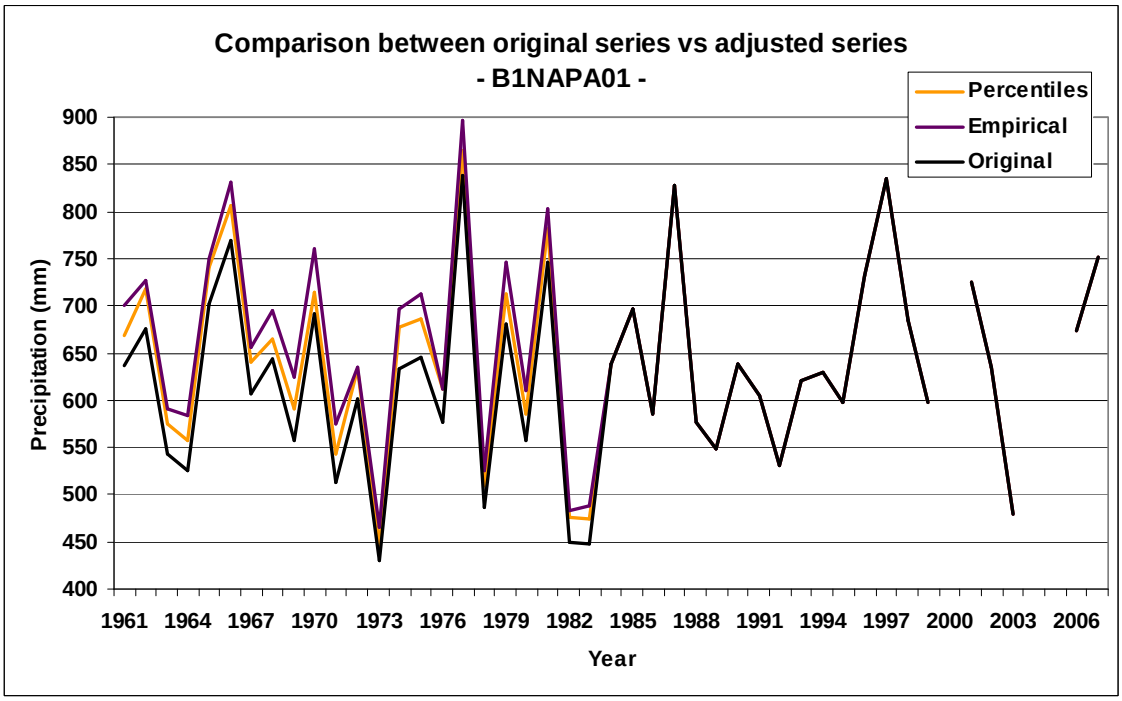

Figure 9. Different adjustments in precipitation series for the period 1961-2007

4. Conclusions

The STSM for two weeks in CHMI in Brno allowed to us to do this study using ProClimDB software. Obviously, it was necessary to introduce us in AnClim and ProClimDB learning to load inputs, showing in dbf/txt, using tools, finding neighbours, using reference series, making adjustments using different methods and other functions (calculations, averages, validations,…). Loading daily temperature (1961-1980) and precipitation (1961-2007) data in ProClimDB of whole Czech Republic, only one series has been selected for creating a known break in 1971 joining another series. Different tests detected the break in new semi-synthetic series and after that, daily temperature series has been homogenized applying six methods (Percentiles, Empirical, R-HOM, R-SPLIDHOM, Vincent's method and Quantile Matching by Wang).

Results showed better adjustments in this kind of break (very big step) applying Percentiles, Empirical, R-HOM and R-SPLIDHOM methods analysing some statistic parameters (AVG anomalies, STD anomalies, RMSE and correlations).

In daily precipitation data one series has been selected knowing a significant break in the middle of the series. Using Percentiles and Empirical methods the inhomogeneity (1985) has been adjusted, so applying the other methods the results were not reliable because precipitation is not a continuous variable.

Finally, the last two days of the STSM the COST benchmark has been finally released (Victor Venema and Jasmina Hadzimustafic) and prepared for running in ProClimDB software.

5. References

Alexandersson, **A**.: 1986: A homogeneity test applied to precipitation data, J. Climatol., 6, 661–675.

Alexandersson, **A**.: 1995: Homogenity testing, multiple breaks and trends, in: Proc. 6th Int. Meeting on Stat. Climatology, Galway, Ireland, 439–441.

Della-Marta, P.M. and H. Wanner, 2006: A Method for homogenising the extremes and mean of daily temperature measurements. *J. of Climate*, **19**, 4179-4197.

Easterling, D. R. and Peterson, **T. C**.: 1995: A new method for detecting undocumented discontinuities in climatological time series, Int. J. Climatol., 15, 369–377.

Gruber, C., Auer, I., Mestre, O. 2008: Homogenization of daily maximum and minimum temperatures – a comparison of methods. 8th Annual Meeting of the EMS/ 7th ECAC.

Mestre, O., Prieur, C., Gruber, C., Caussinus, H., Jourdain, S. A method for homogenization of daily temperature observations. (Submitted).

Potter, K. W.: 1981: Illustration of a New Test for Detecting a Shift in Mean in Precipitation Series, Mon. Weather Rev., 109, 2040– 2045.

Stepanek, P., Zahradnicek P., Skalak, P. 2009: Data quality control and homogenization of air temperature and precipitation series in the area of Czech Republic in the period 1961-2007. Advances in Science and Research, 3, 23-26.

Stepanek, P. 2008: ProClimDB – software for processing climatological datasets, CHMI, regional office Brno, http://www.climahom.eu/ProcData.html

Stepanek, P. 2007: AnClim - software for time series analysis (for Windows). Dept. of Geography, Fac. of Natural Sciences, Masaryk University, Brno. 1.47 MB.

Vincent, L. A., X. Zhang, B.R Bonsal, and W.D. Hogg, 2002: Homogenization of daily temperatures over Canada. *J. Climate*, **15**, 1322-1334.

Wang, X. L., 2010: A quantile matching adjustment algorithm for Gaussian data series (submitted).

Wang, X.L., H. Chen, Y. Wu and Q. Pu, 2009: New techniques for detection and adjustment of shifts in daily precipitation data series. J. Appl. Meteor. Climatol. (submitted).