Data quality control and homogenization of air temperature and precipitation series in the area of the Czech Republic since 1961

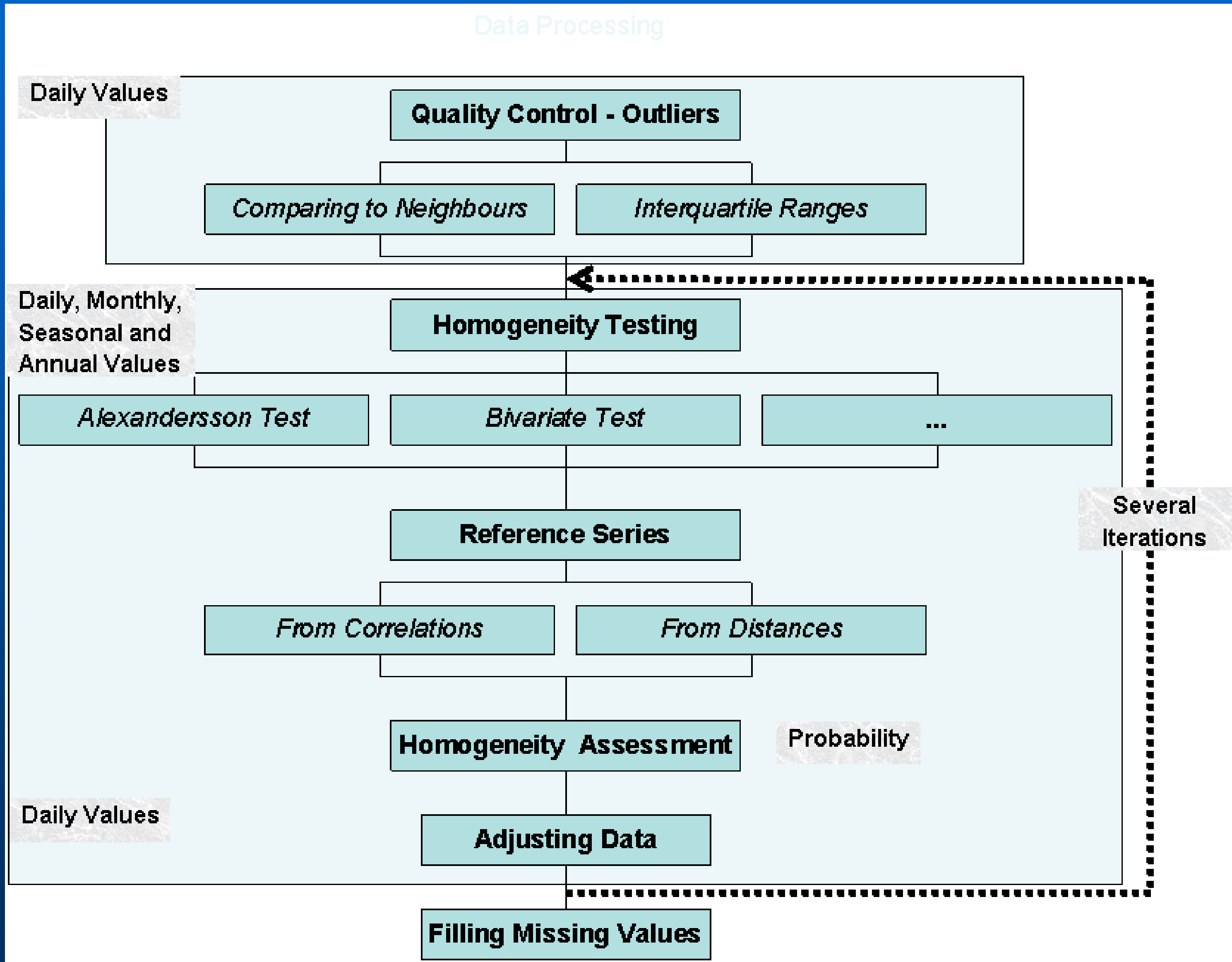
P. Štěpánek [(1)], P. Zahradníček [(1)], P. Skalák [(1)]

[1] Czech Hydrometeorological Institute, Czech Republic

E-mail: petr.stepanek@chmi.cz

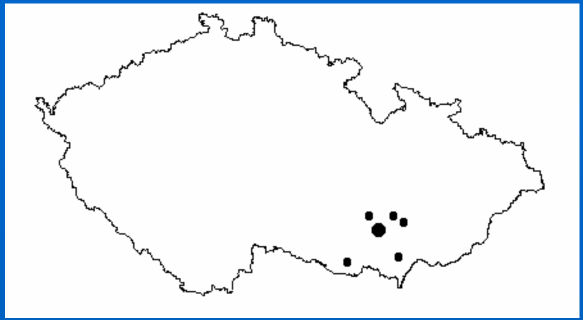
8th Annual Meeting of the EMS / 7th ECAC

Processing before any data analysis



Software AnClim, ProClimDB

Data Quality Control
Finding Outliers



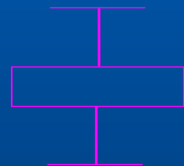
- **Comparing values to values of neighbouring stations**
  - **comparing to <u>min. 3</u> to 10 best correlated (nearest) stations**
  - **calculating series of standardized differences (logarithms of ratios)**
  - **number of cases exceeding 95% confidence limits is counted**
  - **Standardization of neighbours to base station values (AVG, STD, Altitude),**

Data Quality Control
Neighbours values Standardization

- Characteristics calculated **from the standardized values:**
  - **coefficient of Interquartile range** (ranges are estimated from standardized neighbours values)
  - **difference of base station and median**

    from neighbours values (probability):

