Experiences with homogenization of daily and monthly series of air temperature, precipitation and relative humidity in the Czech Republic, 1961-2007

P. Štěpánek[1], P. Zahradníček[1]

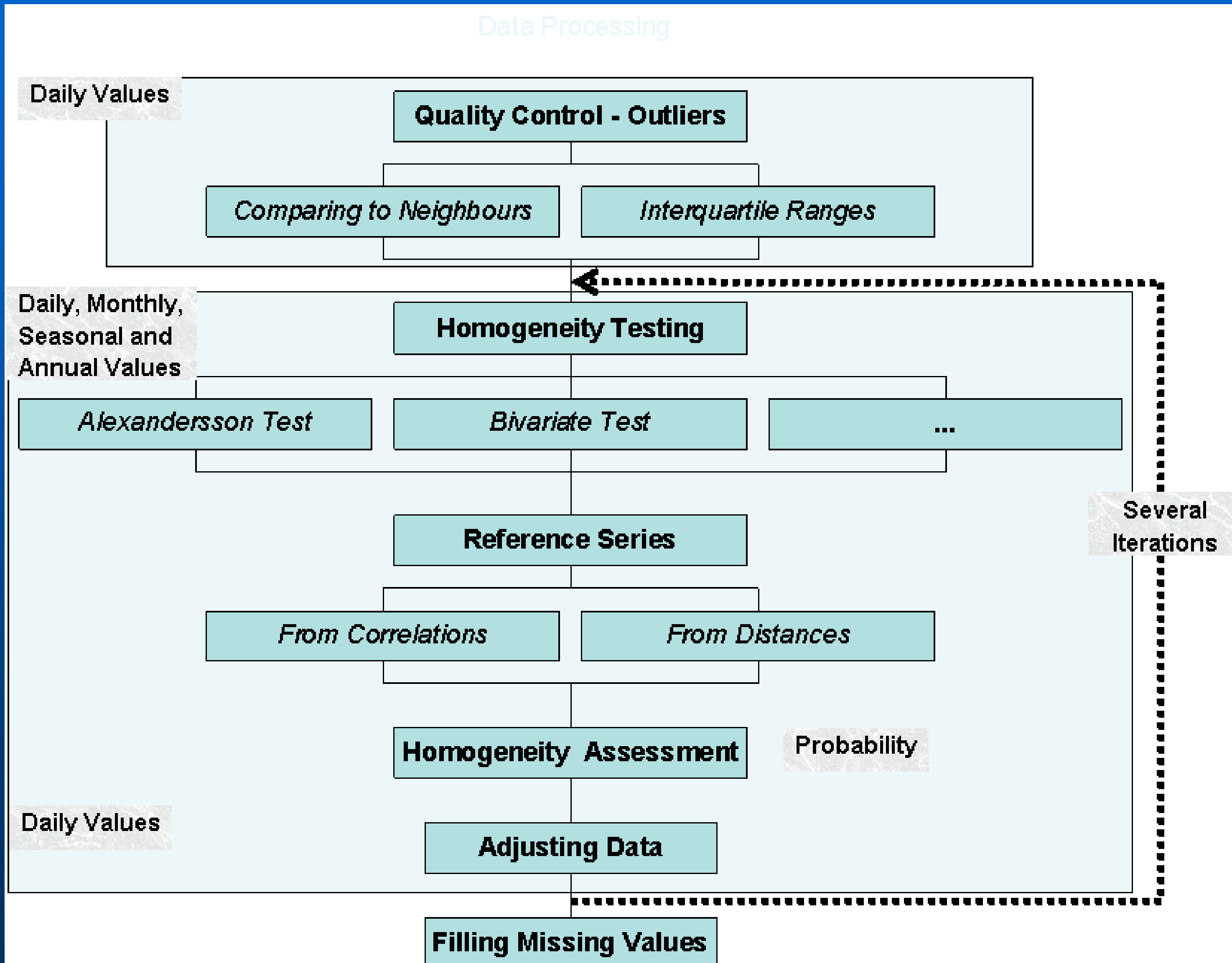
[1] Czech Hydrometeorological Institute, Regional Office Brno, Czech Republic

E-mail: petr.stepanek@chmi.cz

COST-ESO601 meeting and

Sixth Seminar for Homogenization and Quality Control in Climatological Databases

Processing before any data analysis

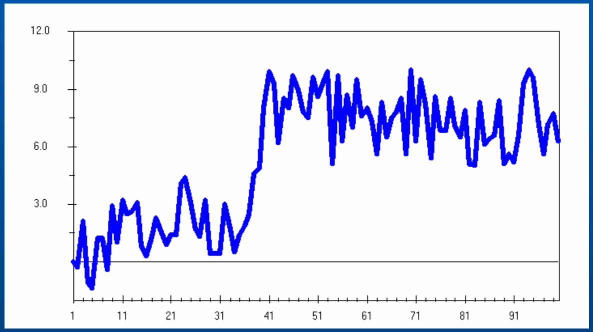


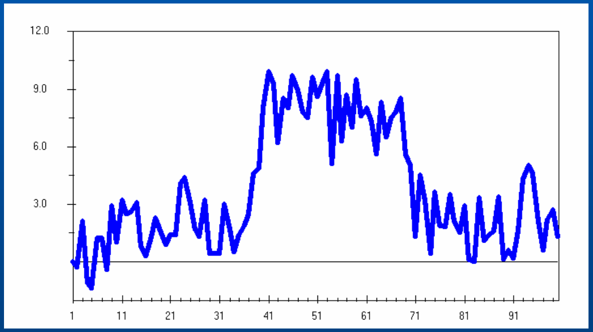
Software AnClim, ProClimDB

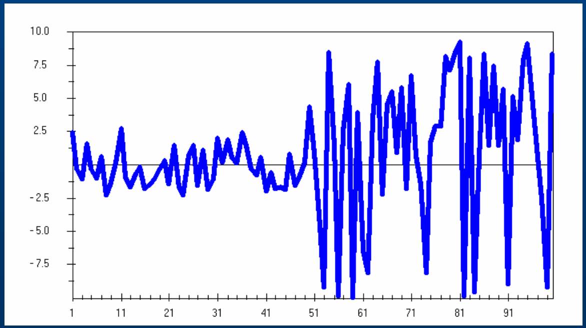
Homogenization

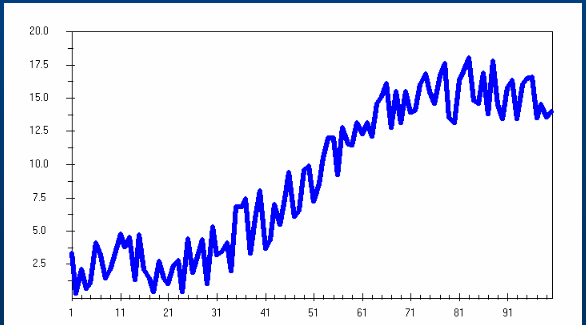
- Change of measuring conditions

→ inhomogeneities





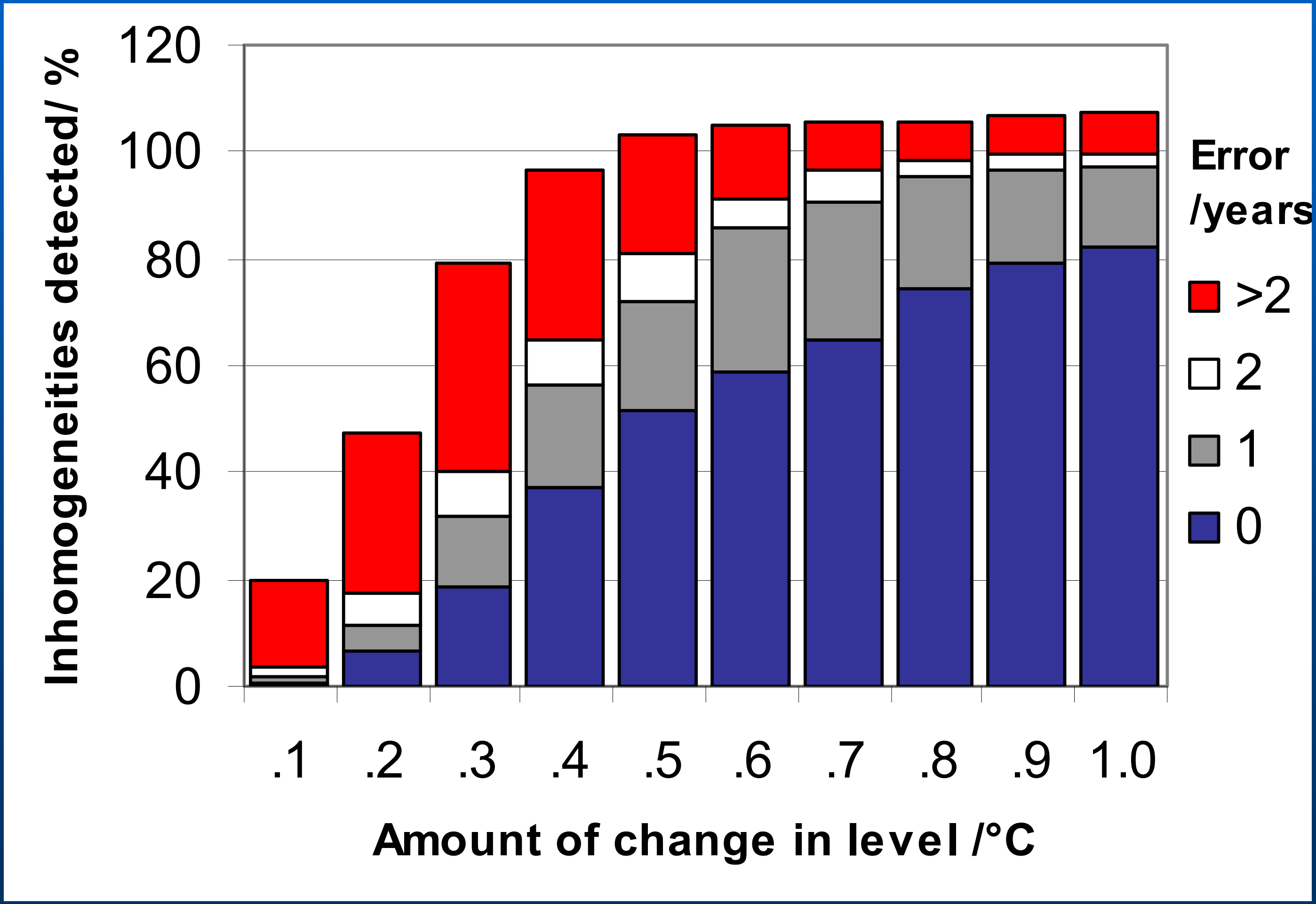




Reliability of Detecting Inhomogeneities by statistical tests (case study)

- **generated series of random numbers (properties of air temperature series for year, summer and winter, CZ)**
- **introduced steps with various amount of change in level**
- **various position of the steps**
- **various lengths of the series**
- 950 series, $p=0.05$

Detecting Inhomogeneities by SNHT (p=0.05, 950 series)



Assessing Homogeneity - Problems

- **most of metadata incomplete**

→ **we depend upon statistical tests results**

Assessing Homogeneity - Problems

- most of metadata incomplete

→ we depend upon statistical tests results

- uncertainty in test results
  - right inhomogeneity detection is problematic
  (for smaller amount of change)

Proposed solution

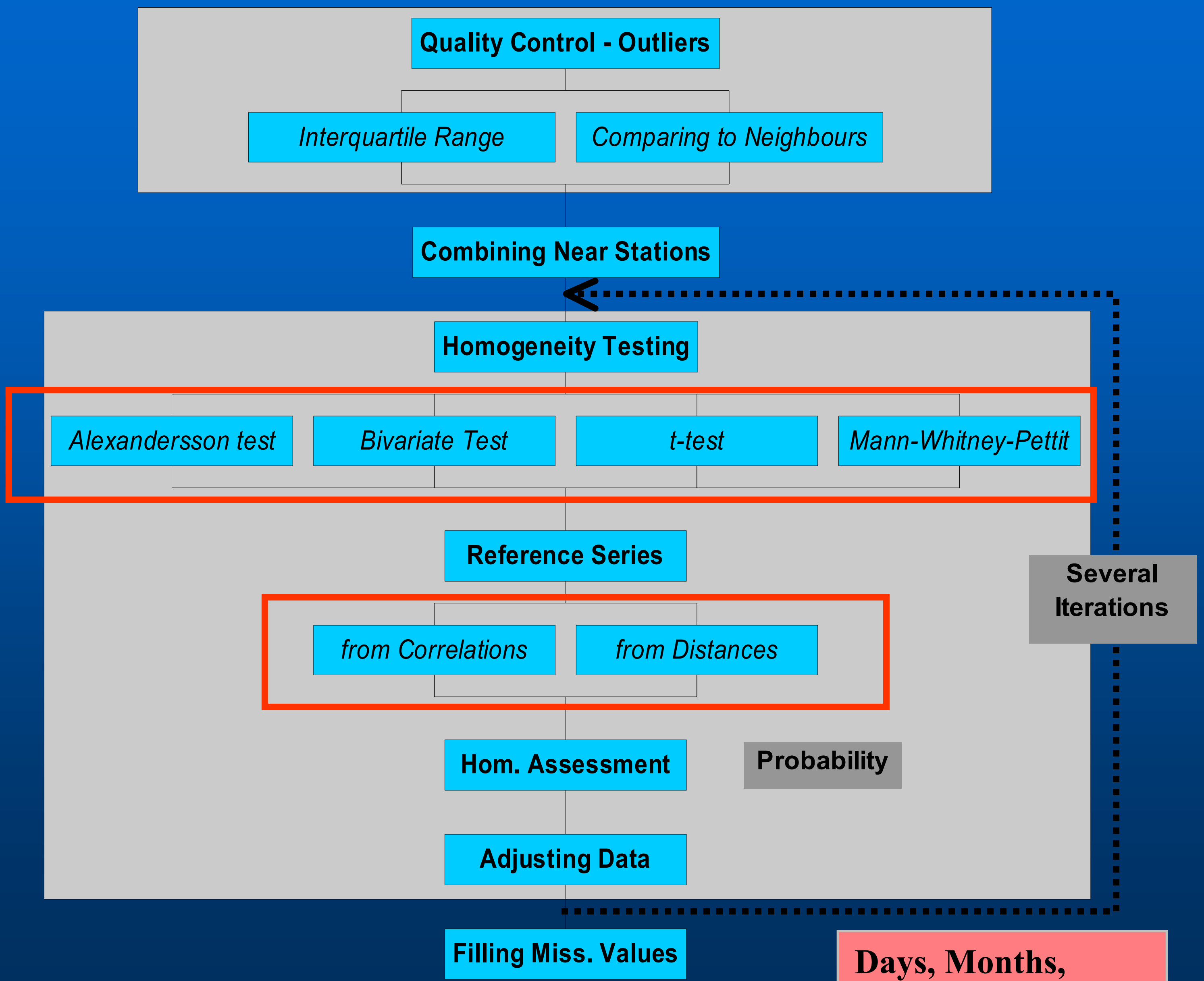
- To get as many test results for each candidate series as possible

→ **„Ensemble“ approach - processing of big amount of test results for each individual series**

Adventages of the „Ensemble“ approach

- we know relevance (probability) of each inhomogeneity
- we can easily assess quality of measurements for series as a whole

How to increase number of test results



Days, Months, seasons, year

Creating Reference Series

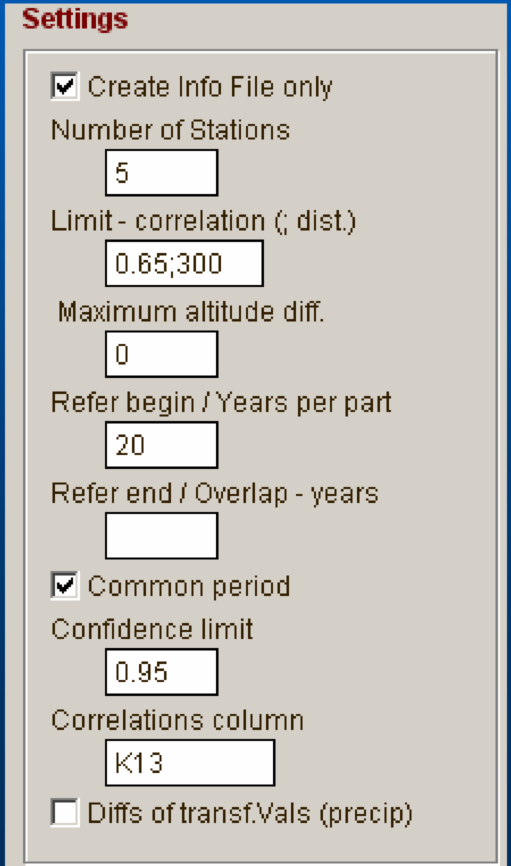
- for monthly, daily data (each month individually)
- weighted/unweighted mean from neighbouring stations
- criterions used for stations selection (or combination of it):
  - best correlated / nearest neighbours

    (correlations – from the first differenced series)
  - limit correlation, limit distance
  - limit difference in altitudes

- neighbouring stations series should be standardized to test series
AVG and / or STD

  (temperature - elevation, precipitation - variance)

  - **missing data are not so big problem then**



Relative homogeneity testing

- **Available tests:**
  - Alexandersson SNHT
  - Bivariate test of Maronna and Yohai
  - Mann – Whitney – Pettit test
  - t-test
  - Easterling and Peterson test
  - Vincent method
  - …

20 year parts of the daily series (40 for monthly series with 10 years overlap),
in SNHT splitting into subperiods in position of detected significant changepoint

(30-40 years per one inhomogeneity)

Homogeneity assessment

Output example: Station Čáslav, 3rd segment, 1911-1950, n=40

Test	Ref	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	Win	Spr	Sum	Aut	Year
A	**avg**	1927	1929	1927	1927	1927	1928	1927	1926	1926	1926	1926	1926	1927	1927	1927	1926	1927
A			1930															
A	**corr**	1927	1927	1927	1927	1927	1928	1927	1926	1926	1926	1926	1926	1927	1927	1927	1926	1927
A				1939		1938	1939	1940	1922						1937	1937		1935
A	**dist**	1927	1928	1927	1927	1927	1928	1927	1926	1926	1926	1926	1926	1927	1927	1927	1926	1927
A			1930								1940							1918
B	**avg**	1927	1928	1927	1927	1927	1928	1927	1926	1926	1926	1926	1926	1927	1927	1927	1926	1927
B									1922									
B	**corr**	1927	1927	1927	1927	1927	1928	1927	1926	1926	1926	1926	1926	1927	1927	1927	1926	1927
B				1936		1938	1939	1944	1922					1935	1937	1937		1935
B									1937									
B	**dist**	1927	1928	1927	1927	1927	1928	1927	1926	1926	1926	1926	1926	1927	1927	1927	1926	1927
B		1930									1940			1931			1913	1918
V	**corr**													1927			1926	
V															1937	1922		1935
V																1937		
V	**dist**													1927	1927	1927		
V																		1918

Homogeneity assessment, Output II example:

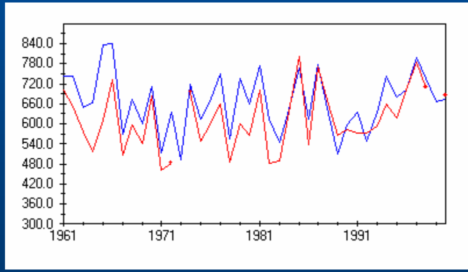
Begin	End	Length	InHomogeneity	Number	% detected inhom	% possible inhom	End	Missing
1911	1950	40		140	100	120		
			1927	60	43	51		
			1926	37	26	32		
			1928	9	6	8		4
			1937	7	5	6		
			1922	4	3	3		
			1935	4	3	3		
			1918	3	2	3		
			1930	3	2	3		
			1939	3	2	3		
			1940	3	2	3		2
			1938	2	1	2		
			1913	1	1	1	3	3
			1929	1	1	1		
			1931	1	1	1		
			1936	1	1	1		
			1944	1	1	1		
1926	1927	2		97	69	83		
1926	1931	6		111	79	95		
1935	1940	6		20	14	17		
1911	1920	10		4	3	3		
1921	1930	10		114	81	97		
1931	1940	10		21	15	18		
1941	1950	10		1	1	1		

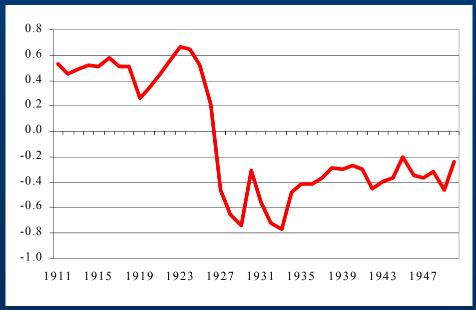
Summed numbers of detections for individual years