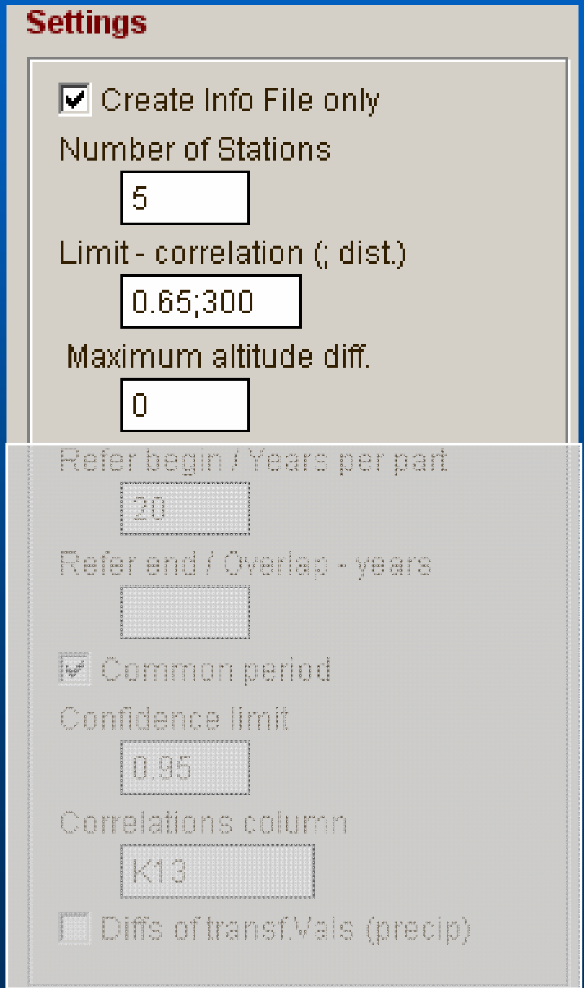
    CDF for ((base station – median_from_standardized_neighbors_values)/STD_base_station)

    
  - **„Expected“ value** (as weighted mean with weights 1/distance or correlations, arbitrary power; possibility of using trimmed mean) and comparison with original value

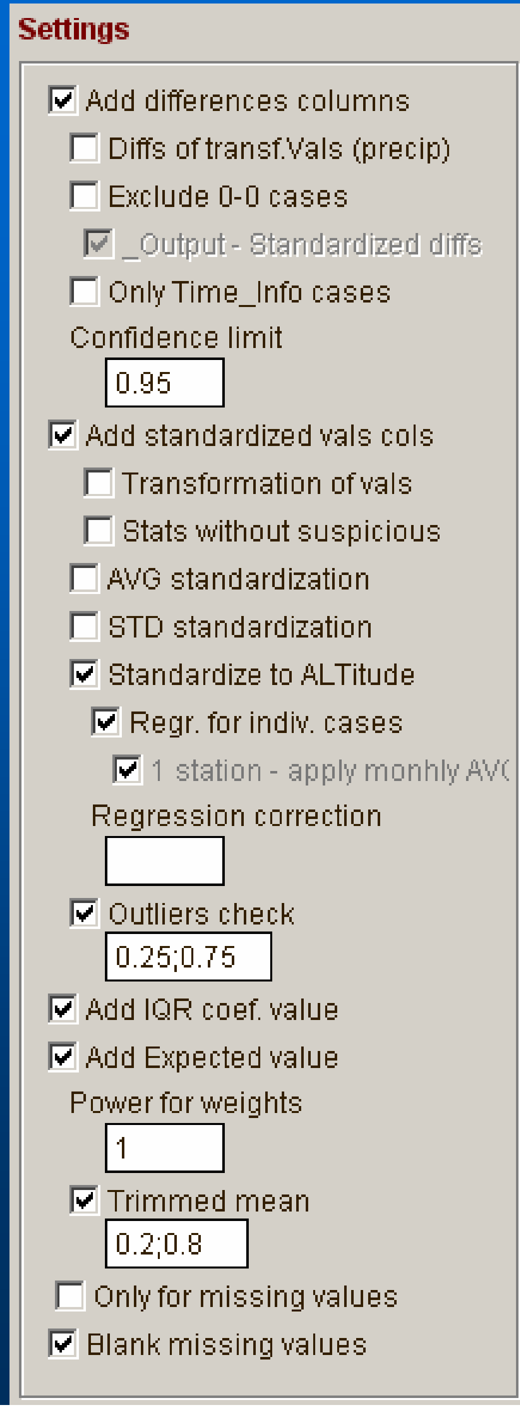
QC, Settings in the software

processing the whole database

1. Finding neighbours:



2. Calculation:



Example of outputs for outliers assessment

Suspicious values

Expected value

Neighbour stations values

List of neighbours

Altitudes

and distances of neighbours

	B	C	D	E	F	G	H	I	J	K	L	M
	ID	YE	MONT	DA	ST_BASE	EXPECT_	REMAR	ST_1	ST_2	ST_3	ST_4	ST_5
0	B2BTUR01_T_03:30				241,00		Altitude	235,00	670,00	203,00	210,00	749,00
0	B2BZAB01_T_03:30						st_1, di	11,58				
0	B1PROT01_T_03:30						st_2, di		36,85			
0	O3PRER01_T_03:30						st_3, di			59,12		
0	O2OLOM01_T_03:30						st_4, di				62,88	
0	O1CERV01_T_03:30						st_5, di					91,95
0	**B2BTUR01_T_03:30**	**2006**	**6**	**25**	**27,30**	**17,28**		**17,30**	**16,10**	**15,50**	**15,80**	**16,10**
5	B2BTUR01_T_03:45				241,00		Altitude	235,00	670,00	203,00	210,00	749,00
5	B2BZAB01_T_03:45						st_1, di	11,58				
5	B1PROT01_T_03:45						st_2, di		36,85			
5	O3PRER01_T_03:45						st_3, di			59,12		
5	O2OLOM01_T_03:45						st_4, di				62,88	
5	O1CERV01_T_03:45						st_5, di					91,95
5	**B2BTUR01_T_03:45**	**2006**	**6**	**25**	**26,50**	**17,26**		**17,30**	**16,30**	**15,80**	**15,60**	**16,20**
0	B2BTUR01_T_04:00				241,00		Altitude	235,00	670,00	203,00	210,00	749,00
0	B2BZAB01_T_04:00						st_1, di	11,58				
0	B1PROT01_T_04:00						st_2, di		36,85			
0	O3PRER01_T_04:00						st_3, di			59,12		
0	O2OLOM01_T_04:00						st_4, di				62,88	
0	O1CERV01_T_04:00						st_5, di					91,95
0	**B2BTUR01_T_04:00**	**2006**	**6**	**25**	**26,30**	**17,41**		**17,30**	**16,50**	**16,50**	**15,90**	**16,20**
0	B2BTUR01_T_05:00				241,00		Altitude	235,00	670,00	203,00	210,00	749,00
0	B2BZAB01_T_05:00						st_1, di	11,58				
0	B1PROT01_T_05:00						st_2, di		36,85			
0	O3PRER01_T_05:00						st_3, di			59,12		
0	O2OLOM01_T_05:00						st_4, di				62,88	
0	O1CERV01_T_05:00						st_5, di					91,95
0	**B2BTUR01_T_05:00**	**2006**	**6**	**25**	**24,70**	**17,52**		**17,30**	**17,20**	**17,30**	**16,30**	**17,20**