Homogeneity assessment

- **combining several outputs (sums of detections in individual years, metadata, graphs of differences/ratios, …)**

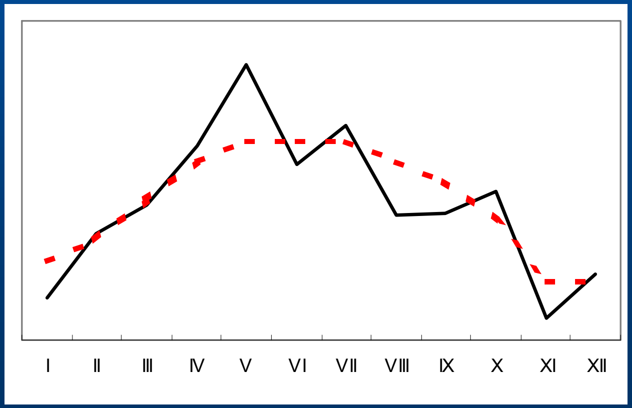
	ID	EL	YEAR	BEGI	END	YEAR_COUN	Y_POSSIBL	YEA	MIS	X_BEGIN_D	X_END_DA	X	X	L	L	A	B	REMARK	C	C
x	B1BOJK01	x	1985			41	14.24		12	23.3.1984	31.3.2003	#	#				E	change		
	B1BOJK01	x	1985			41	14.24		12	23.3.1984	31.12.9999	#	#					obs	V	B
	B1BYSH01	x	1978			37	12.85													
?	B1BYSH01	x	1979			33	11.46													
?	B1BYSH01	x	1980			43	14.93													
?	B1HLHO01	x	1965			31	10.76	4	1											
	B1HOLE01	x	1976			33	11.46													
	B1KROM01	x		1977	1978	31	10.76													
x	B1RADE01	x	1994			44	15.28		2	1.1.1994	31.12.9999	#	#				F	change		
	B1RADE01	x	1994			44	15.28		2	1.1.1994	31.12.9999	#	#					obs	J	B
x	B1RYCH01	x	1973			49	17.01			1.5.1973	28.2.1991	#	#				V	change		
	B1RYCH01	x	1973			49	17.01			1.9.1972	28.2.1991	#	#					obs	M	B
xx?	B1STRZ01	x	1987			53	18.40													
	B1STRZ01	x	1988			30	10.42													
	B1UHBR01	x	1983			31	10.76			18.2.1984	31.1.1999	#	#				U	change		
	B1UHBR01	x	1983			31	10.76			18.2.1984	12.5.1993	#	#					obs	J	B
x	B1UHBR01	x	1984			77	26.74			18.2.1984	31.1.1999	#	#				U	change		
	B1UHBR01	x	1984			77	26.74			18.2.1984	12.5.1993	#	#					obs	J	B
	B1VELI01	x	1978			31	10.76													
?	B1VELI01	x		1977	1978	44	15.28													
?	B1VKLO01	x	1984			29	10.07													
x	B1VYSK01	x	1999			32	11.11	-1		1.4.1998	31.12.9999	#	#				V	change		
	B1VYSK01	x	1999			32	11.11	-1		1.4.1998	31.12.9999	#	#					obs	V	B
	B2BOSK01_	x	1968			33	11.46													
	B2BREC01	x	1968			35	12.15													
	B2BRUM01	x	1989			51	17.71			1.2.1989	31.3.1994	#	#				E	change		
	B2BRUM01	x	1989			51	17.71			1.2.1989	31.3.1994	#	#					obs	M	B





Adjusting monthly data

- using reference series based on correlations
- adjustment: from differences/ratios 20 years before and after a change, monhtly
- smoothing monthly adjustments (low-pass filter for adjacent values)



Example:

Adjusting values - evaluation

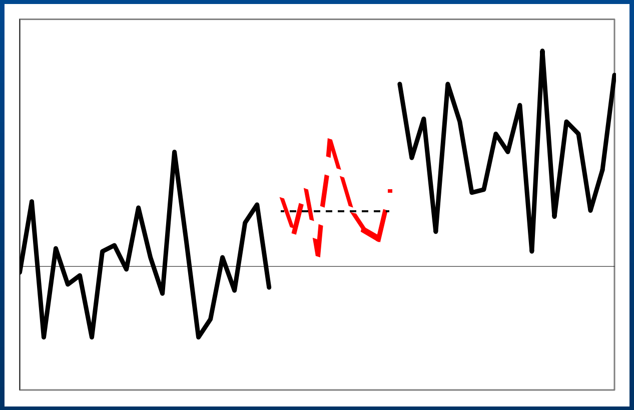
ID_1	p	II	BEGIN	END	YEAR	MONTH	REMARK	C(	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	K11	K12
B1RYCH01		E	1961	1992	1973	5	ADJust		1.135	1.197	1.155	1.333	1.149	1.070	1.088	1.354	1.145	1.116	1.136	1.265
B1RYCH01							DIFF1		0.905	0.875	0.912	0.813	0.906	0.956	0.896	0.786	0.912	0.956	0.908	0.855
B1RYCH01							DIFF2		1.027	1.048	1.053	1.084	1.041	1.024	0.975	1.064	1.045	1.067	1.032	1.081
B1RYCH01							corr		0.964	0.930	0.963	0.915	0.888	0.870	0.866	0.927	0.961	0.952	0.956	0.875
B1RYCH01							corr+		0.007	0.017	0.006	0.026	0.014	0.006	0.008	-0.001	-0.002	0.017	0.010	0.033
B1RYCH01							t		1.904	2.144	2.443	3.897	1.957	0.936	0.874	3.424	1.937	1.507	2.252	3.415
B1RYCH01							t_crit		2.042	2.048	2.045	2.045	2.045	2.045	2.042	2.042	2.042	2.042	2.042	2.045
B1RYCH01							Std_1		0.171	0.184	0.108	0.216	0.206	0.168	0.274	0.146	0.241	0.255	0.139	0.159
B1RYCH01							Std_2		0.178	0.235	0.181	0.169	0.175	0.209	0.232	0.256	0.146	0.164	0.157	0.185
B1RYCH01							t2		1.923	2.252	2.730	3.685	1.884	0.985	0.837	3.904	1.718	1.351	2.325	3.569
B1RYCH01							t2_crit		1.960	1.961	1.960	1.961	1.961	1.960	1.961	1.960	1.961	1.961	1.960	1.960
B1RYCH01							No_1		12	12	12	12	12	12	12	12	12	12	12	11
B1RYCH01							No_2		20	18	19	19	19	19	20	20	20	20	20	20
B1RYCH01							b1_1		-0.015	-0.016	0.002	0.017	0.028	0.002	-0.035	0.002	0.035	0.040	0.015	-0.012
B1RYCH01							b1_2		-0.007	-0.024	-0.002	0.001	-0.008	0.018	-0.022	-0.002	-0.007	-0.016	-0.014	-0.024
B1RYCH01		**> 2n:0.479,0.233**			**1973**	**5**	**ADJ_sm**		**1.180**	**1.178**	**1.206**	**1.238**	**1.172**	**1.107**	**1.149**	**1.229**	**1.185**	**1.138**	**1.162**	**1.199**
B1RYCH01							corr		0.964	0.930	0.963	0.915	0.888	0.870	0.866	0.927	0.961	0.952	0.956	0.875
B1RYCH01							corr+(AD		0.007	0.016	0.003	0.026	0.014	0.006	0.009	0.010	-0.005	0.019	0.009	0.030

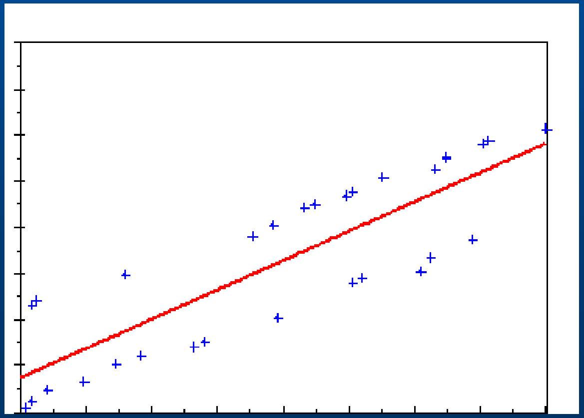
Iterative homogeneity testing

- **several iteration of testing and results evaluation**
  - **several iterations of homogeneity testing and series adjusting (3 iterations should be sufficient)**
  - **question of homogeneity of reference series is thus solved:**
    - **possible inhomogeneities should be eliminated by using averages of several neighbouring stations**
    - **if this is not true: in next iteration neighbours should be already homogenized**

Filling missing values

- **Before homogenization: influence on right inhomogeneity detection**
- **After homogenization: more precise - data are not influenced by possible shifts in the series**



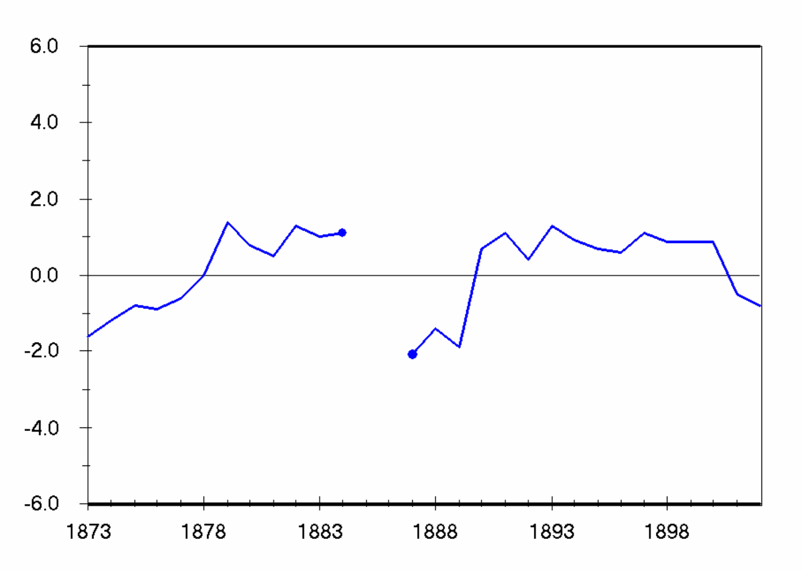


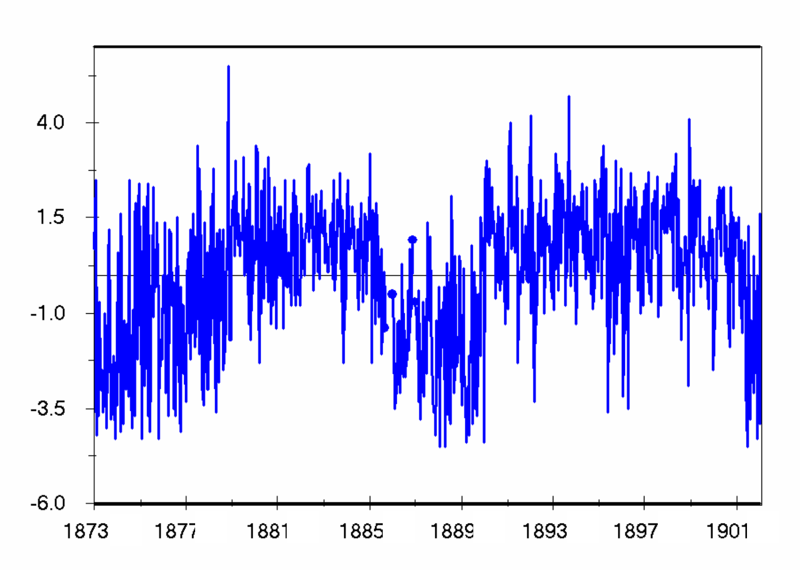
Dependence of tested series on reference series

Using daily data for inhomogeneities detection

- **Additional information to monthly, seasonal and annual values testing**
- **Advantageous in case of breaks appears near ends of series**
- **Missing values – no such influence like in case of monthly data**
- **Problems (normal distribution or autocorellations) but can be handled to some extend**
- **Correlation coefficients** (tested versus reference series) **are slightly lower** (compared to monthly data), **but still high enough** (around 0.9 even in case precipitation)

Using daily data for inhomogeniety detection





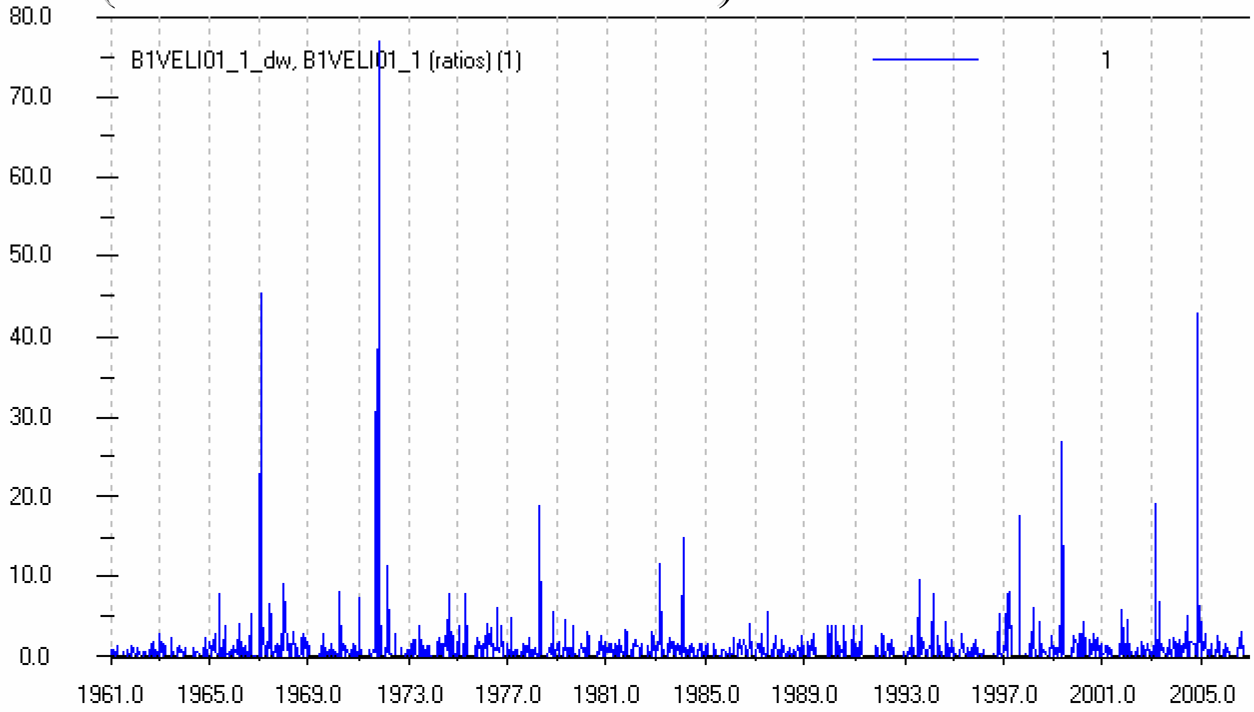
Homogenization of daily values – **precipitation** series

- **working with individual monthly values** **(to get rid of annual cycle)**
- **It is still needed to adapt data to approximate to normal distribution**
- **One of the possibilities: consider values above 0.1 mm only**
- **Additional transformation of series of ratios (e.g. with square root)**

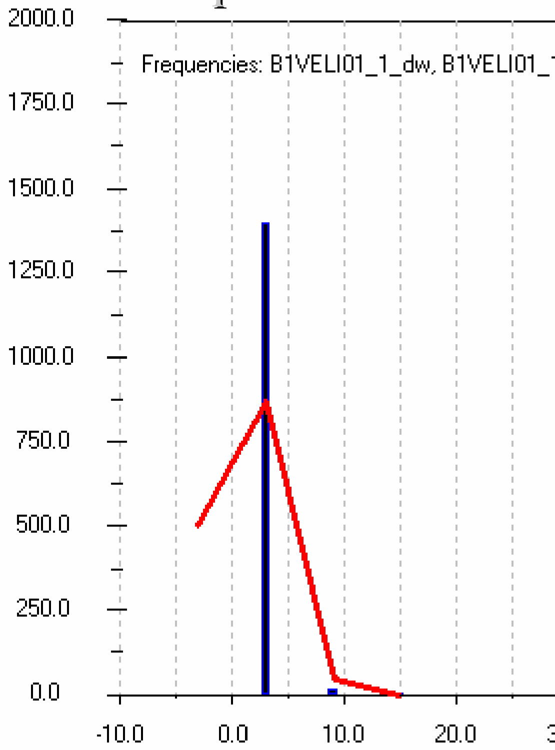
Homogenization of precipitation – daily values

Original values - far from normal distribution

(ratios tested/reference series)



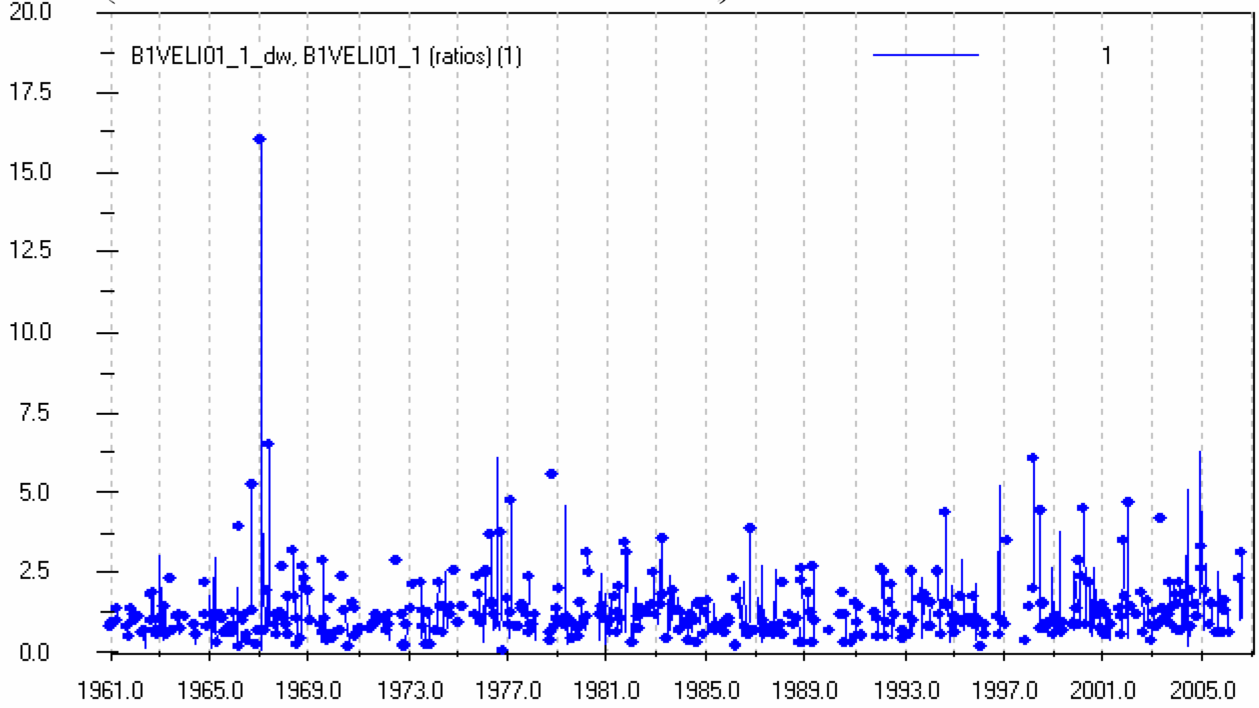
Frequencies



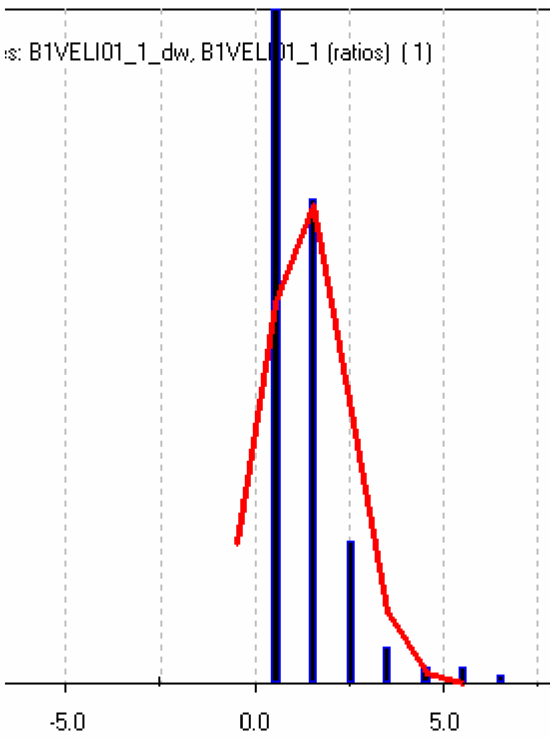
Homogenization of precipitation – daily values