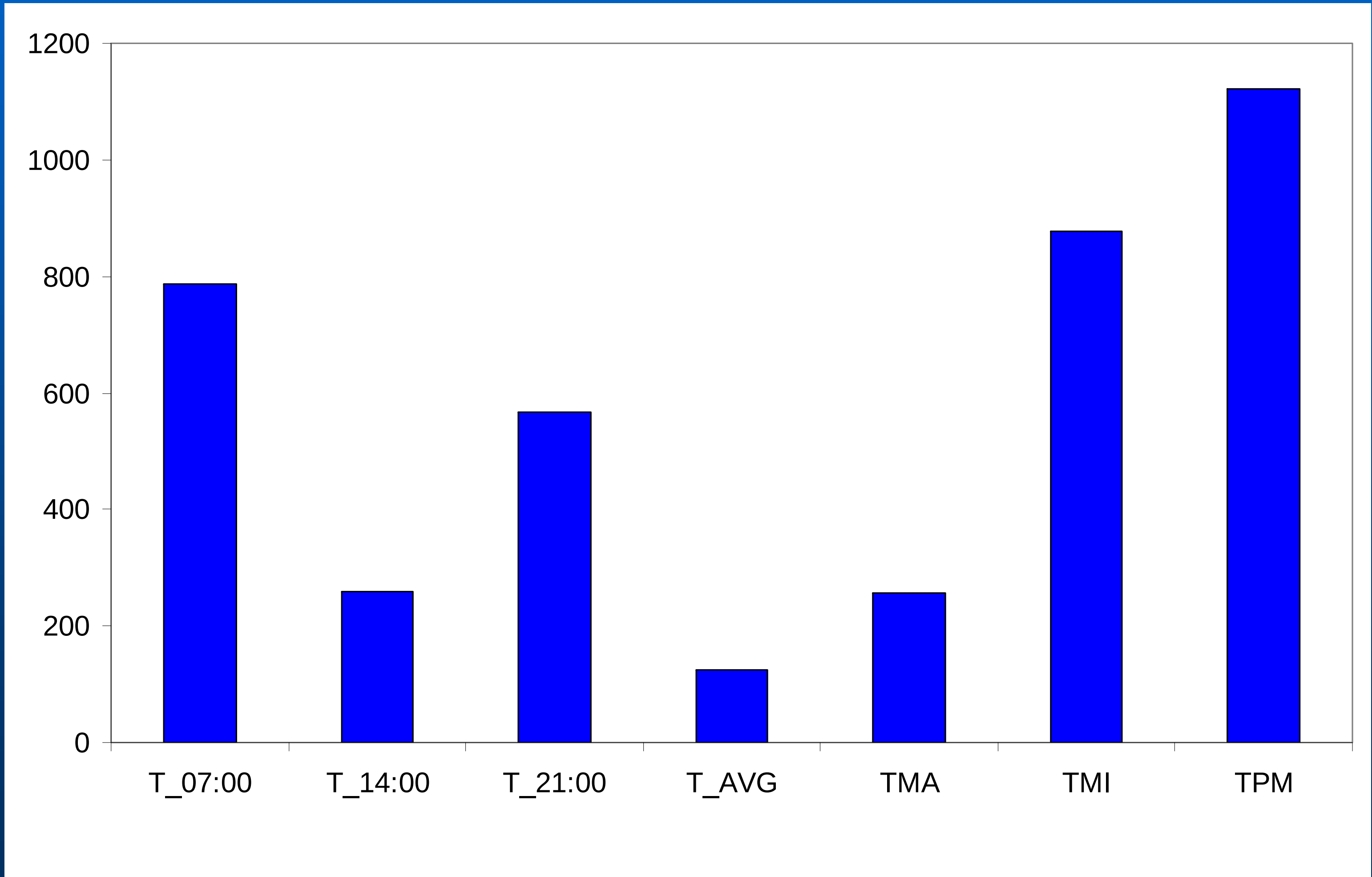
Quality control

- Run for period 1961-2007, daily data (measured values in observation hours)
- All stations (200 climatological stations, 800 precipitation stations)
- All meteorological elements (T, TMA, TMI, TPM, SRA, SCE, SNO, E, RV, H, F) – parameters set individually

- Historical records will follow now

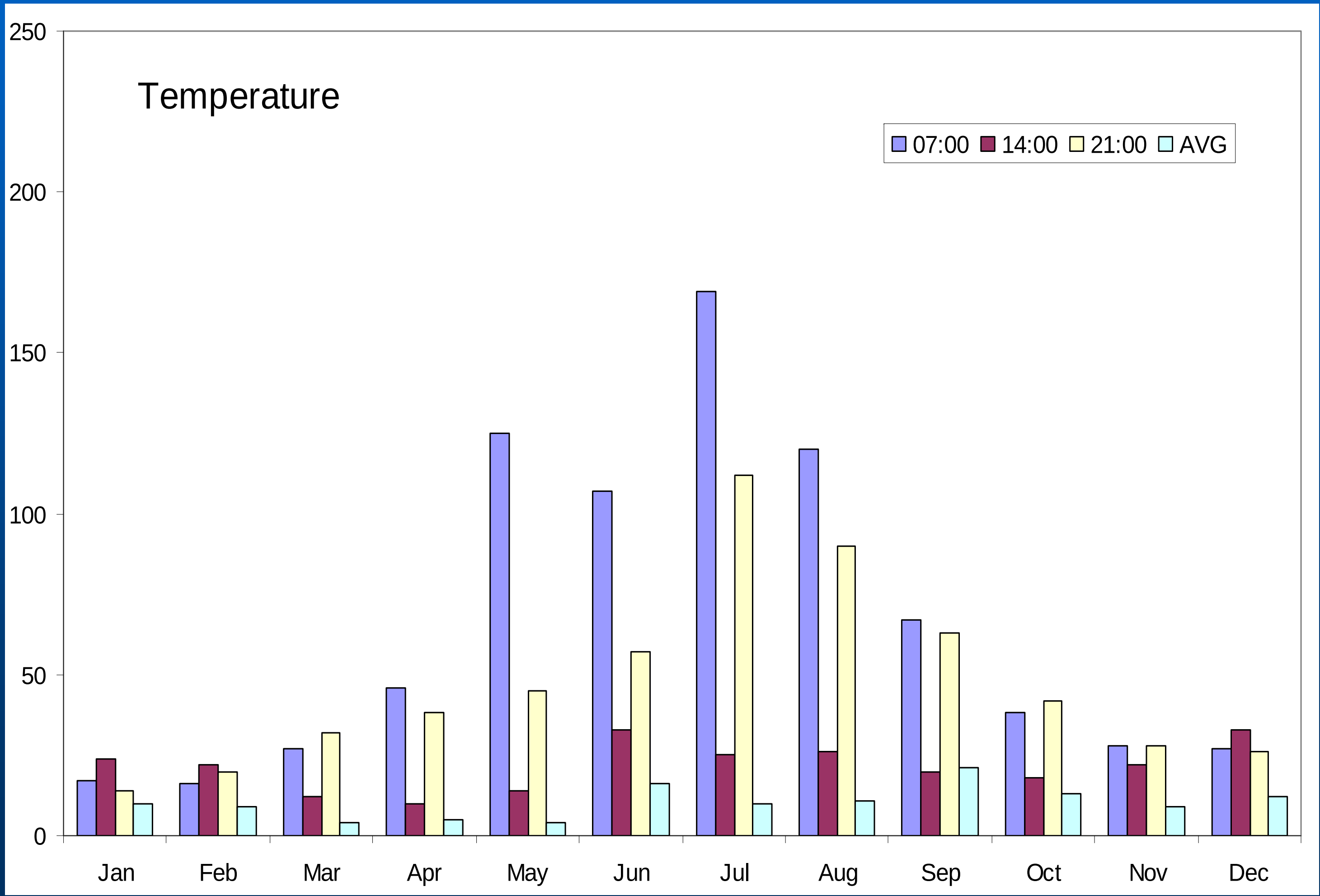
Air temperature, number of outliers 1961-2007, from 3.431.000 station-days

T – air temperature at obs. hour, TMA – daily maximum temp., TMI – daily min. temp., TPM – daily ground minimum temp.



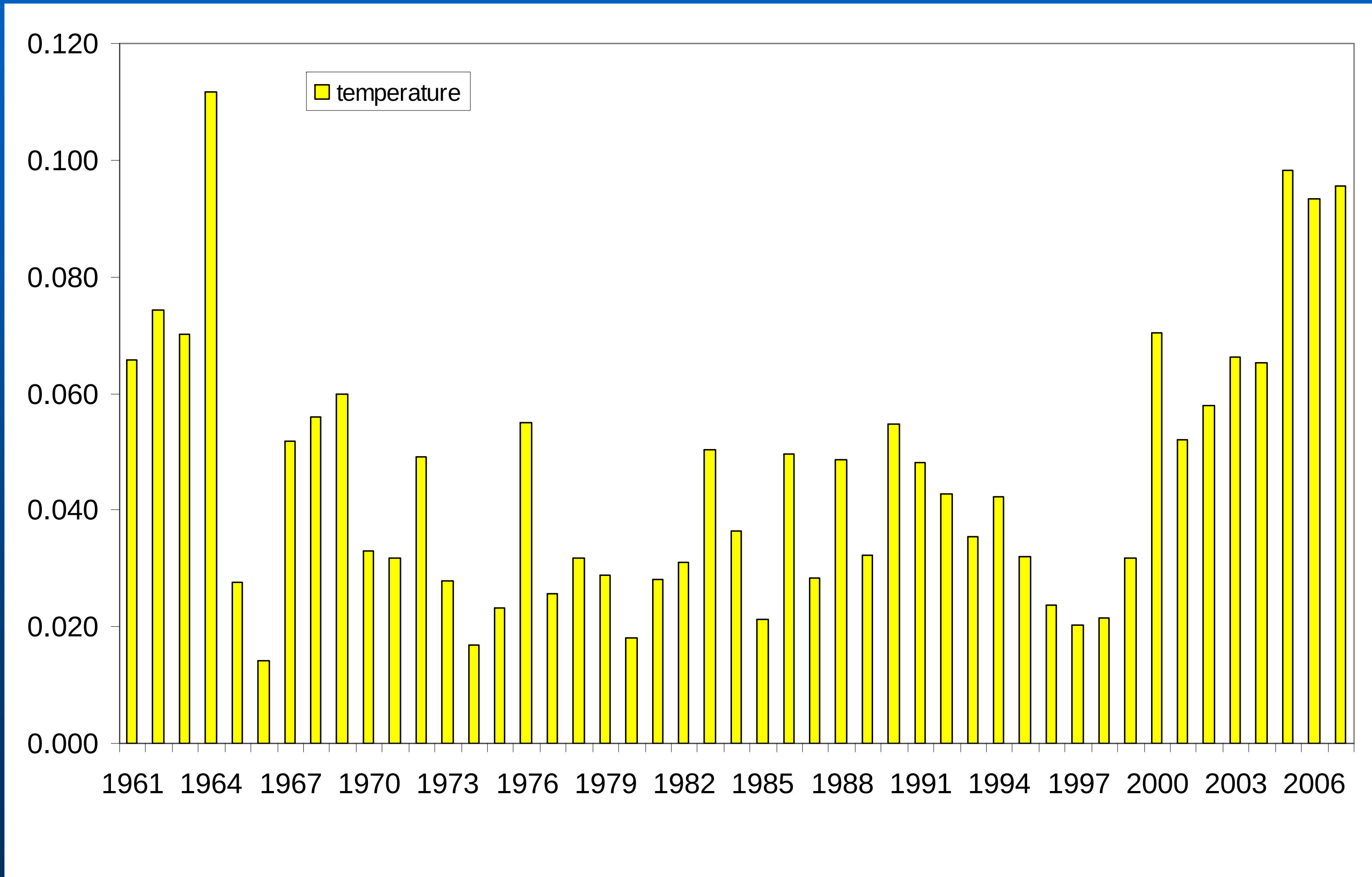
Air temperature, number of outliers 1961-2007, from 3.431.000 station-days

Air temperature at obs. hour, AVG – daily average temp.