- **Limit value 0.1 mm**

(ratios tested/reference series)



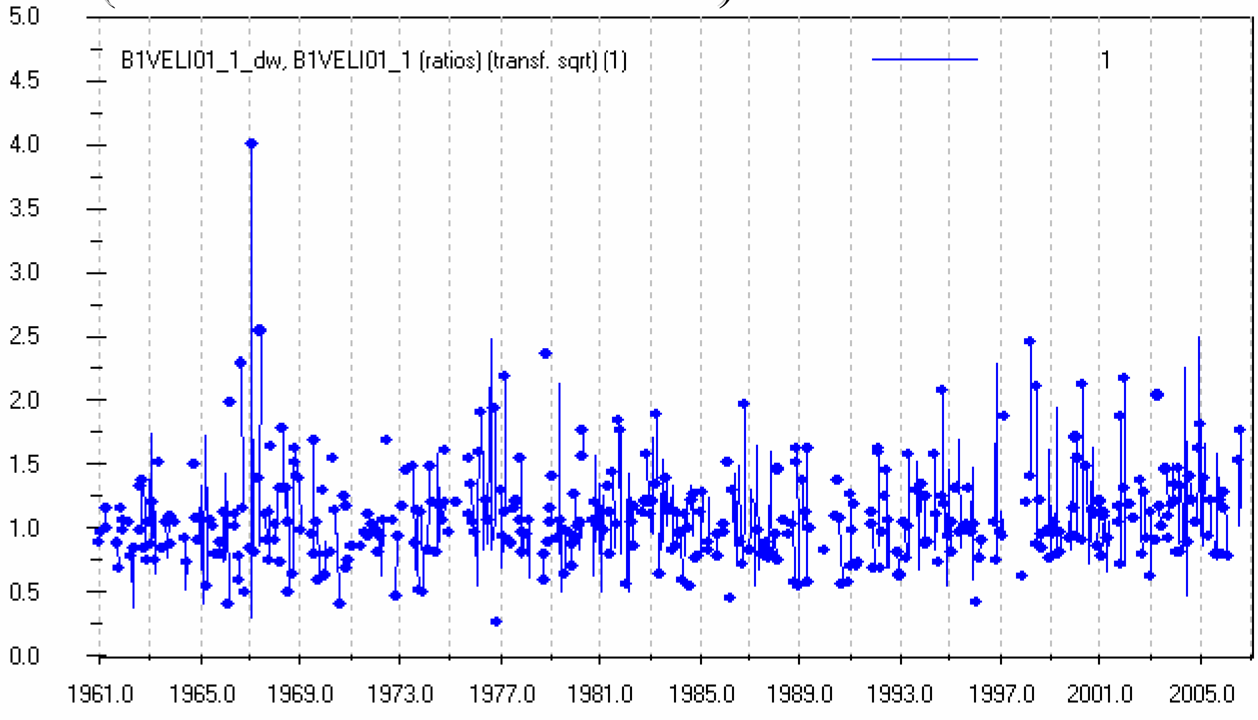
Frequencies



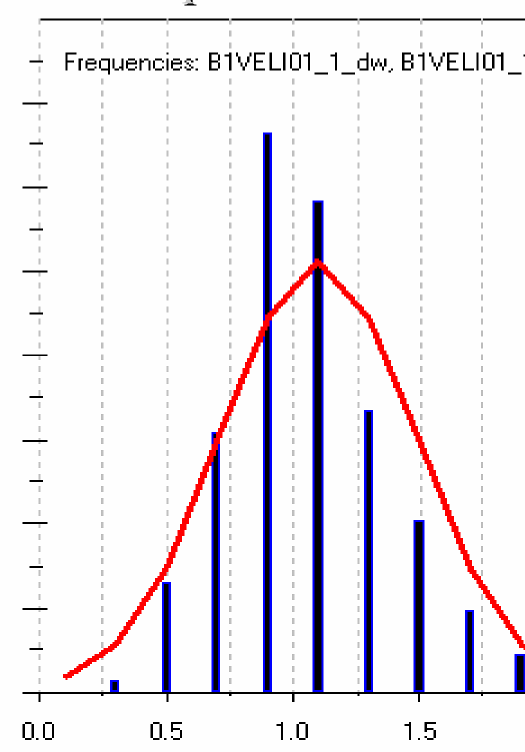
Homogenization of precipitation – daily values

- **Limit value 0.1 mm, square root transformation** (of ratios)

(ratios tested/reference series)

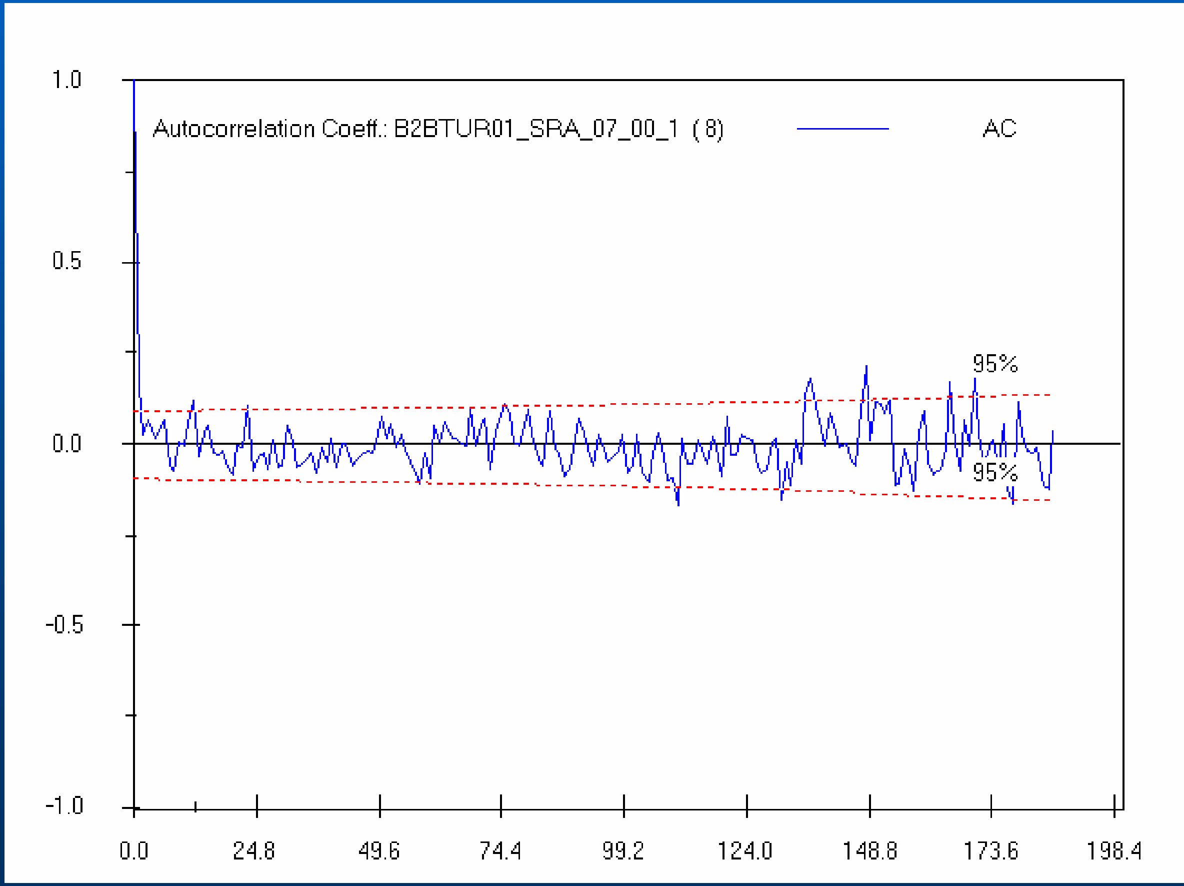


Frequencies



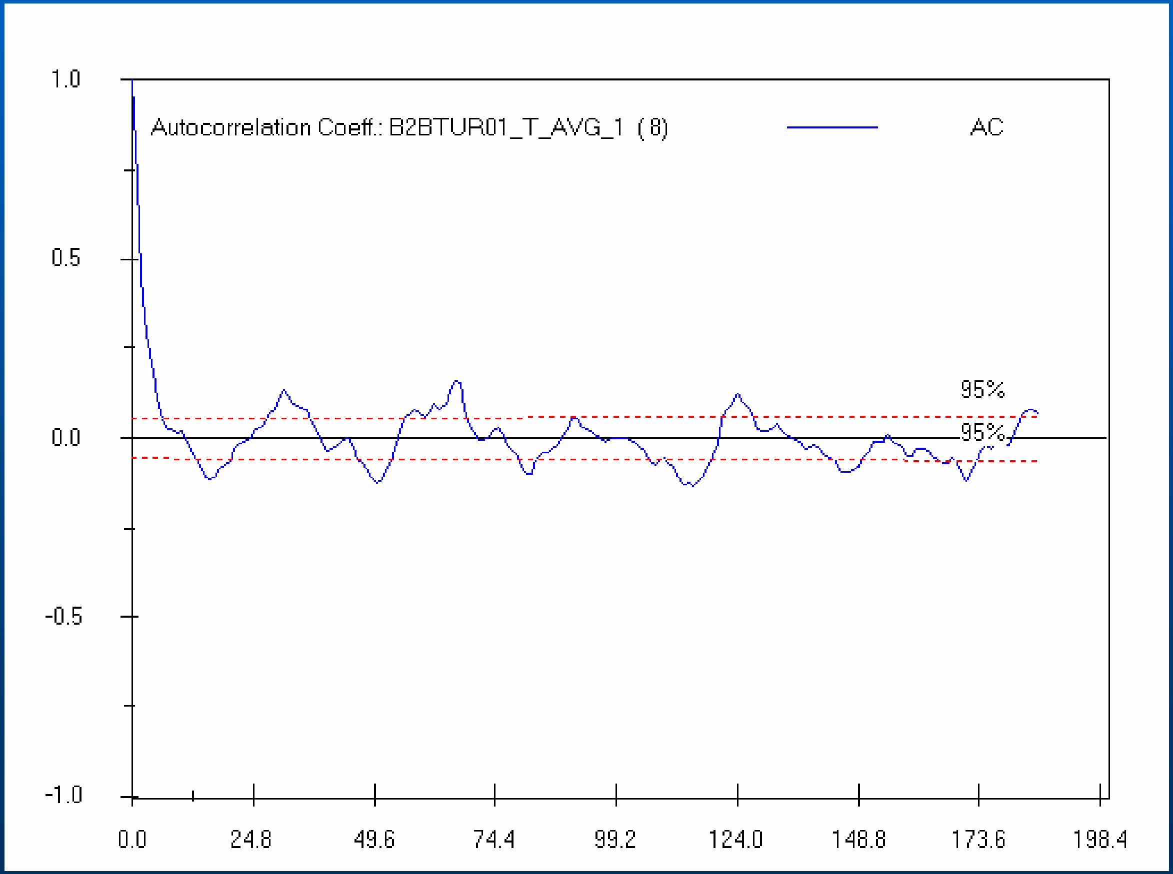
Problem of independence, **Precipitation** above 1 mm

- **August, Autocorrelations**



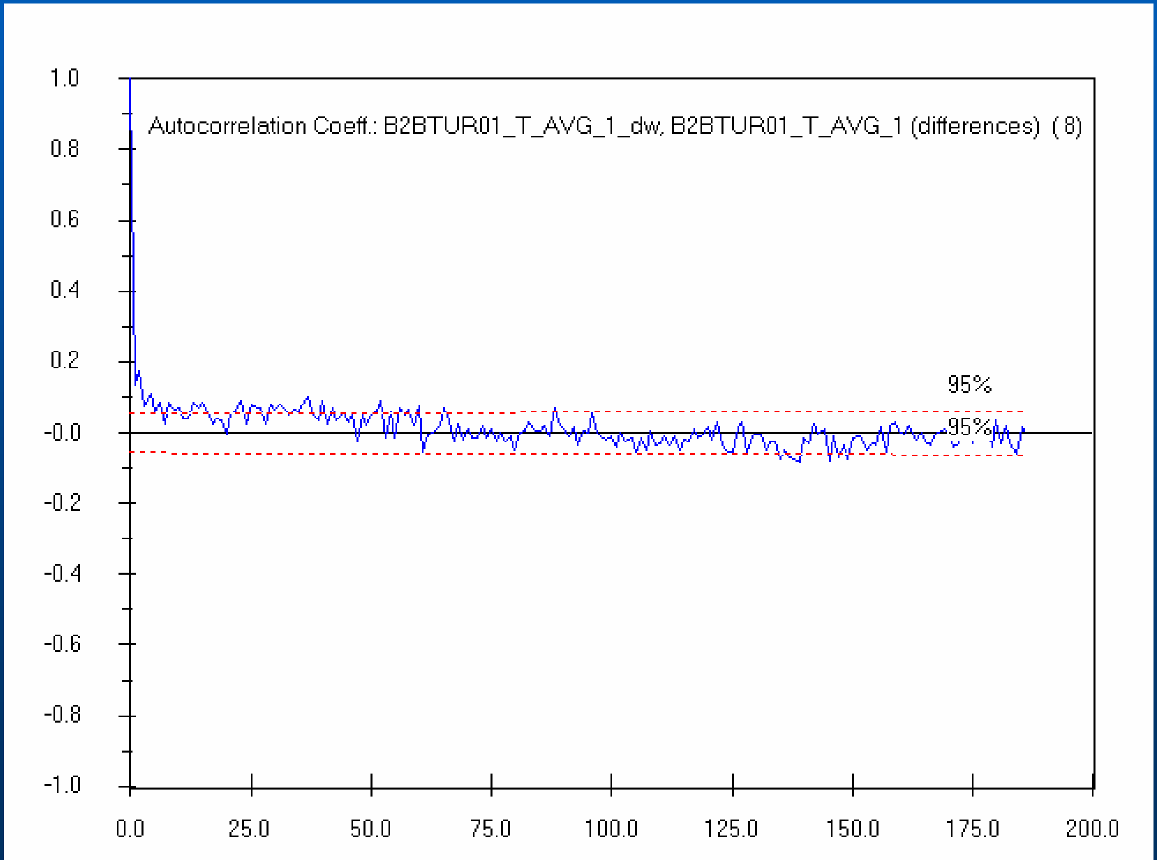
Problem of independece, Temperature

- **August, Autocorrelations**

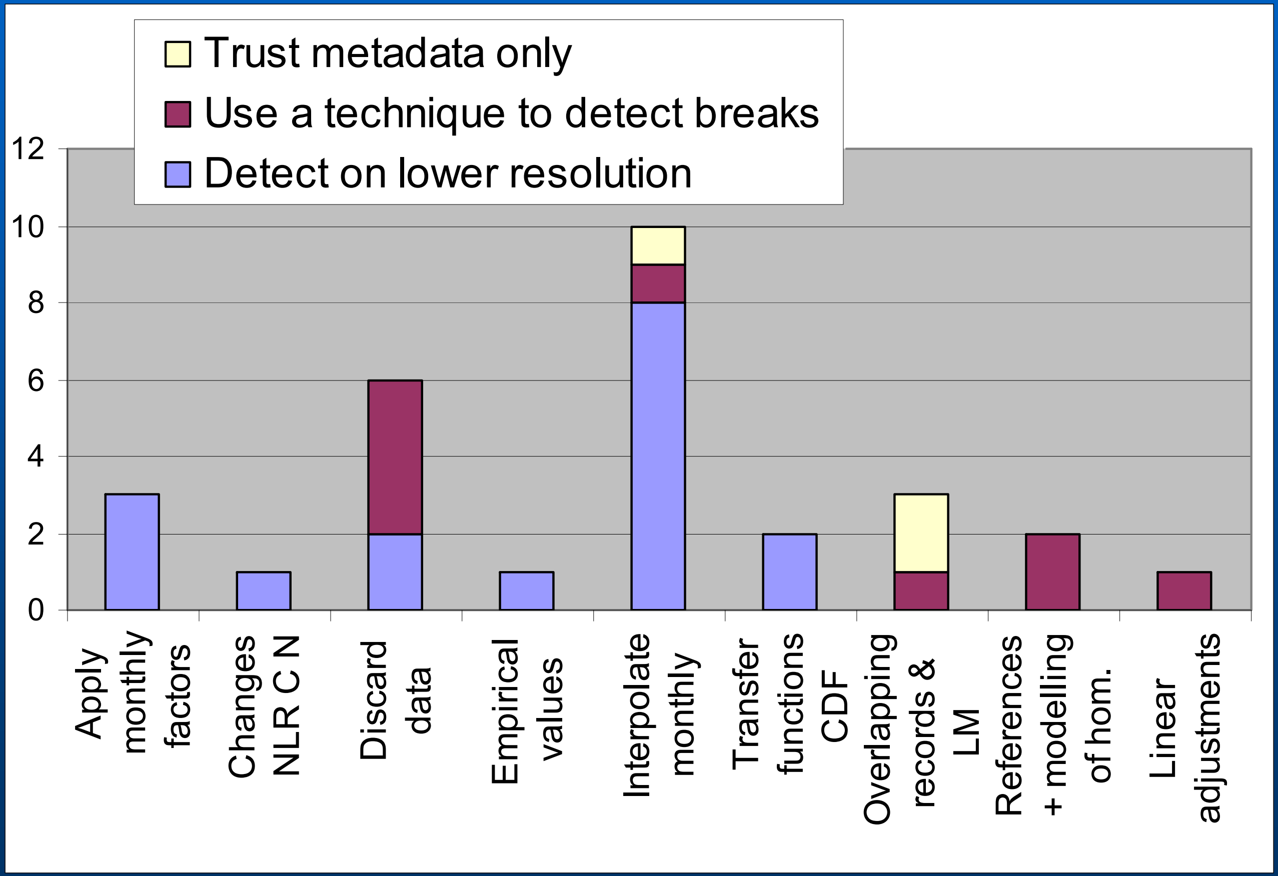


Problem of independece, Temperature differences

- **August, Autocorrelations**



WP1 SURVEY (Enric Aguilar)
Daily data - <u>Correction</u> (WP4)



- **Very few approaches actually calculate special corrections for daily data.**
- **Most approaches either**
  - **Do nothing (discard data)**
  - **Apply monthly factors**
  - **Interpolate monthly factors**
- **The survey points out several other alternatives that WG5 needs to investigate**

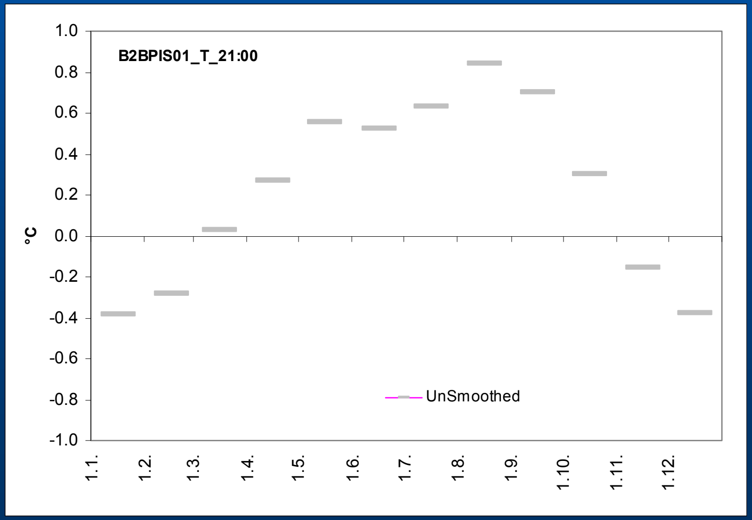
WG1 PROPOSAL TO WG4.
Methods

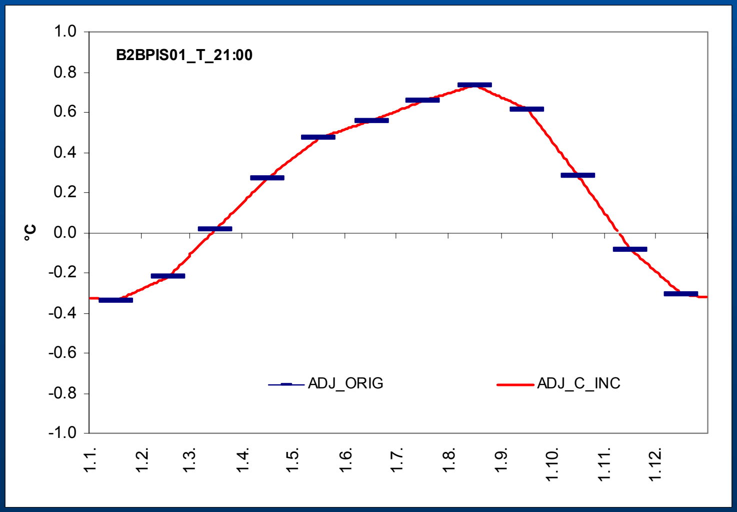
- **Interpolation of monthly factors**
  - **MASH**
  - **Vincent *et al* (2002)**
- **Nearest neighbour resampling models, by Brandsma and Können (2006)**
- **Higher Order Moments (HOM), by Della Marta and Wanner (2006)**
- **Two phase non-linear regression (Mestre)**

Adjusting daily values for inhomogeneities, from **<u>monthly</u>** versus **<u>daily</u>** adjustments („delta" method)

Adjusting from monthly data

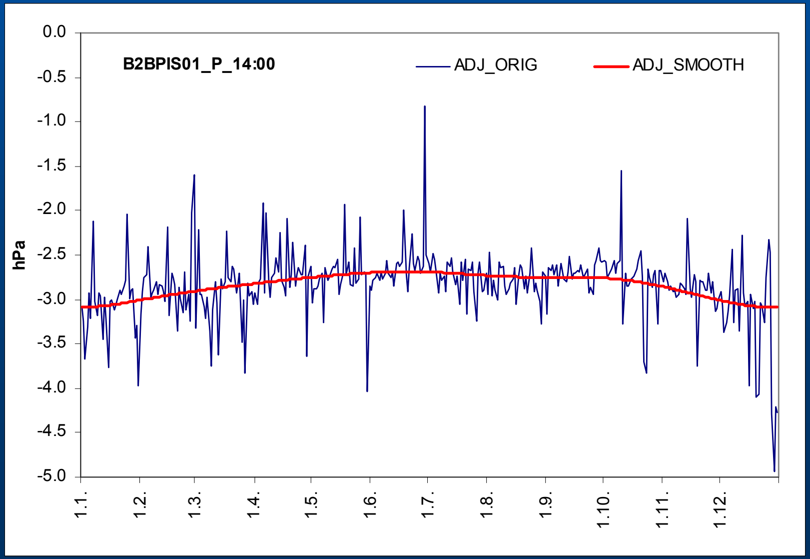
- monthly adjustments smoothed with Gaussian low pass filter (weights approximately 1:2:1)
- smoothed monthly adjustments are then evenly distributed among individual days





Adjusting straight from **daily** data

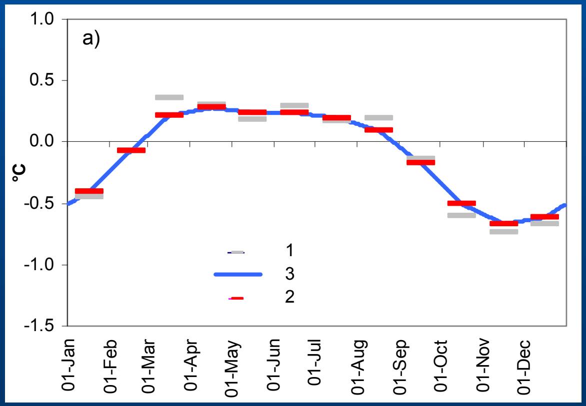
- **Adjustment estimated for each individual day (series of 1st Jan, 2nd Jan etc.)**
- **Daily adjustments smoothed with Gaussian low pass filter for 90 days (annual cycle 3 times to solve margin values)**

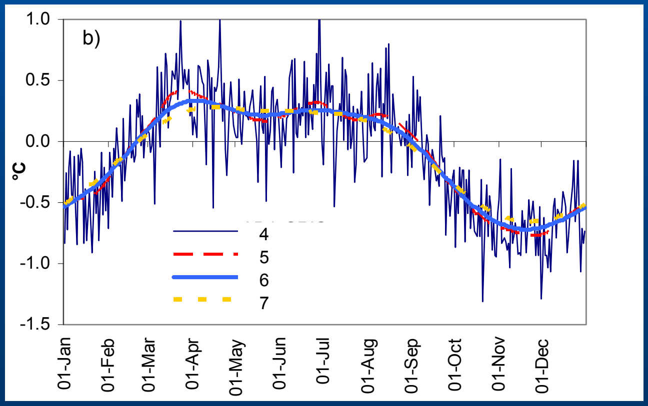


Adjustments (Delta method)

- The same final adjustments may be obtained from either monthly averages or through direct use of daily data

(for the daily-values-based approach, it seems reasonable to smooth with a low-pass filter for 60 days. The same results may be derived using a low-pass filter for two months (weights approximately 1:2:1) and subsequently distributing the smoothed monthly adjustments into daily values)



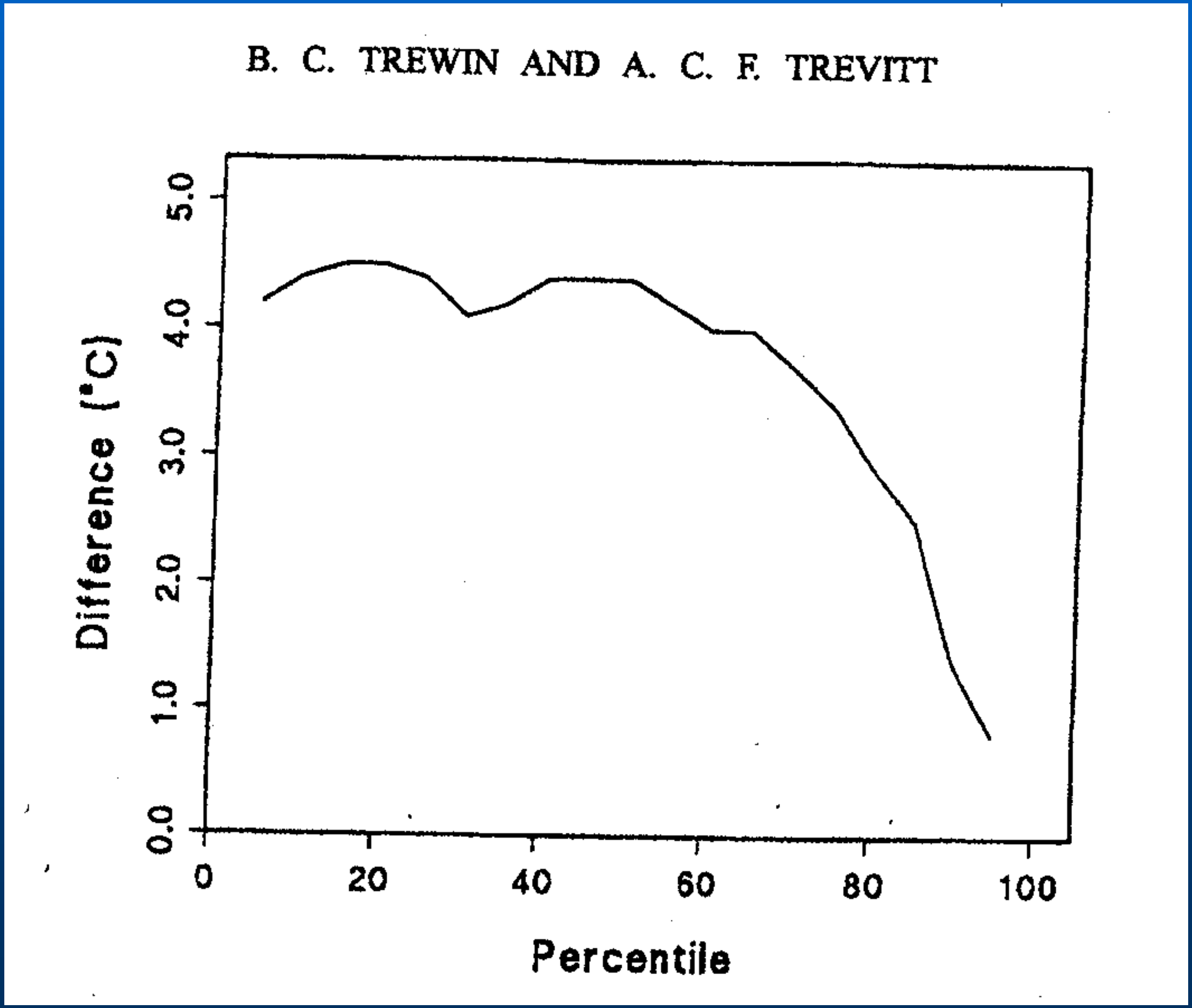


(1 – raw adjustments, 2 – smoothed adjustments, 3 – smoothed adjustments distributed into individual days), b) daily-based approach (4 – individual calendar day adjustments, 5 – daily adjustments smoothed by low-pass filter for 30 days, 6 – for 60 days, 7 – for 90 days)

Variable correction

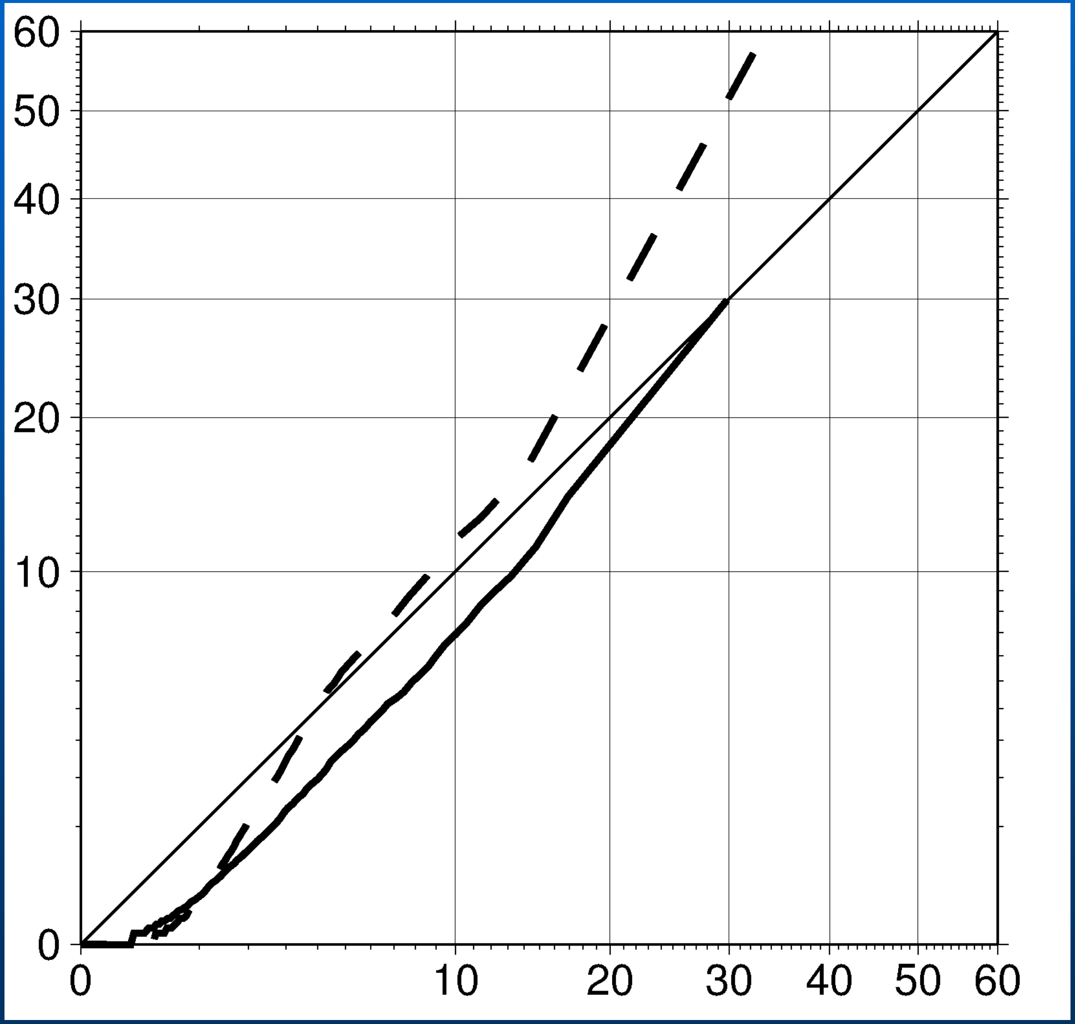
- f(C(d)|R), function build with the reference dataset R, d – daily data
- cdf, and thus the pdf of the adjusted candidate series C*(d) is exactly the same as the cdf or pdf of the original candidate series C(d)

Variable correction



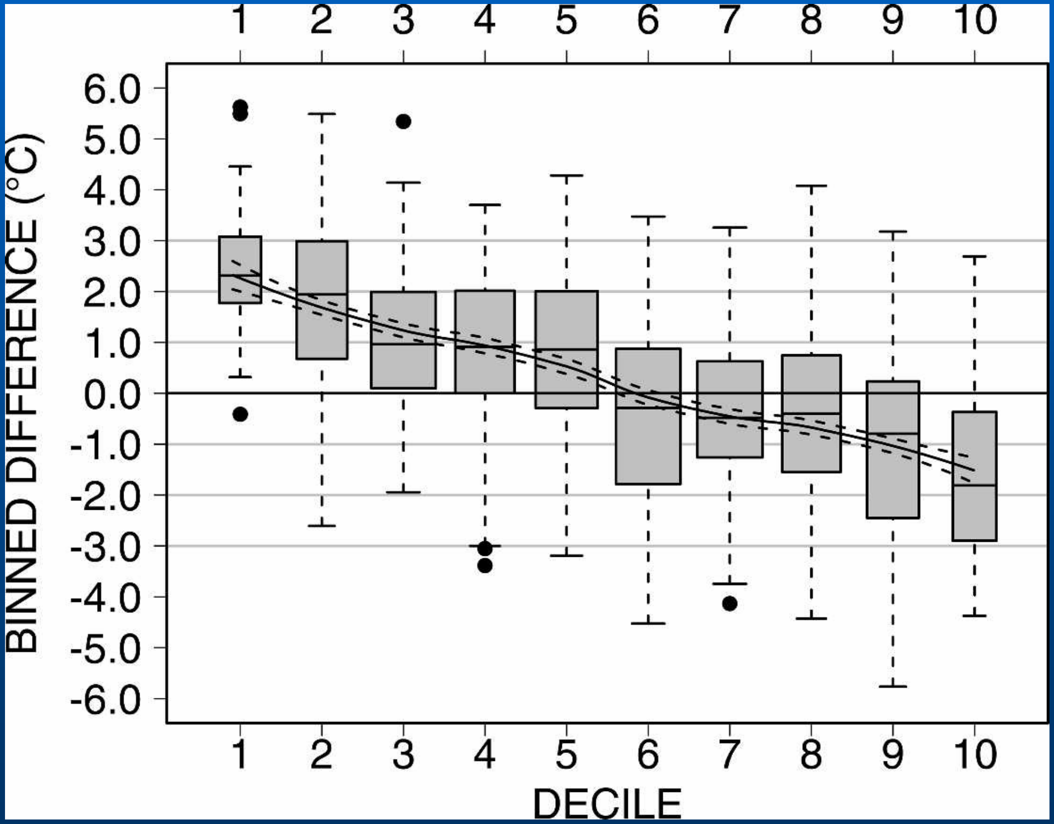
1996

Variable correction, **q-q function**



Michel Déqué, Global and Planetary Change 57 (2007) 16–26

Variable correction,
The higher-order moments method



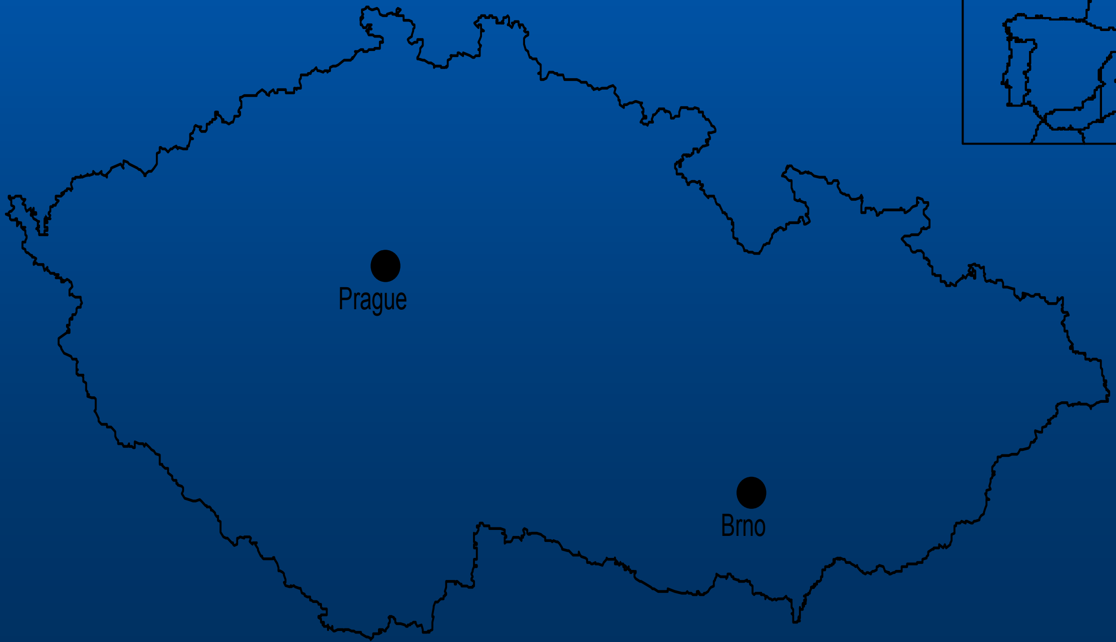
DELLA-MARTA AND WANNER, JOURNAL OF CLIMATE 19 (2006) 4179-4197

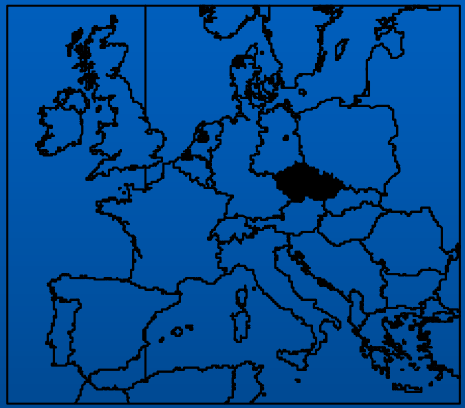
Remarks
Homogenization without metadata –
recommendations how to increase its confidence

- Daily, monthly, seasonal, annual data
- Various reference series
- Various statistical tests
- 40 year periods (20 for daily data), some overlap
- Several steps - iterations