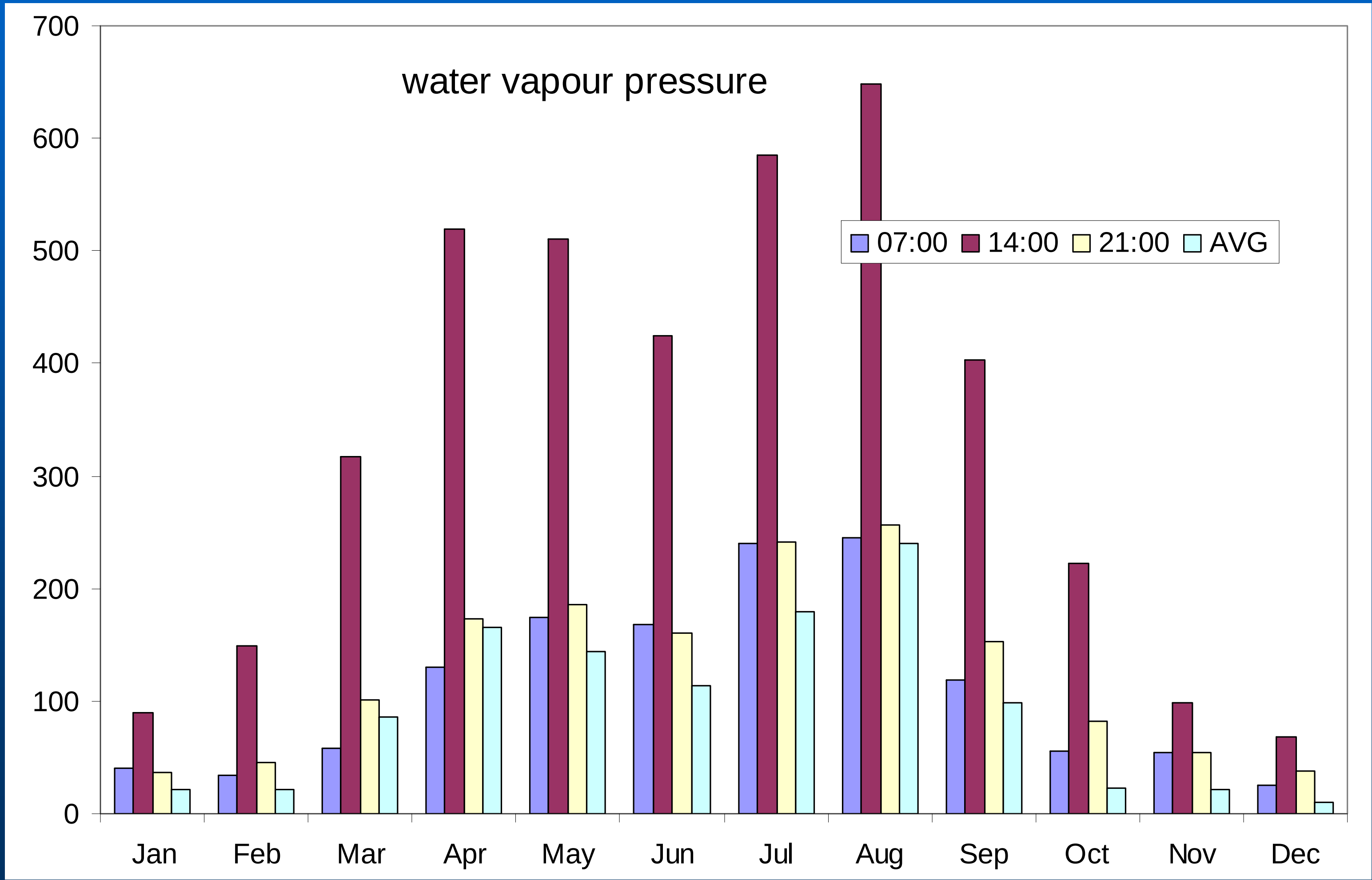
Air temperature, number of outliers 1961-2007,

Number of outliers per one station (all observation hours, AVG)



Water vapor pressure, number of outliers 1961-2007, from 3.431.000 station-days

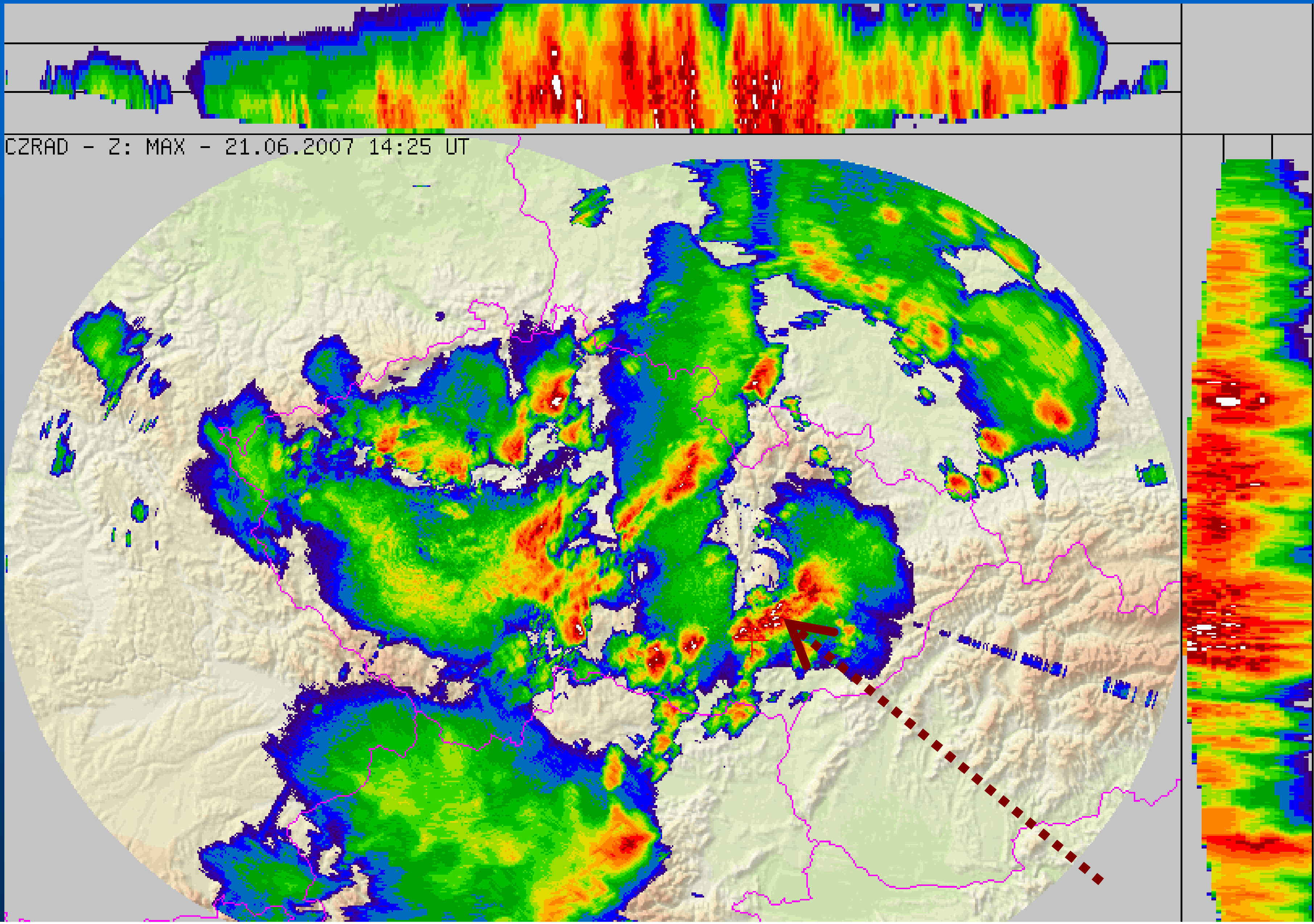
Water vapor pressure at obs. hour, AVG – daily average