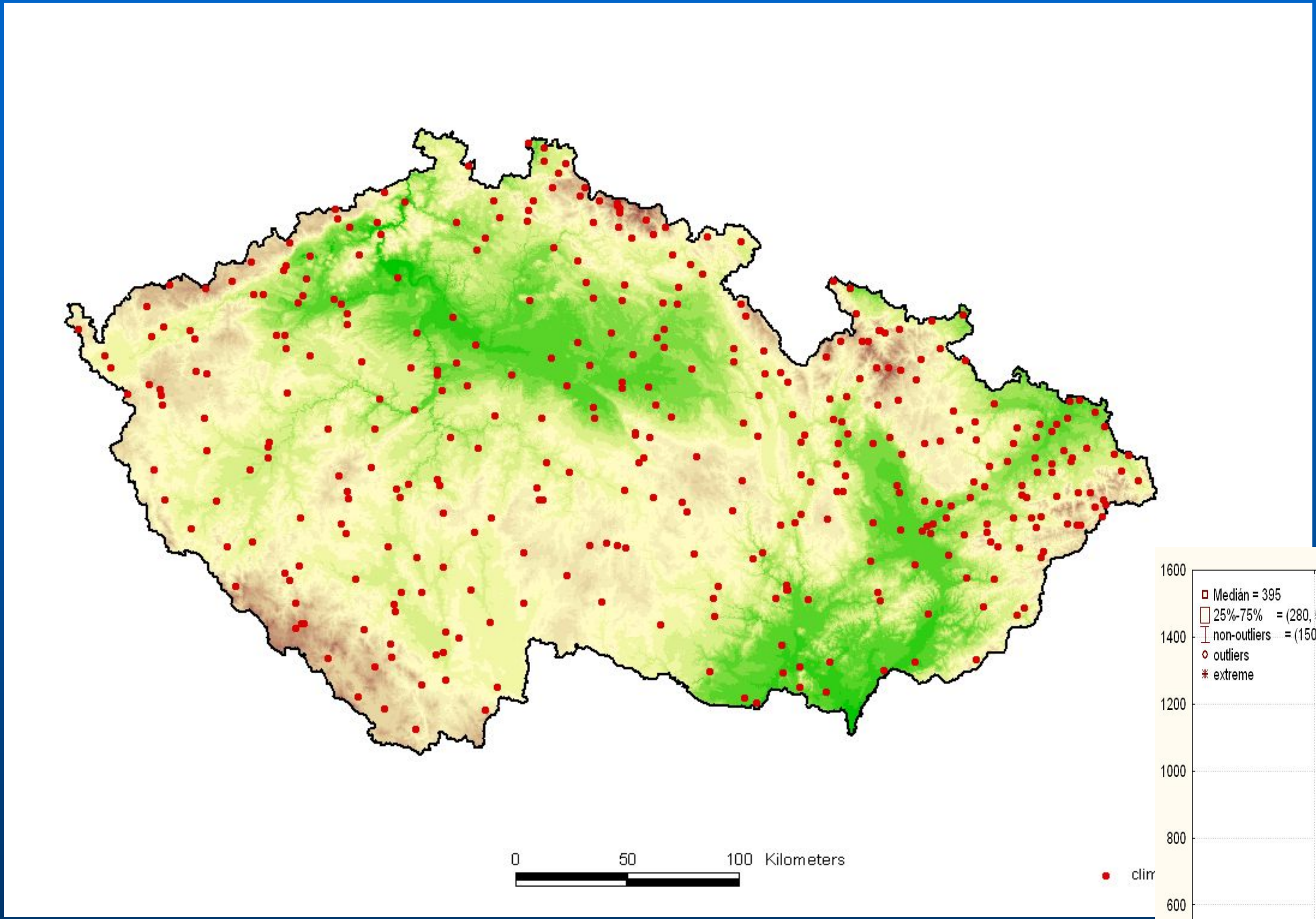
Homogenization

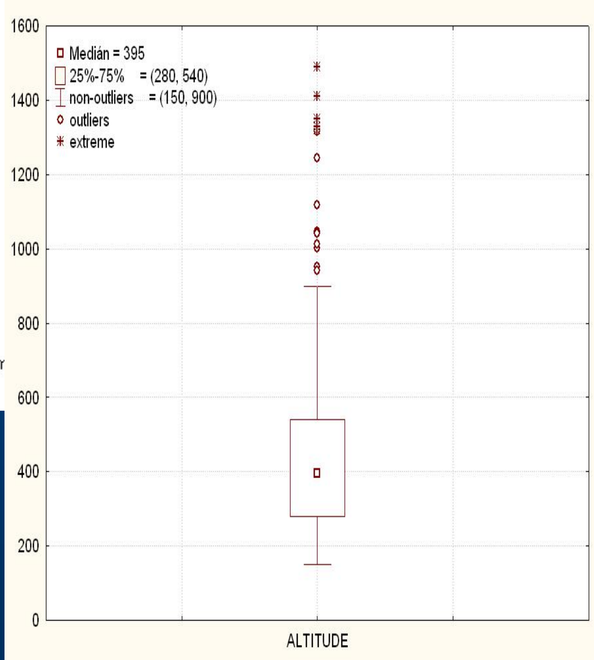
of the series in the Czech Republic





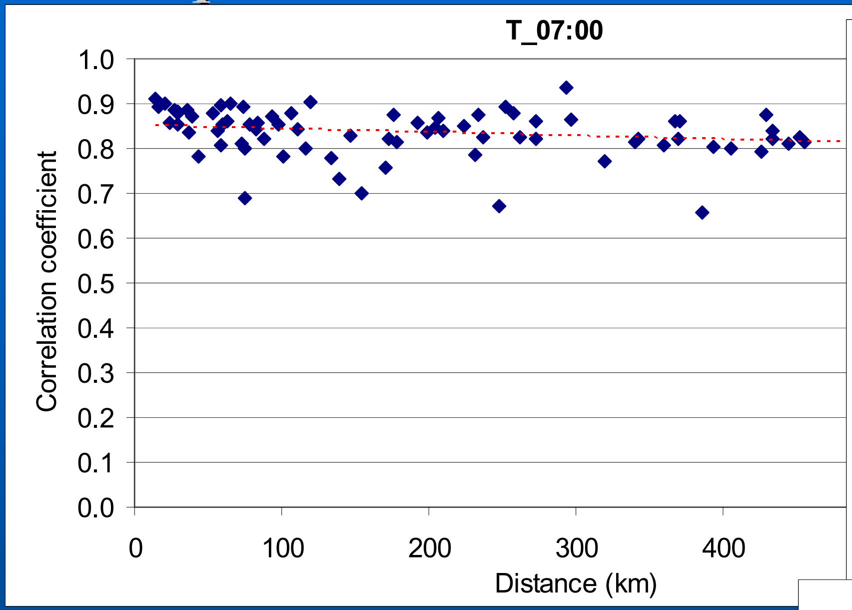
Spatial distribution of climatological stations

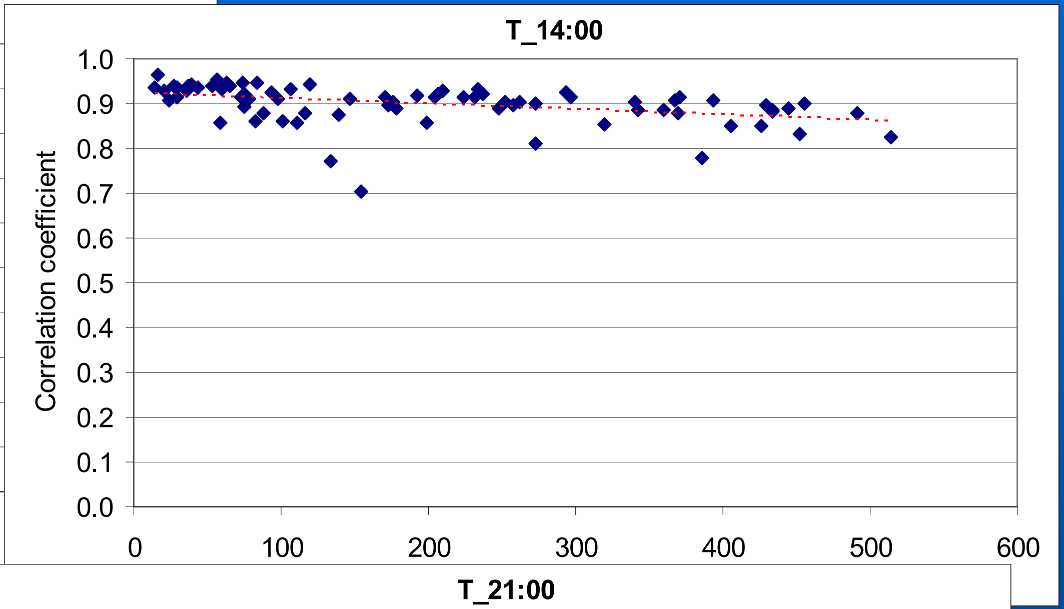


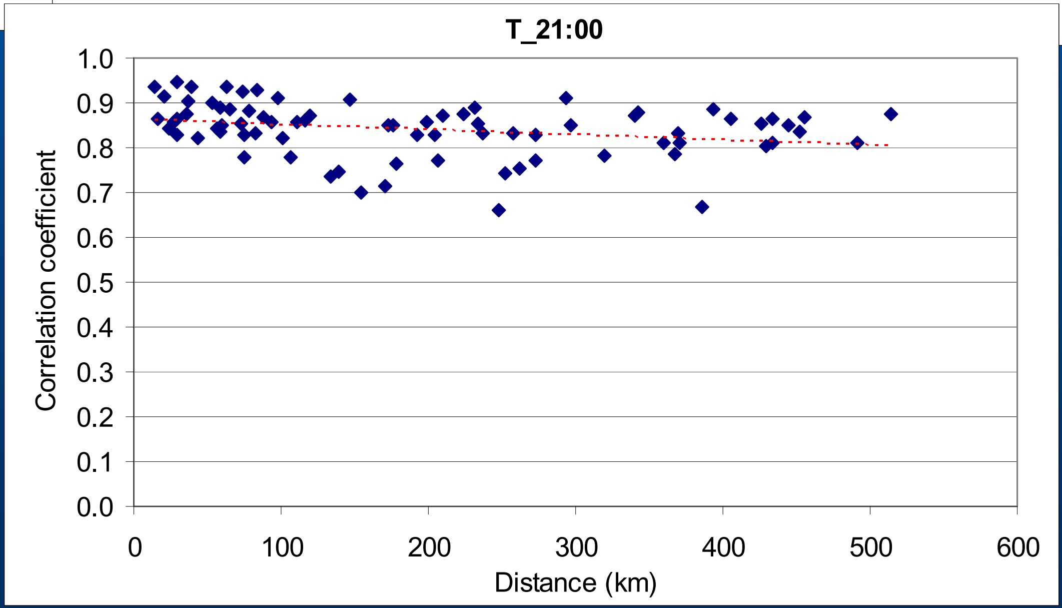


- period 1961-2007
- 200 stations
- mean minimum distance: 12 km

Correlation coefficients, change in space, monthly air temperature

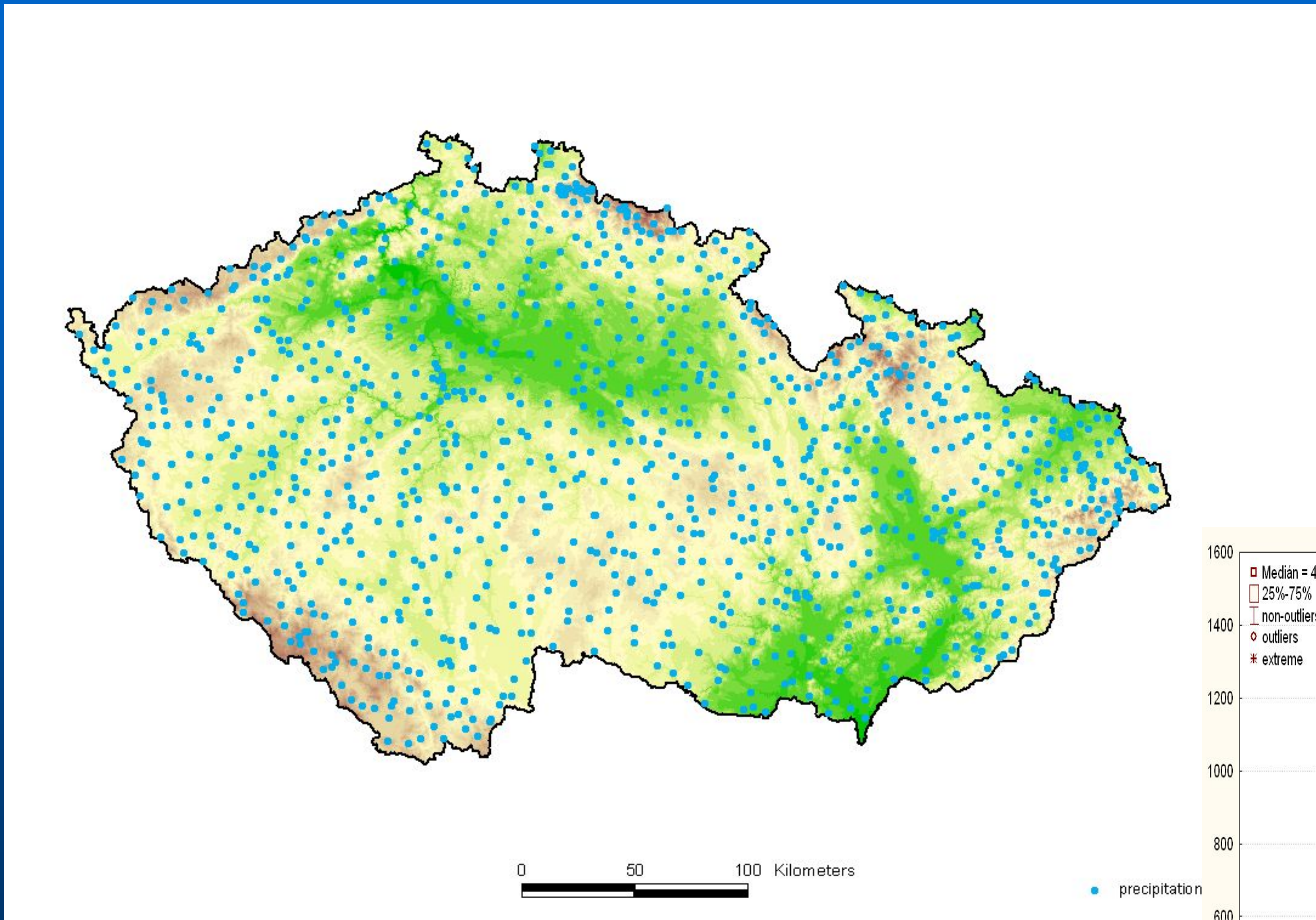


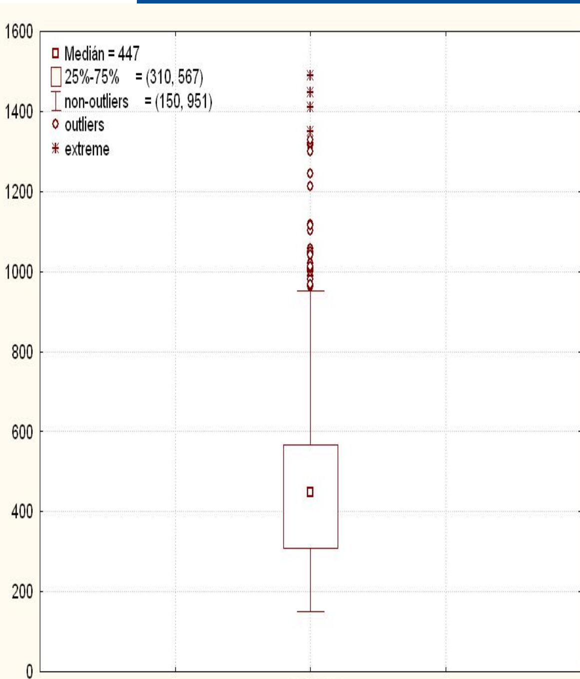




Average of monthly correlation coefficients, 1961-2000, individual observation hours

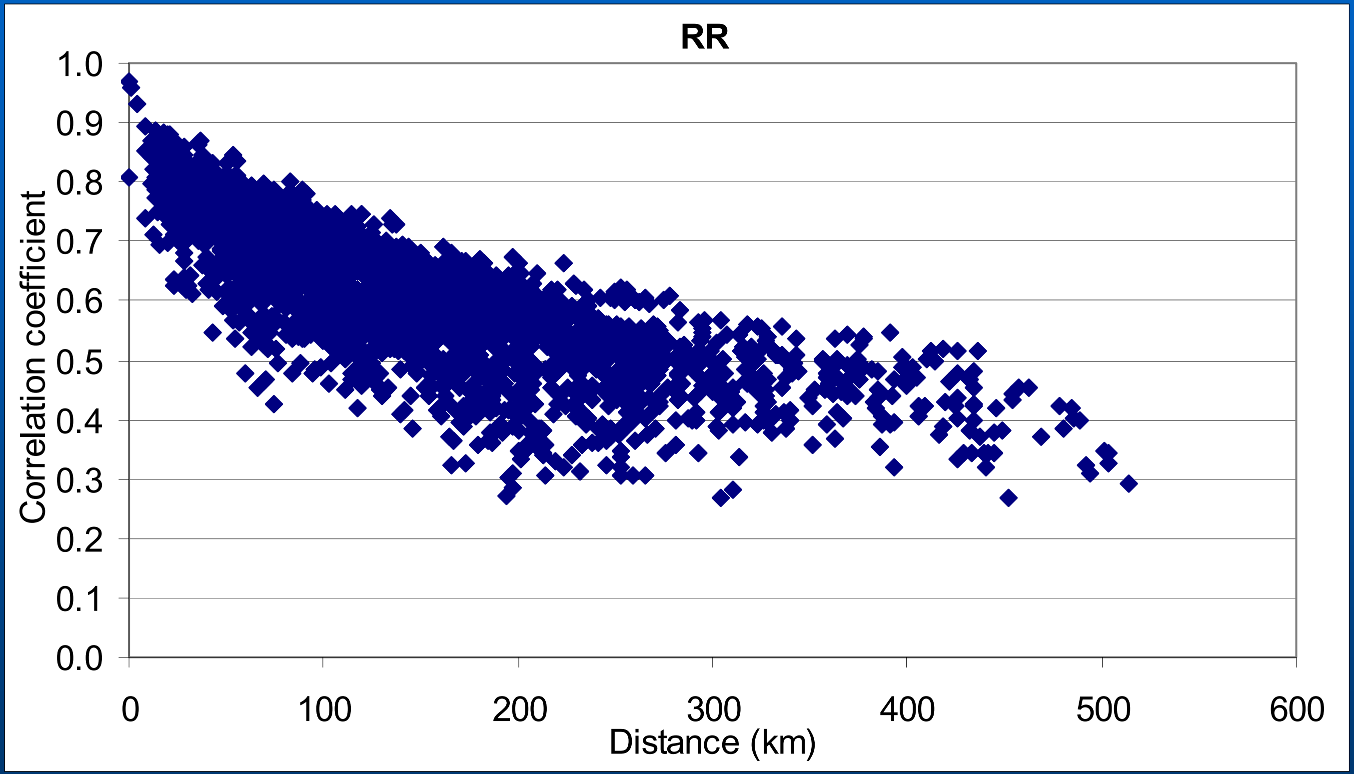
Spatial distribution of precipitation stations





- period 1961-2007
- 600 stations
- mean minimum distance: 7.5 km

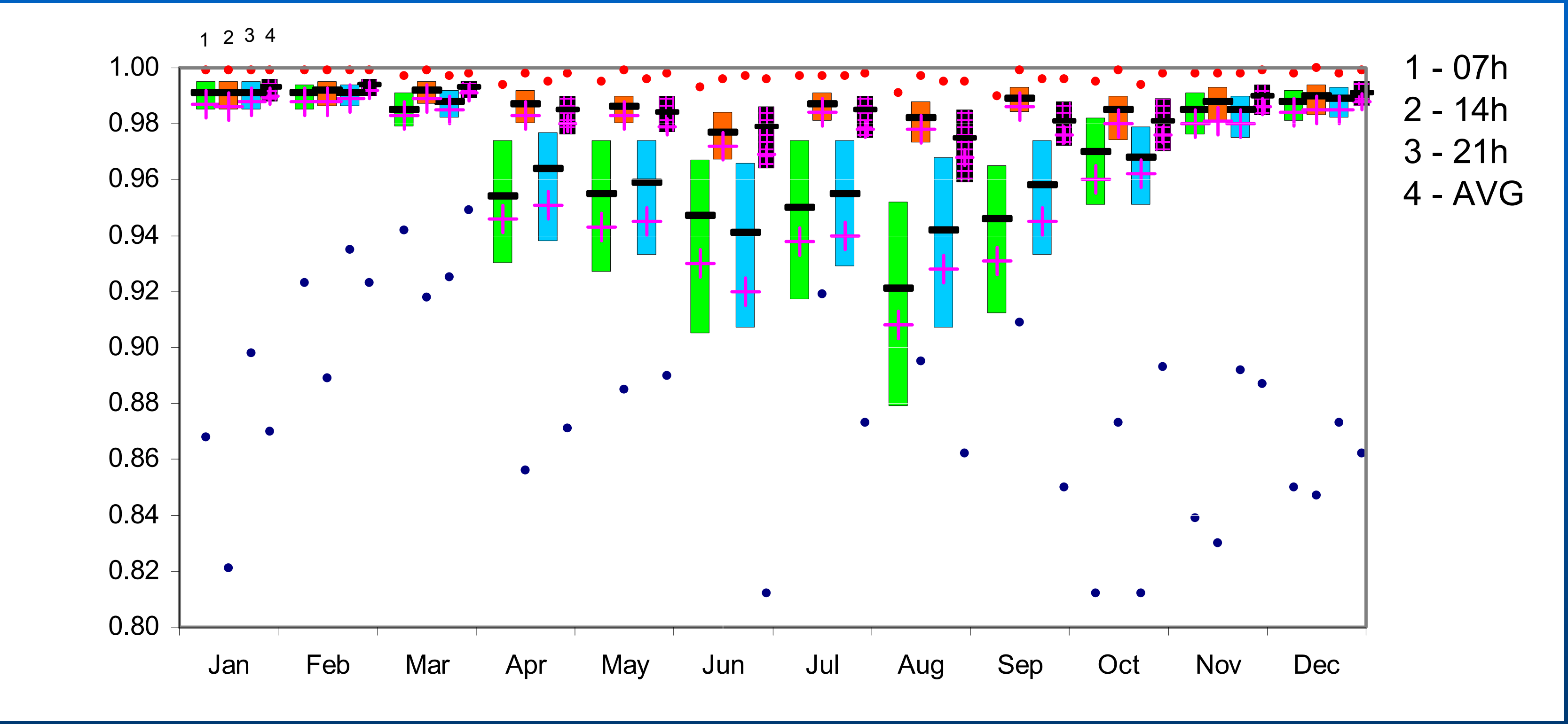
Correlation coefficients, change in space, monthly precipitation