Problematic detections - heavy rainfall

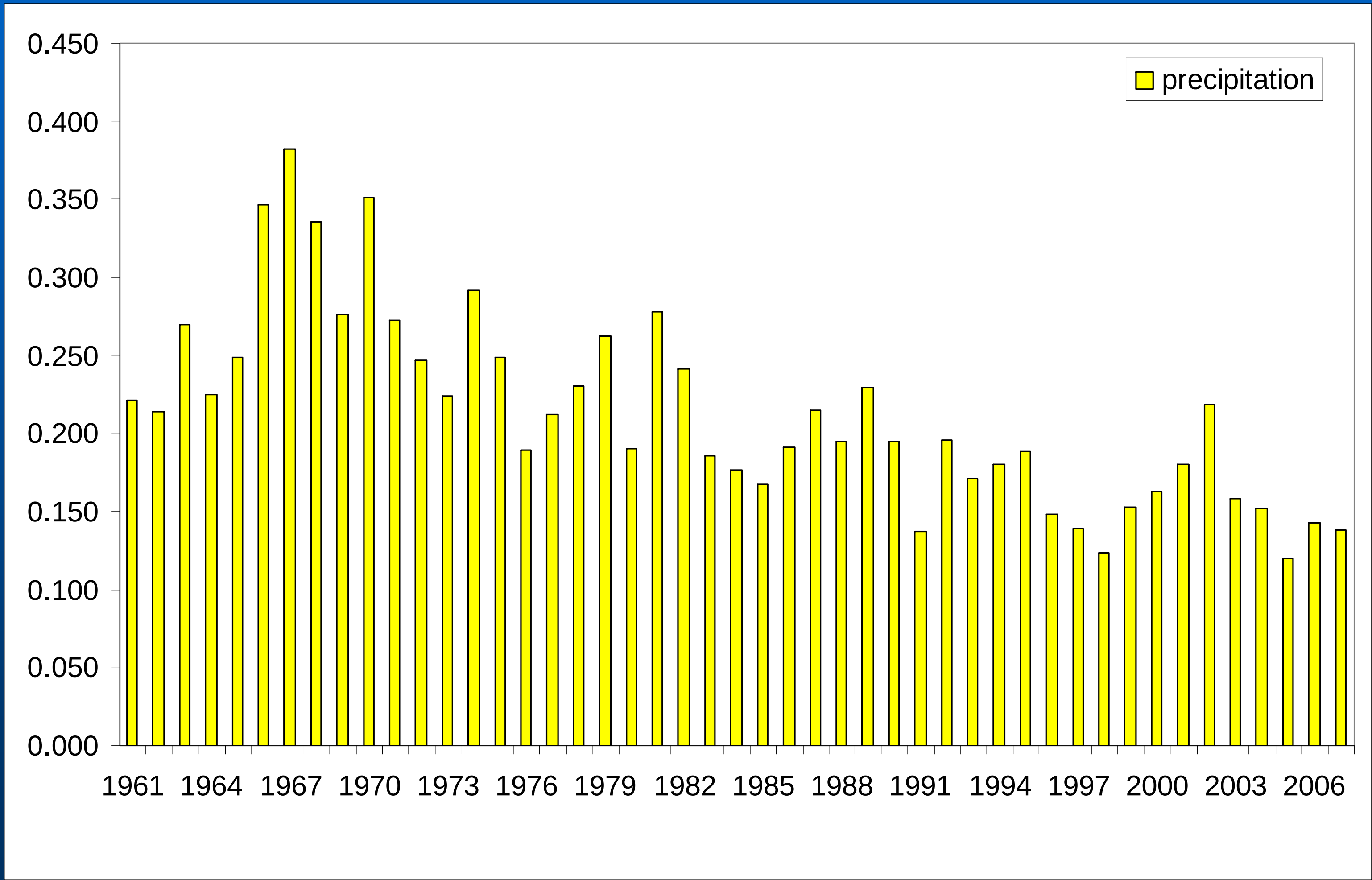
ID	YEAR	MONTH	DAY	ST_BASE	EXPECT_VAL	REMARK	ST_1	ST_2	ST_3	ST_4	ST_5	D
B2BTUR01_SRA3H_16:00				241,00		Altitude	235,00	670,00	203,00	210,00	749,00	
B2BZAB01_SRA3H_16:00						st_1, di	11,58					
B1PROT01_SRA3H_16:00						st_2, di		36,85				
O3PRER01_SRA3H_16:00						st_3, di			59,12			
O2OLOM01_SRA3H_16:00						st_4, di				62,88		
O1CERV01_SRA3H_16:00						st_5, di					91,95	
B2BTUR01_SRA3H_16:00	**2005**	**4**	**6**	**10,00**	**1,47**		**1,50**	**0,00**	**0,20**	**0,00**	**0,30**	
B2BTUR01_SRA3H_16:00	**2006**	**7**	**14**	**8,70**	**0,32**		**0,30**	**0,50**	**0,20**	**0,00**		
B2BTUR01_SRA3H_16:00	**2006**	**8**	**13**	**7,00**	**0,13**		**0,10**	**0,70**	**0,00**	**0,00**	**0,00**	
B2BTUR01_SRA3H_16:00	**2007**	**6**	**21**	**21,70**	**0,66**		**0,70**		**3,00**	**4,70**	**0,10**	
B2BTUR01_SRA3H_16:00	**2007**	**7**	**11**	**9,40**	**0,04**		**0,00**	**0,60**	**0,00**	**0,00**	**1,40**	
B2BTUR01_SRA3H_19:00				241,00		Altitude	235,00	670,00	203,00	210,00	749,00	
B2BZAB01_SRA3H_19:00						st_1, di	11,58					
B1PROT01_SRA3H_19:00						st_2, di		36,85				
O3PRER01_SRA3H_19:00						st_3, di			59,12			
O2OLOM01_SRA3H_19:00						st_4, di				62,88		
O1CERV01_SRA3H_19:00						st_5, di					91,95	
B2BTUR01_SRA3H_19:00	**2005**	**5**	**23**	**8,00**	**0,03**		**0,00**	**0,20**	**0,00**	**0,00**	**0,00**	
B2BTUR01_SRA3H_19:00	**2005**	**7**	**23**	**7,00**	**1,73**		**1,80**	**1,00**	**0,00**	**0,00**	**0,00**	
B2BTUR01_SRA3H_19:00	**2006**	**5**	**13**	**4,40**	**0,02**		**0,00**	**0,00**	**0,00**	**0,00**	**0,10**	
B2BTUR01_SRA3H_19:00	**2006**	**7**	**8**	**13,70**	**-0,04**		**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	
B2BTUR01_SRA3H_19:00	**2006**	**8**	**7**	**5,90**	**0,25**		**0,20**	**0,90**	**0,90**	**0,00**	**0,00**	
B2BTUR01_SRA3H_19:00	**2007**	**1**	**1**	**3,40**	**0,69**		**0,70**	**0,60**	**0,30**	**0,00**	**1,10**	
B2BTUR01_SRA3H_19:00	**2007**	**6**	**14**	**9,00**	**0,03**		**0,00**	**0,00**	**0,30**	**0,00**	**0,00**	
B2BTUR01_SRA3H_22:00				241,00		Altitude	235,00	670,00	203,00	210,00	749,00	
B2BZAB01_SRA3H_22:00						st_1, di	11,58					
B1PROT01_SRA3H_22:00						st_2, di		36,85				
O3PRER01_SRA3H_22:00						st_3, di			59,12			
O2OLOM01_SRA3H_22:00						st_4, di				62,88		
O1CERV01_SRA3H_22:00						st_5, di					91,95	
B2BTUR01_SRA3H_22:00	**2005**	**4**	**25**	**1,90**	**0,39**		**0,40**	**0,10**	**0,20**	**0,00**	**0,10**	
B2BTUR01_SRA3H_22:00	**2005**	**6**	**25**	**20,00**	**7,68**		**7,70**		**8,00**	**8,60**	**8,00**	

Problematic detections (heavy rainfall), Radar information



Precipitation, number of outliers 1961-2007,

Number of outliers per one station



Presented method can be further applied for

- **Filling missing values (the "expected" value)**
- **Calculation of technical series (e.g. for grid points - to be used for RCM validations or correction,** EC FP6 project CECILIA**), …**