2483 values, average of monthly correlation coefficients

Correlations between tested and reference series

Air temperature



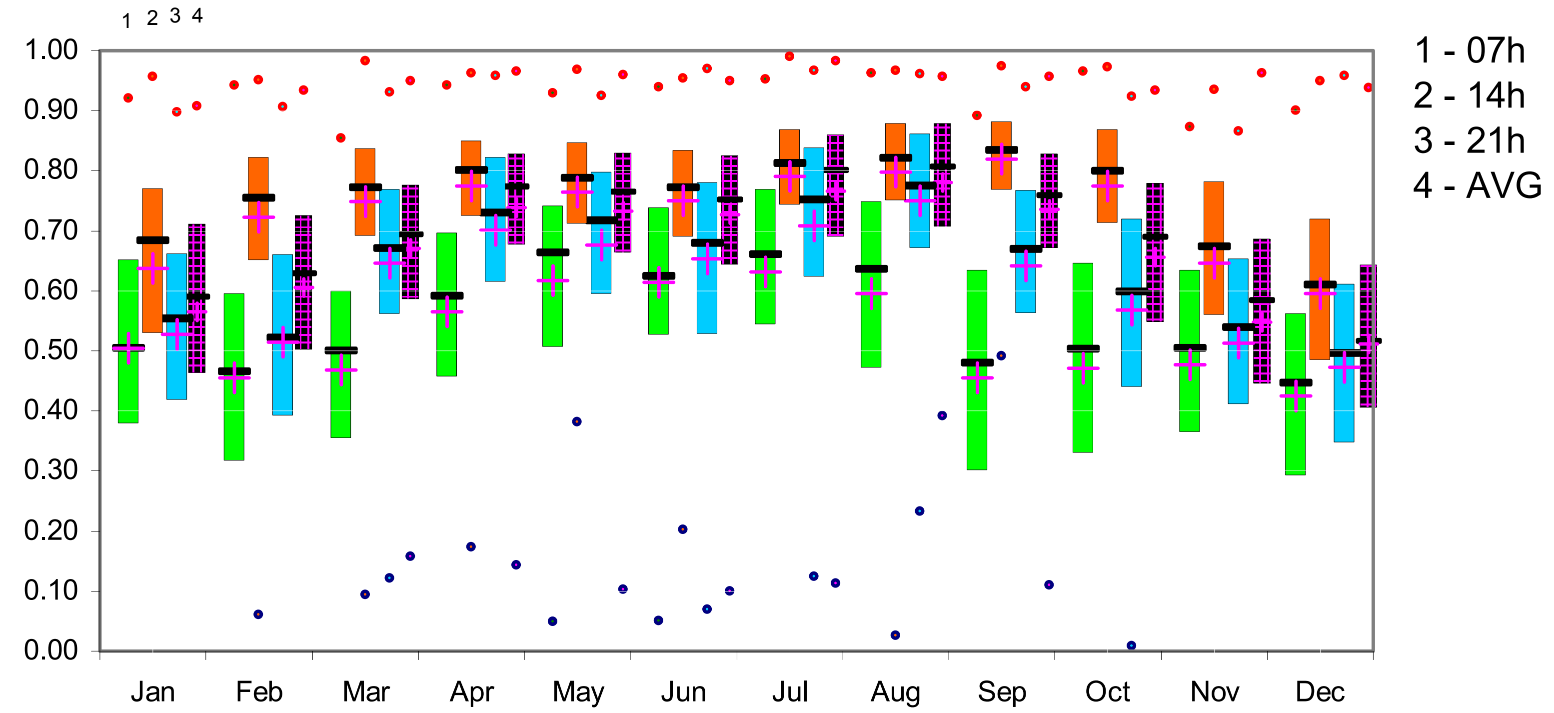
Boxplots:

- Median
- Upper and lower quartiles

(for 200 testes series)

Correlations between tested and reference series

Relative Humidity



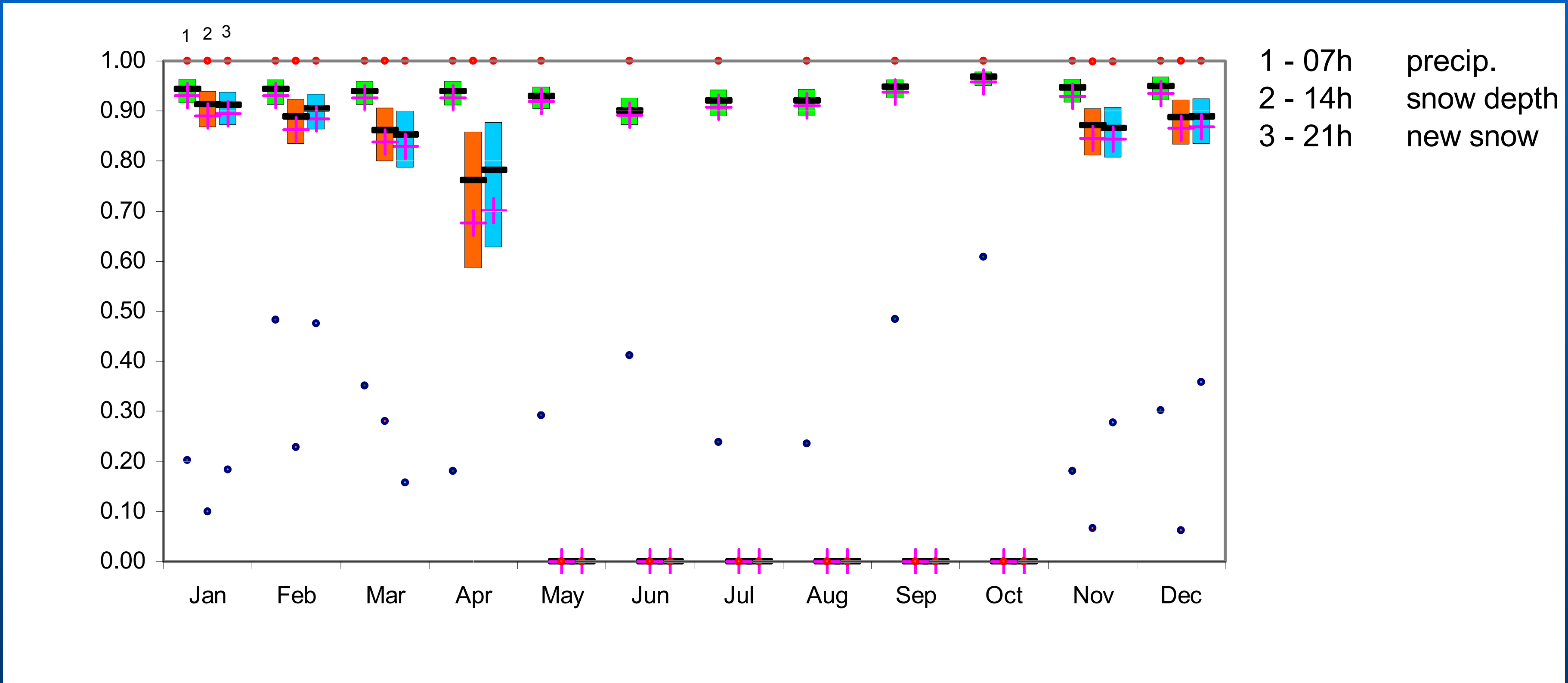
Boxplots:

- Median
- Upper and lower quartiles

(for 200 testes series)

Correlations between tested and reference series

Precipitation, snow depth, new snow



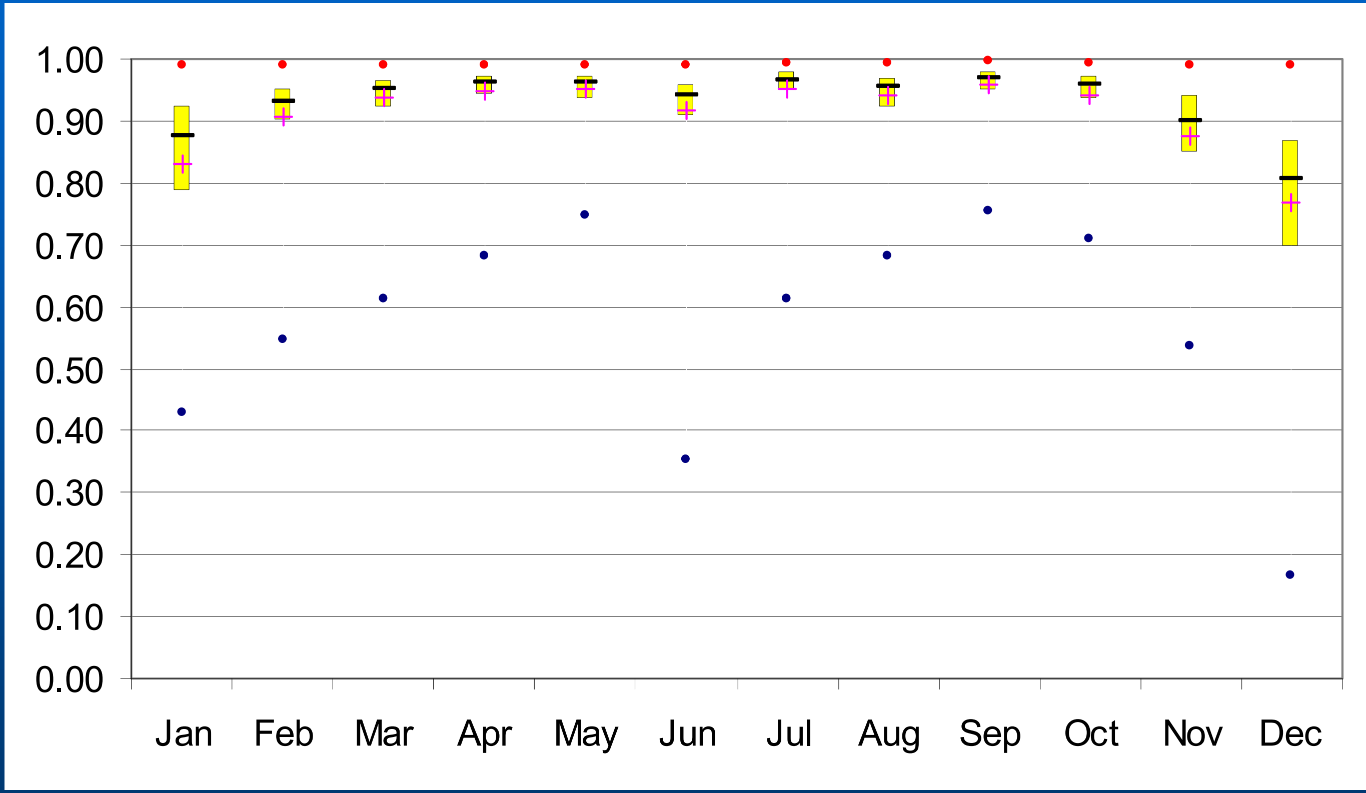
Boxplots:

- Median
- Upper and lower quartiles

(for 800 testes series)

Correlations between tested and reference series

Sunshine duration



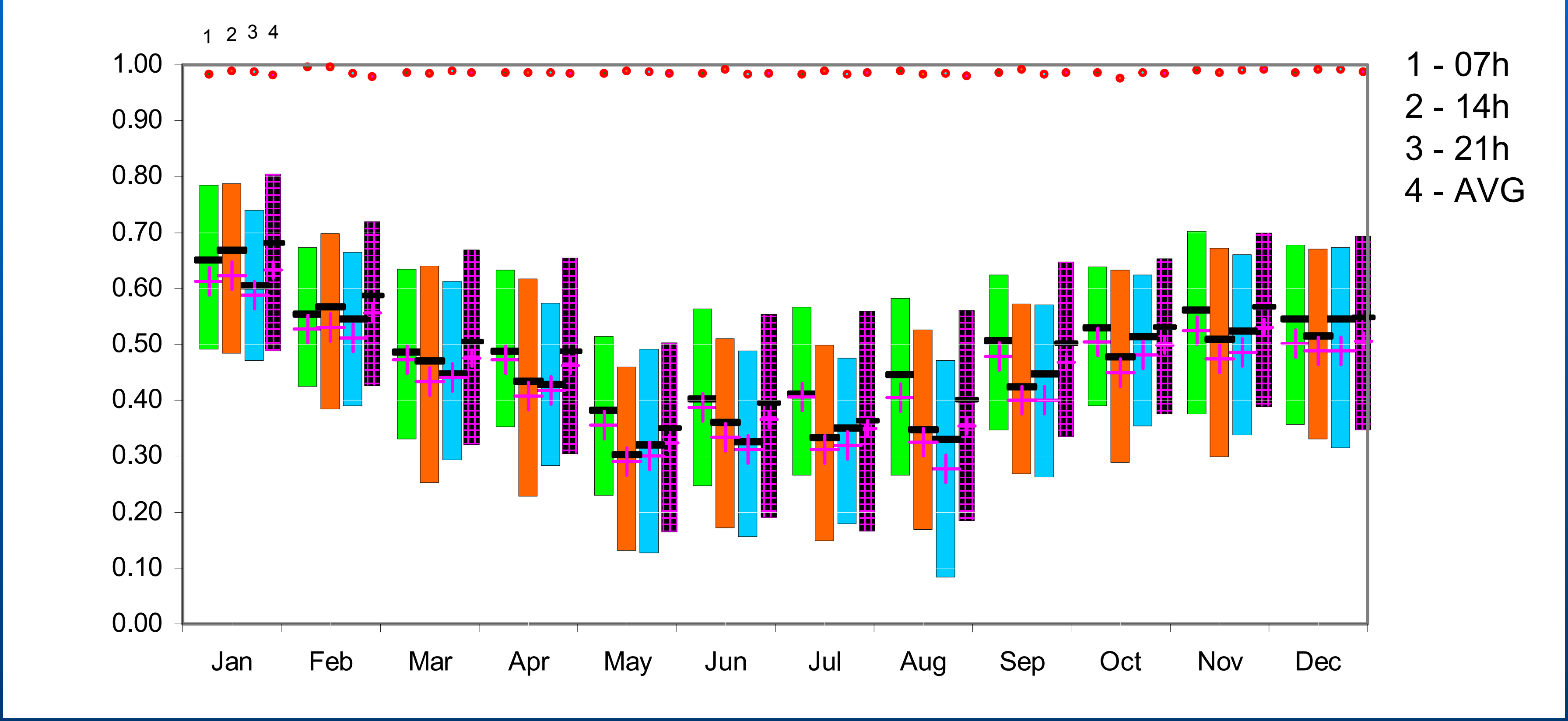
Boxplots:

- Median
- Upper and lower quartiles

(for 100 testes series)

Correlations between tested and reference series

Wind speed



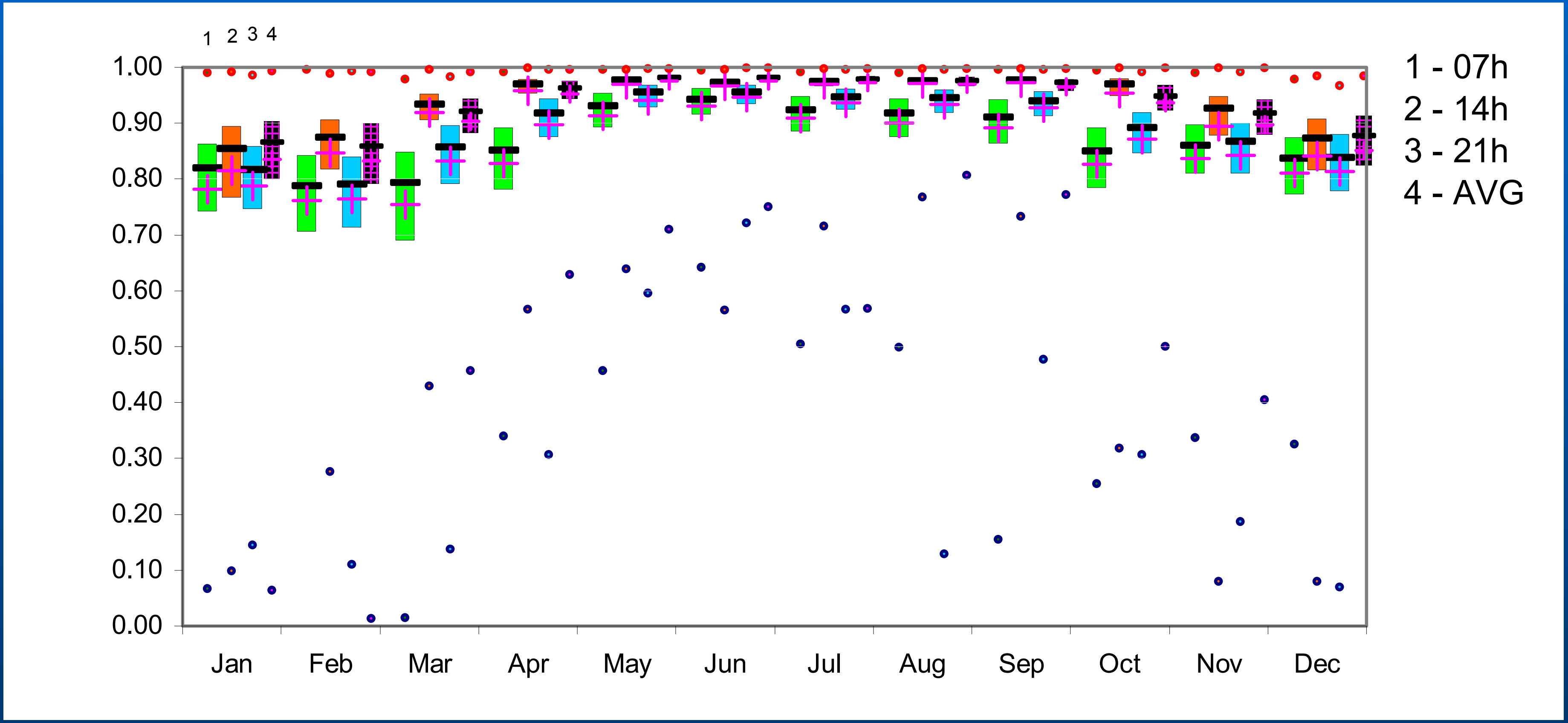
Boxplots:

- Median
- Upper and lower quartiles

(for 200 testes series)

Correlations between tested and reference series

Temperature, daily values



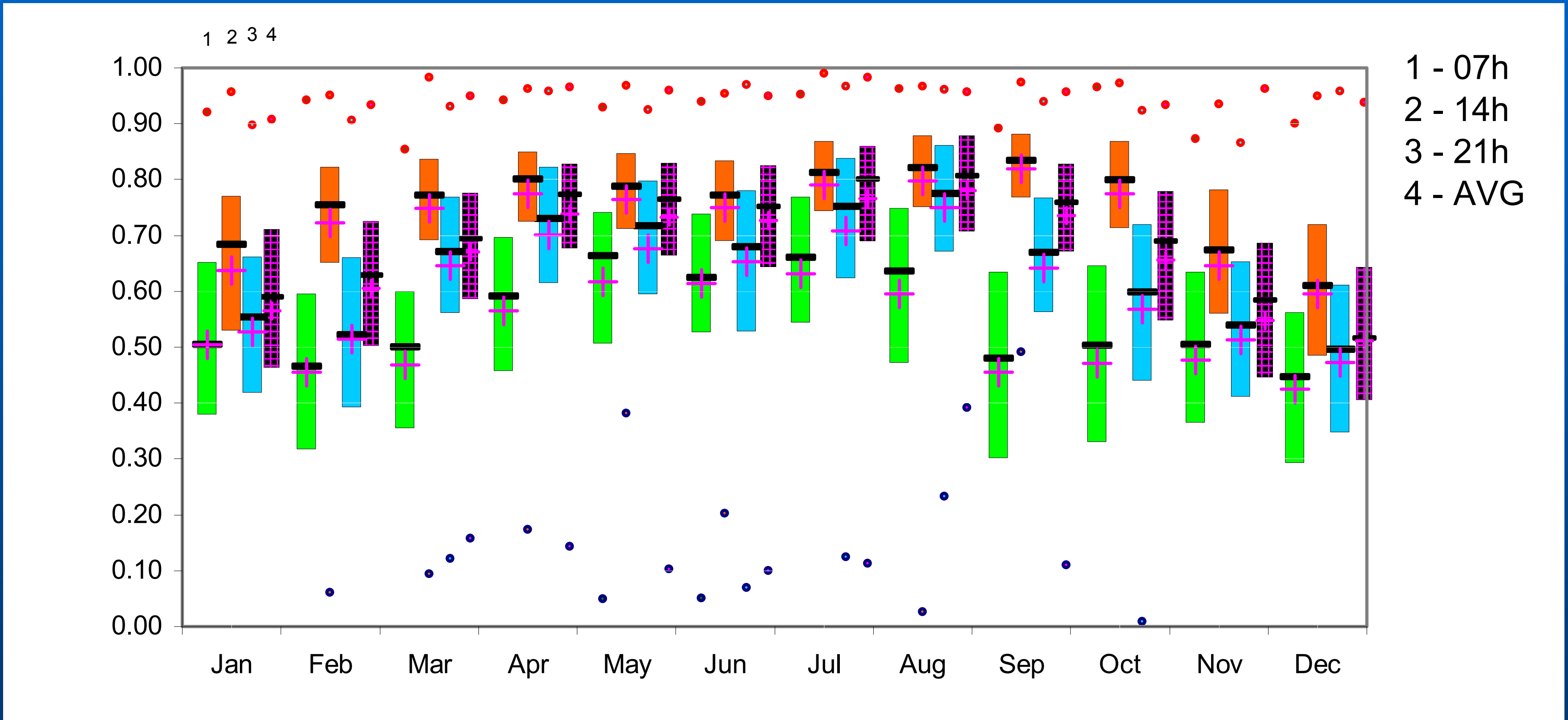
Boxplots:

- Median
- Upper and lower quartiles

(for 200 testes series)

Correlations between tested and reference series

Relative humidity, daily values



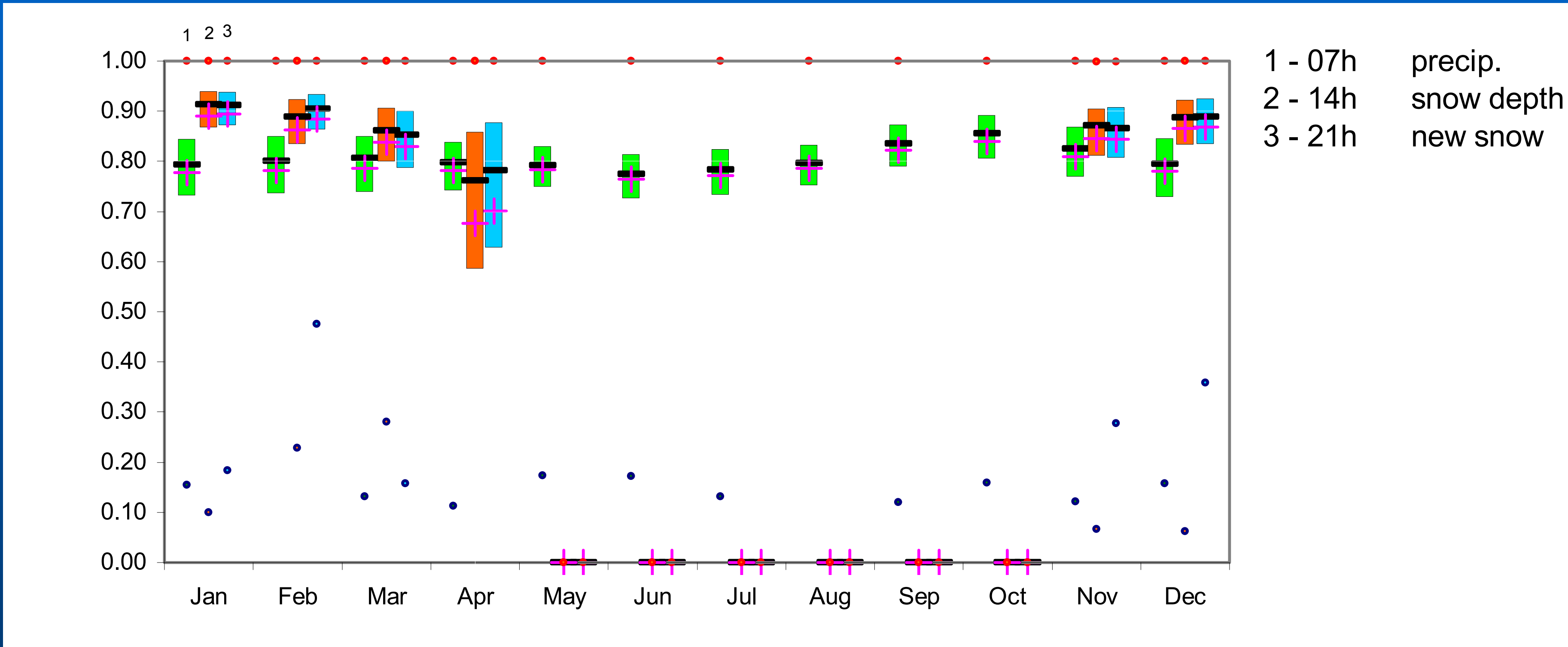
Boxplots:

- Median
- Upper and lower quartiles

(for 200 testes series)

Correlations between tested and reference series

Precipitation, daily values (**>0.1, ln transformation**)



Boxplots:

- Median
- Upper and lower quartiles

(for 200 testes series)

Using RCM simulations data as a reference series ALADIN-CLIMATE/CZ

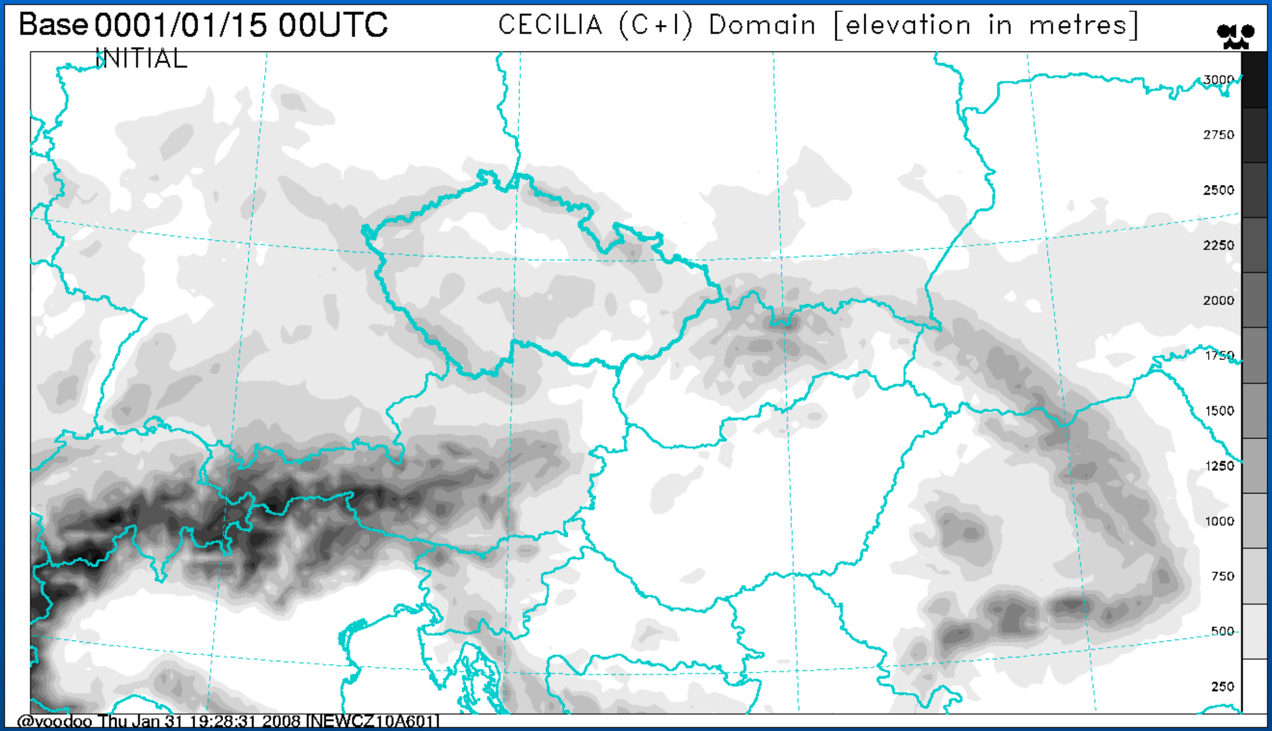
- **NWP LAM ALADIN – being developed by consortium of European and N. African countries led by Météo-France**
- **ALADIN-CLIMATE/CZ based on CY28 NWP version**
- **Physical parameterizations package (pre-ALARO) based partly on EC FP5 MFSTEP development**
- **Used in FP6 projects ENSEMBLES, CECILIA & several national research projects**
- **At CHMI used at NEC-SX6 central computer**
- **To be superceded by CY32 version with ALARO physics (addressing the 5-7km resolution) and first tests to be run during spring 2008**

EC FP6 CECILIA

Climate modeling part (WP2):

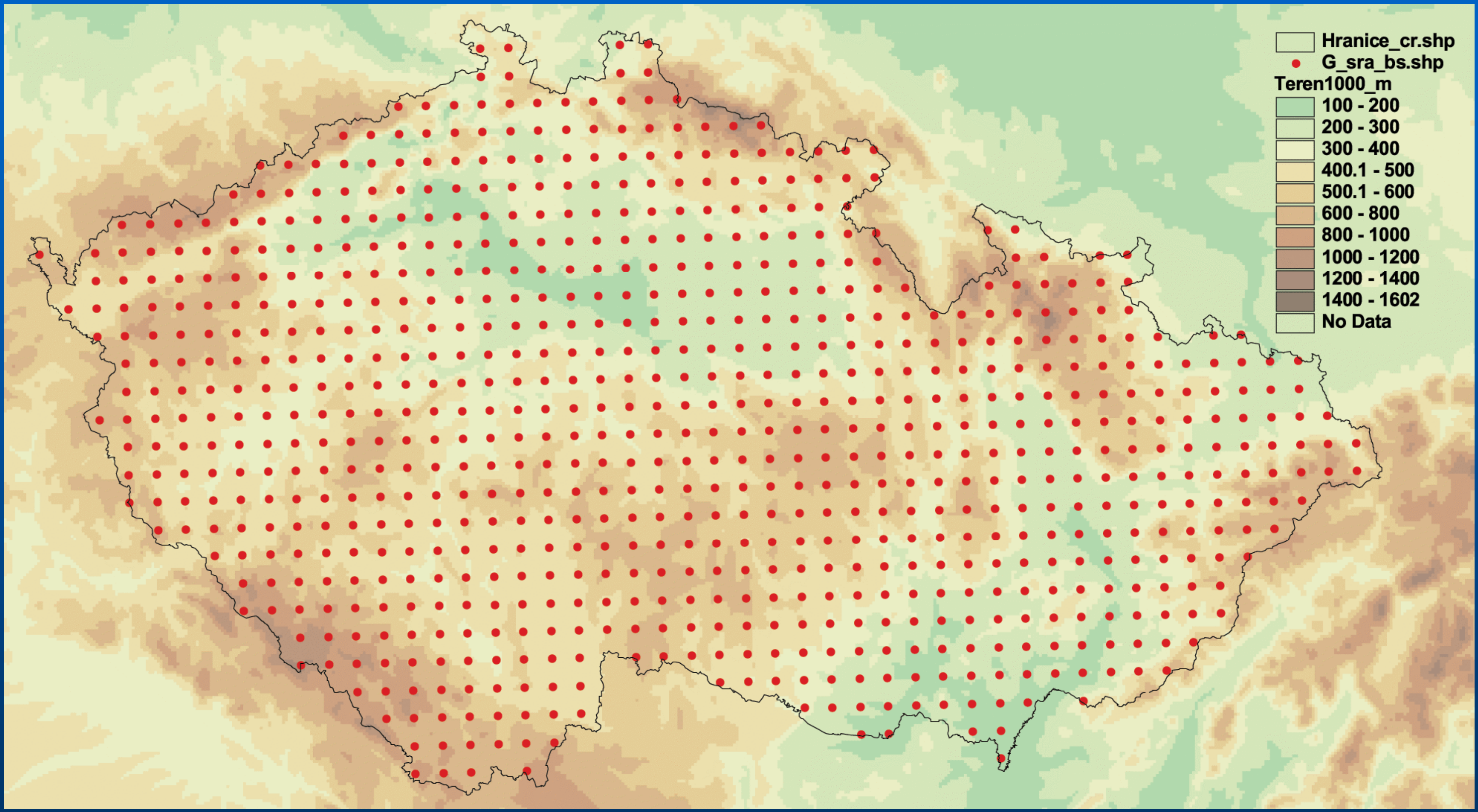
- **CHMI ALADIN CLIMATE/CZ + ARPEGE-CLIMATE**
- **1961 – 2000 ECMWF ERA-40 run** ***(finished …)***
- **1960 – 1990 “present time” slice (*finished …*)**
- **2020 – 2050 “near future” slice** ***(finished …)***
- **2070 – 2100 “distant future” slice** ***(being calculated)***

CECILIA experiments …



- 10 km spatial step
- 450 seconds time step
- 43 atmosphere levels
- one month integration ~20.000 s. at NEC computer in Prague
- 164 x 90 points (LON x LAT)

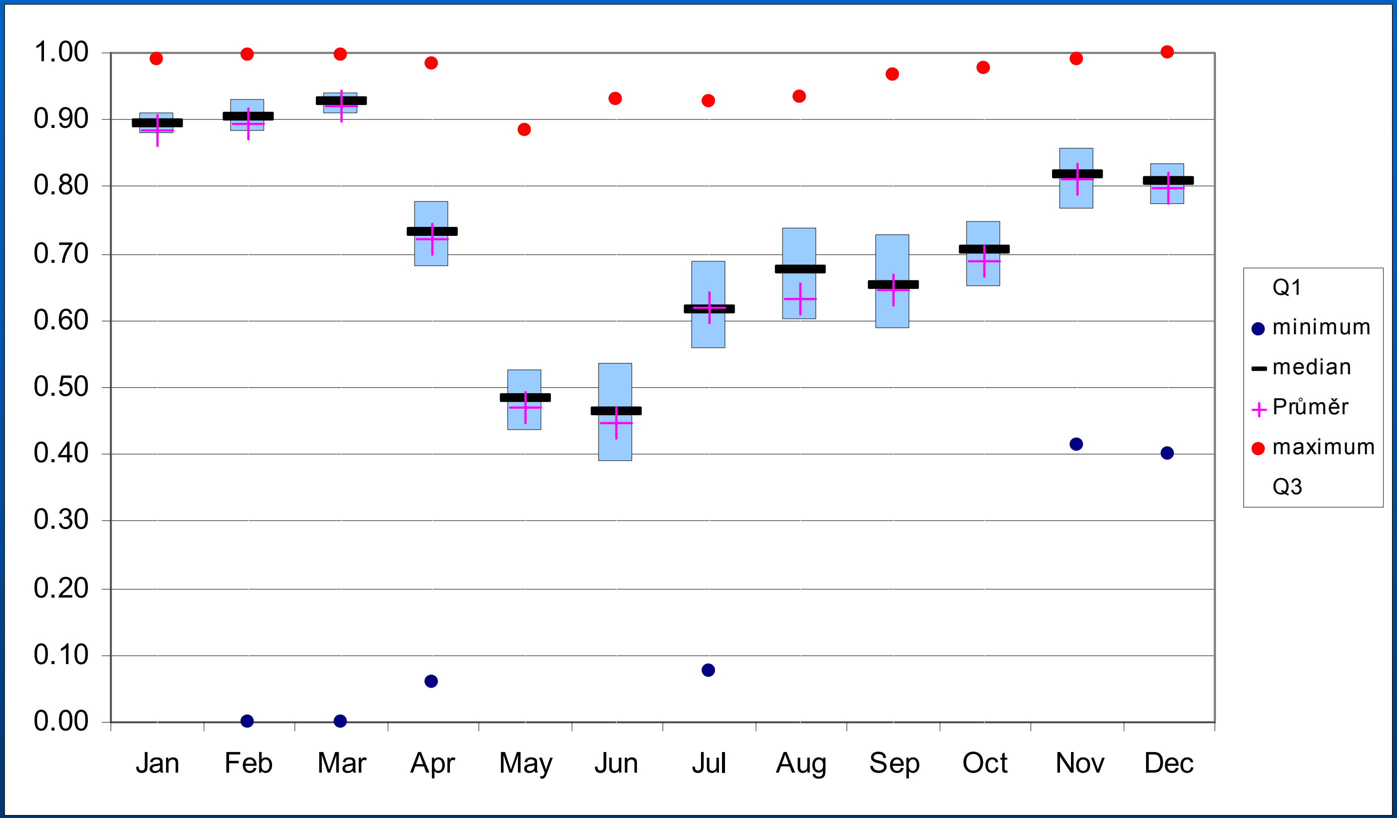
ALADIN CLIMATE/CZ
Grid points over the Czech Republic



10 km model resolution = 789 grid points in total => similar to precipitation station network density

Correlations between tested and reference series

Air temperature, RCM reference series



Boxplots:

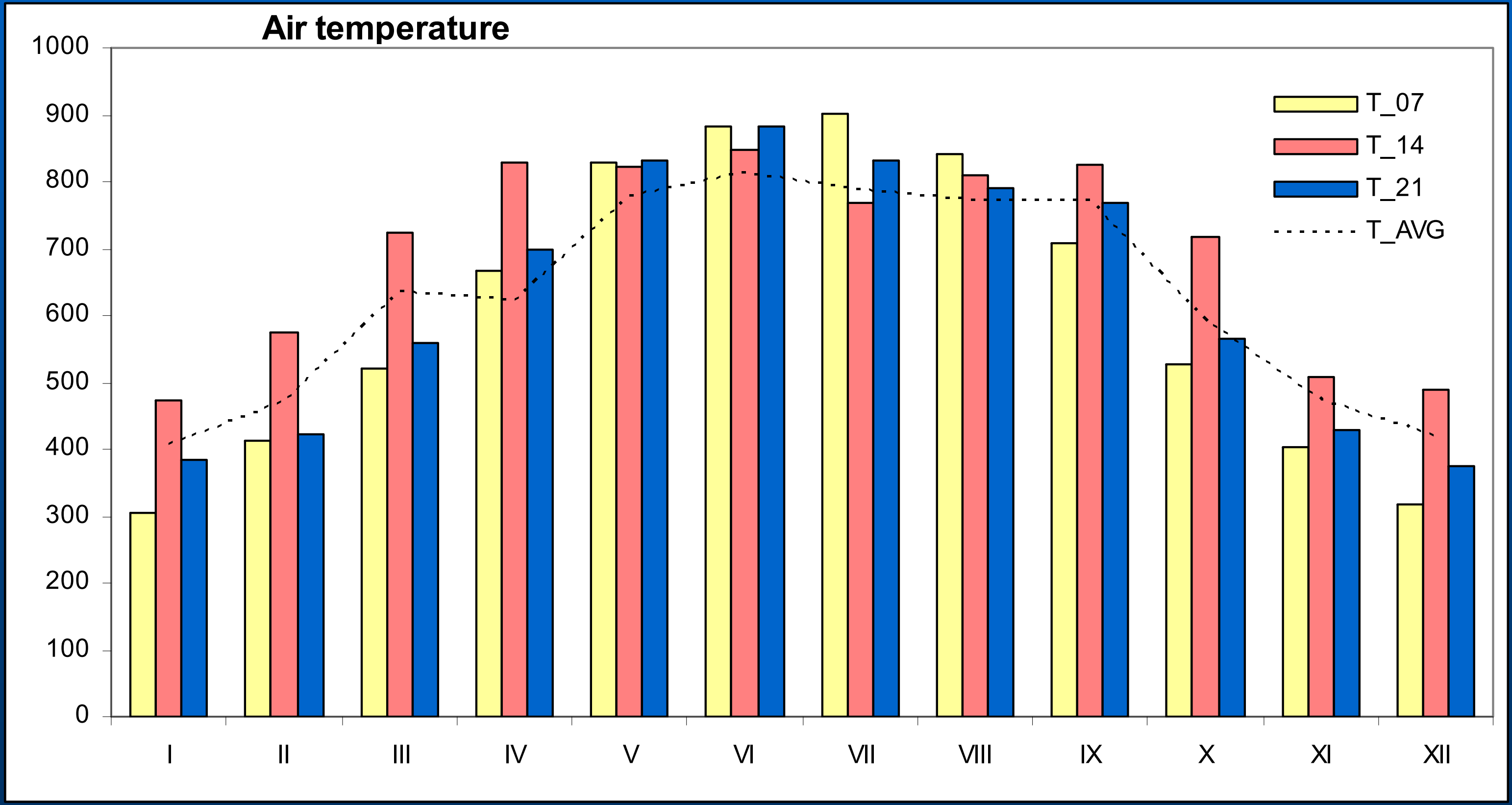
- Median
- Upper and lower quartiles

(for 400 testes series)

Homogeneity testing results
Air temperature

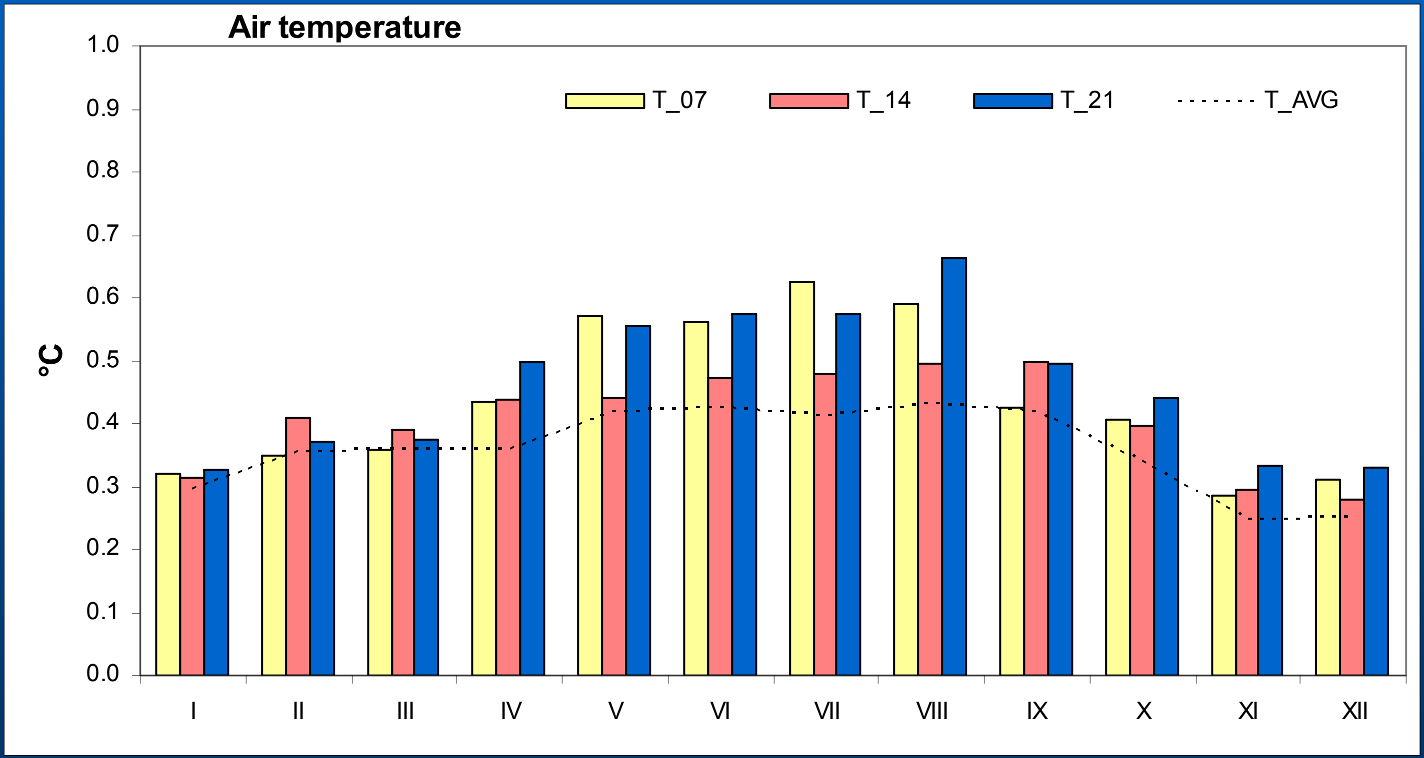
Number of significant inhomogeneities (0.05) detected by used tests

(*A*, *B* tests, *c* and *d* reference series, alltogether)



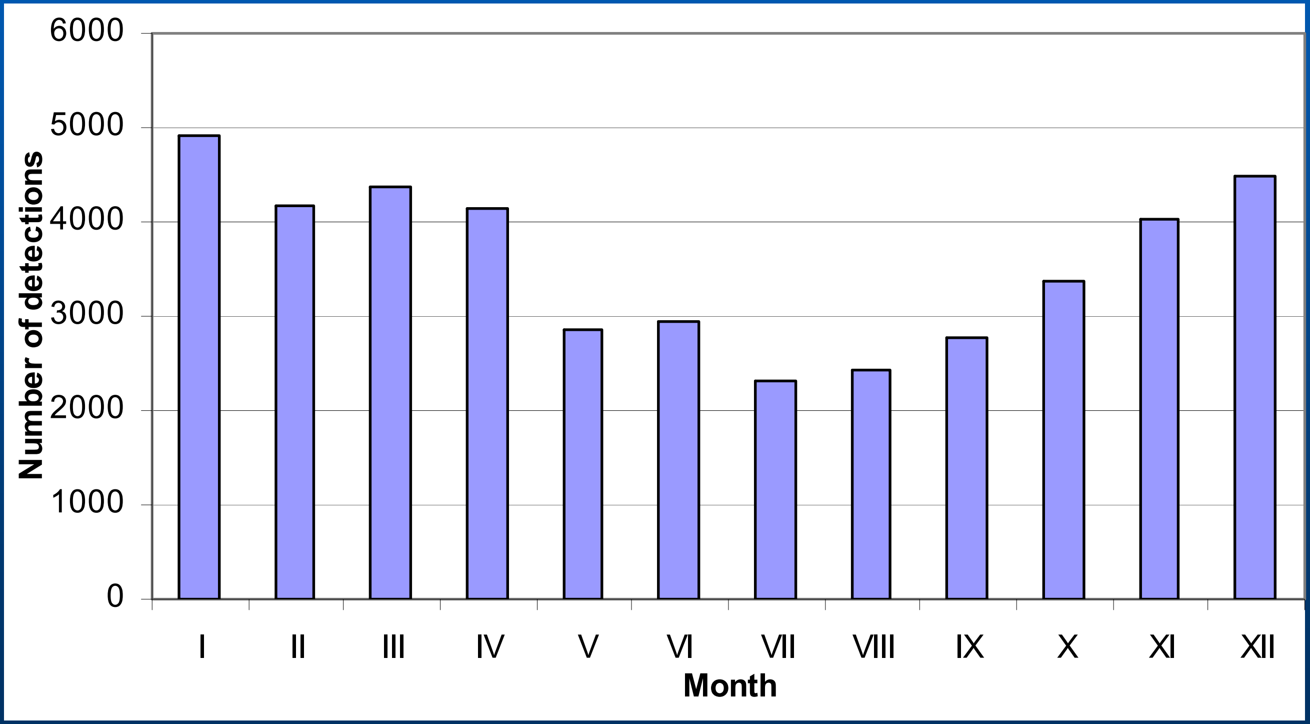
Homogeneity testing results
Air temperature

Amount of adjustments, averages of absolute values, T_AVG



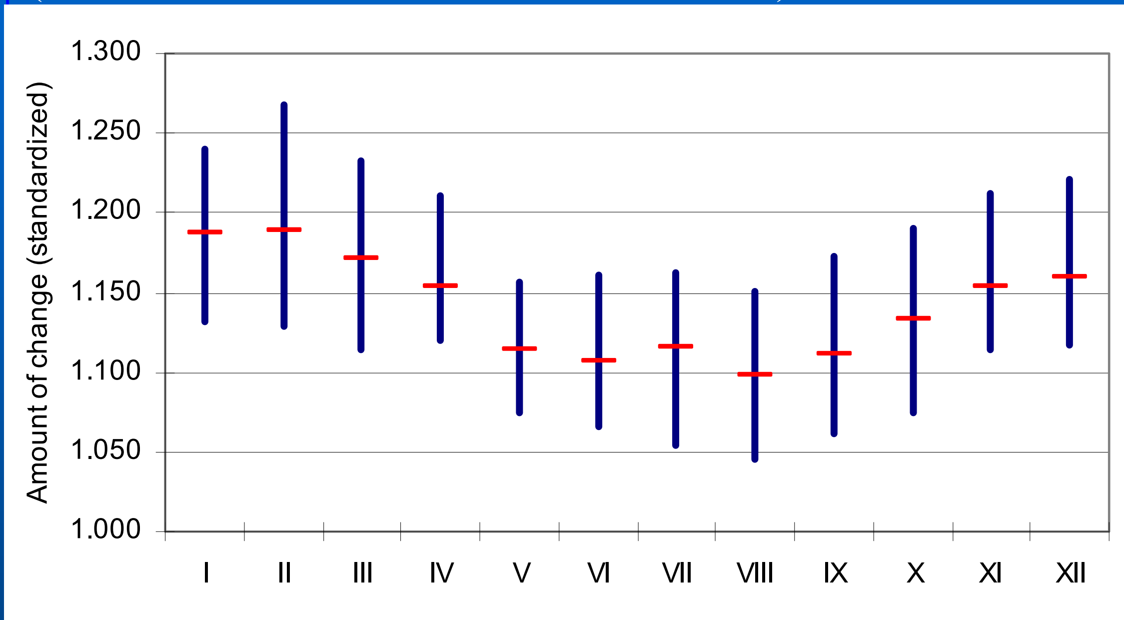
Homogeneity testing results Precipitation

- **4 tests, 4 reference series, 12 months + 4 seasons and year**
- **Number of detected inhomogeneities (significant)**



Amount of change (ratios – standardized to be >1.0), precipitation

(reference series calculation based on correlations)

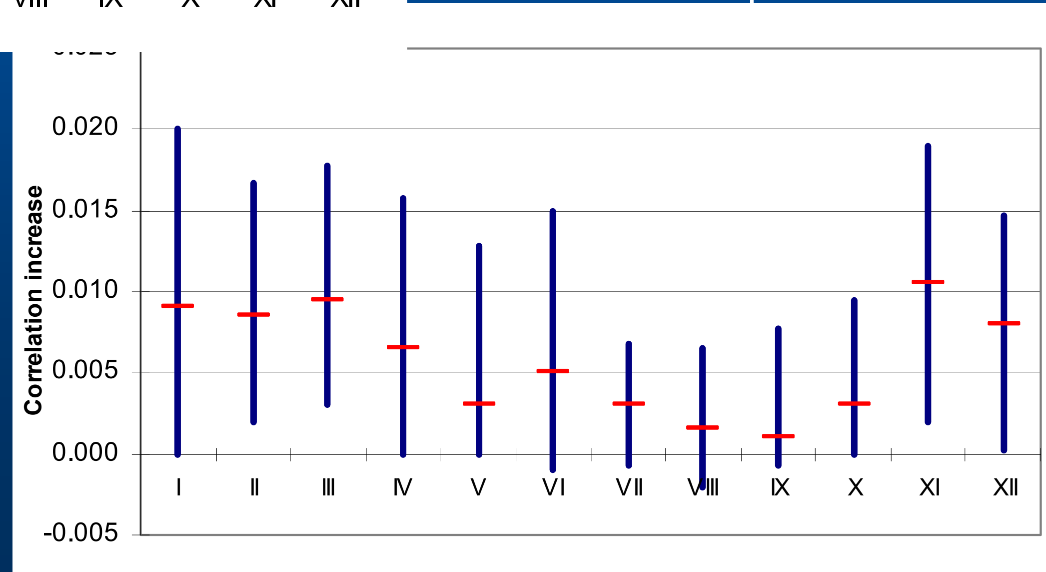


Boxplots:

- Median
- Upper and lower quartiles

(for 589 testes series)

Correlation improvement



Inhomogeneities in summer versus in winter, Air **temperature**

- Change of measuring conditions at the station (relocation etc.) is manifested in the series mainly in **summer**
- in winter: active surface role is diminished, prevailng circulation factors, in summer: active surface role increases, prevailing radiation factors

Inhomogeneities in summer versus in winter, **Precipitation**

- Change of measuring conditions at the station (relocation etc.) is manifested in the series mainly in **winter**
- in winter: errors of measurement (solid precipitation - wind, …)

Homogenization
Final remarks, recommendations 1/3

- **data quality control before homogenization is of very importance (if it is not part of it)**
- **Using series of observation hours (complementarily to daily AVG) is highly recommended (different manifestation of breaks)**
- **be aware of annual cycle of inhomogeneities, adjustments, …**
- **to know behavior of spatial correlations (of element being processed) to be able to create reference series of sufficient quality …**

Homogenization
Final remarks, recommendations 2/3

- Because of **Noise in the time series** it makes sense:
- - „Ensemble“ approach to homogenization (combining information from different statistical tests, time frames, overlapping periods, reference series, meteorological elements, …)
- - more information for inhomogeneities assessment – higher quality of homogenization in case metadata are incomplete

Homogenization of **daily values**, remarks 3/3

- **Correlation coefficients** (tested versus reference series) **are slightly lower** (compared to monthly data), **but still high enough** (around 0.9 even in case precipitation)
- **Advantage: reliable inhomogeneities detection near the ends of series**
- **Complementary information to monthly and seasonal values detections (but problems with distribution, autocorrelations, …)**

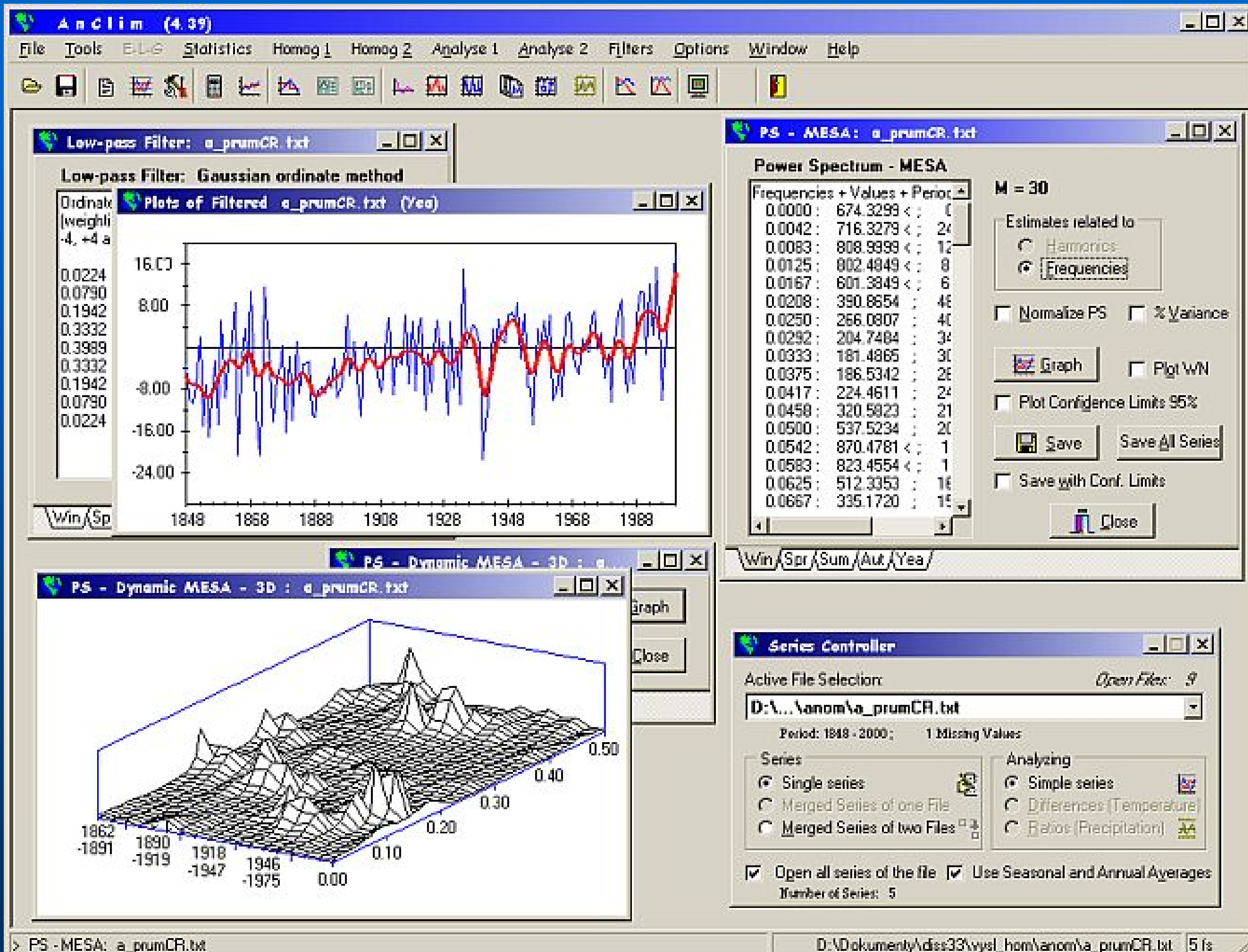
- **Correction of daily data:**
  - **"delta" method, if applied, it should be discriminated with regard to other parameters like cloudiness, …**
  - **Variable correction (such as HOM) seems to be a good choice … (preserving CDF)**

Software used for data processing

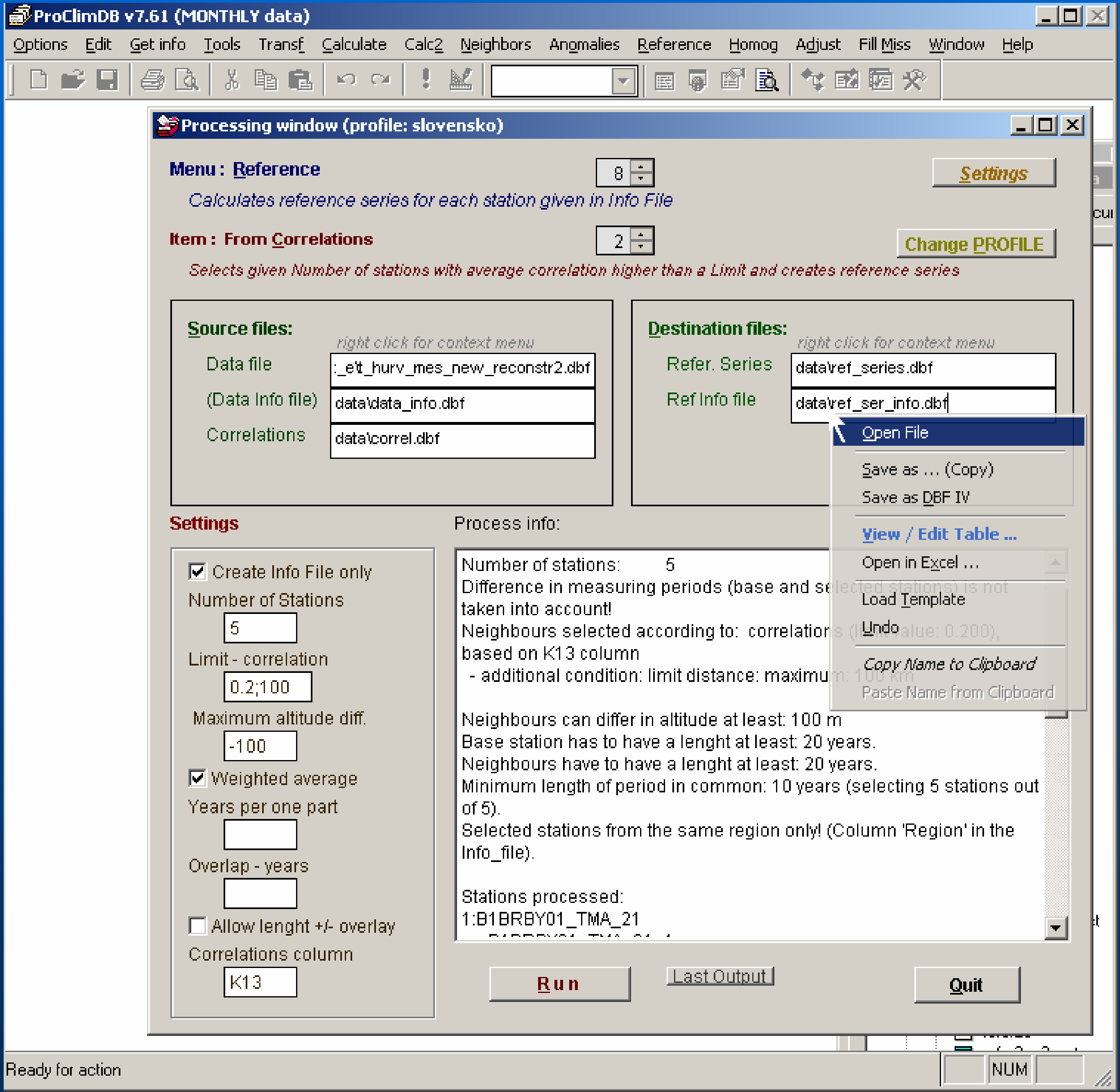
- **LoadData - application for downloading data from central database (e.g. Oracle)**
- **ProClimDB software for processing whole dataset (finding outliers, combining series, creating reference series, preparing data for homogeneity testing, extreme value analysis, RCM outputs validation, correction, …)**
- **AnClim software for homogeneity testing**

http://www.climahom.eu

AnClim software



ProClimDB software



http://www.climahom.eu