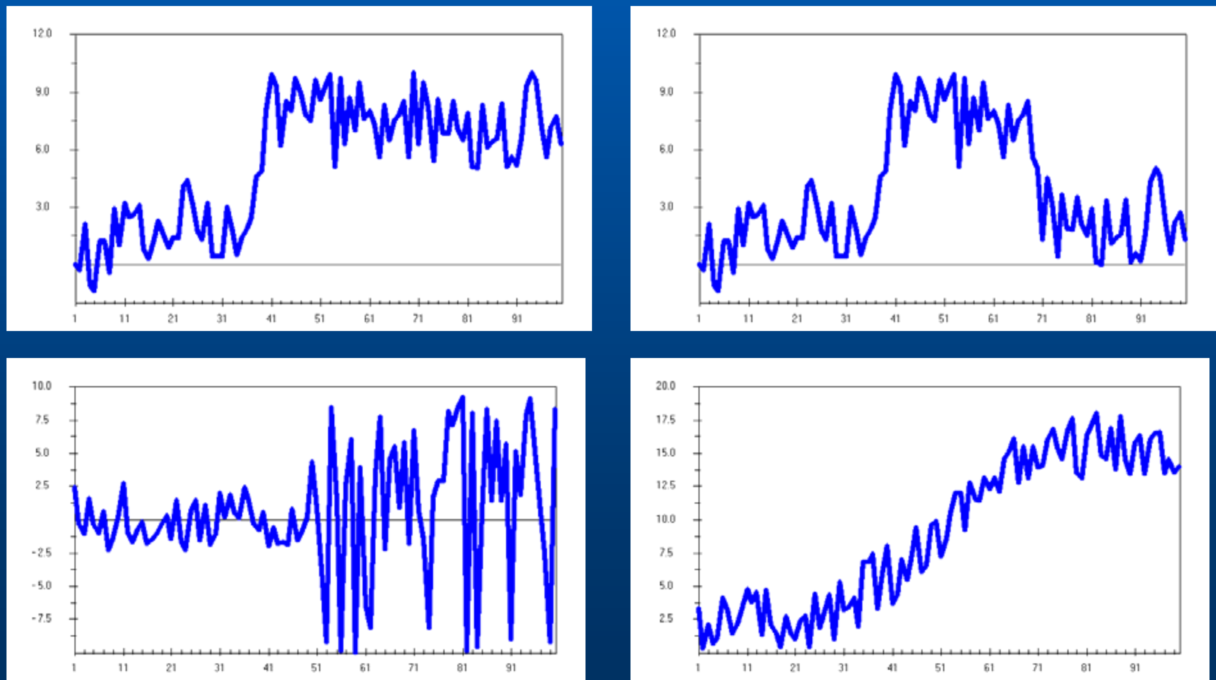
Remarks for QC

- **Only combination of several methods for outliers detection leads to satisfying results ("real" outliers detection, supressing fault detection -> Emsemble approach)**
- **Parameters (settings) has to be found individually for each meteorological element, maybe also region (terrain complexity) and part of a year (noticeable annual cycle in number of outliers)**
- **it is important to use measured value (e.g. from observation hours) - outliers are masked in daily average (and even more in monthly or annual ones)**
- Errors found in all elements and investigated countries (AT, CZ, SK, HU)

Homogenization

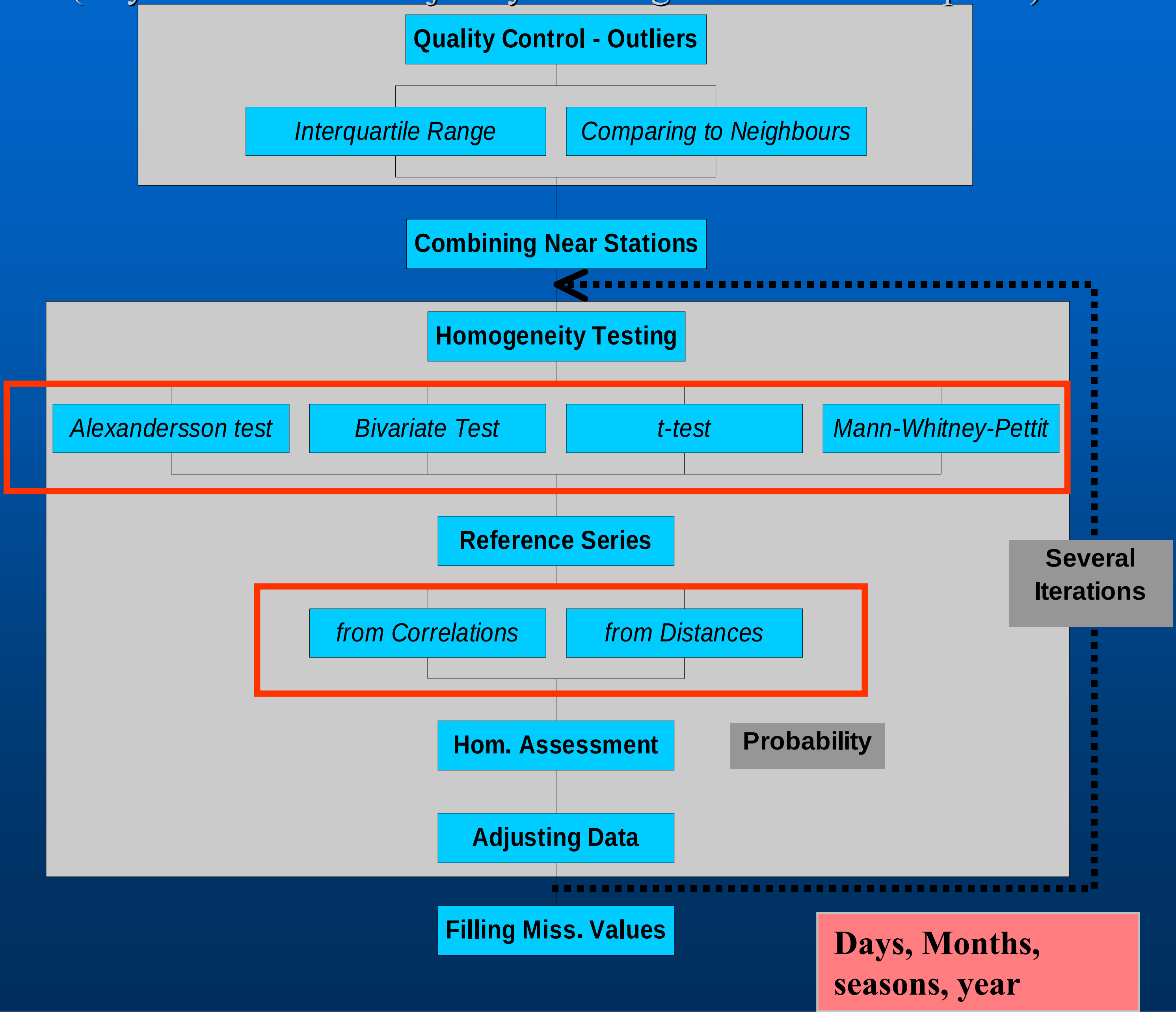
- **Change of measuring conditions**

  ⟶ **inhomogeneities**



How to increase number of test results

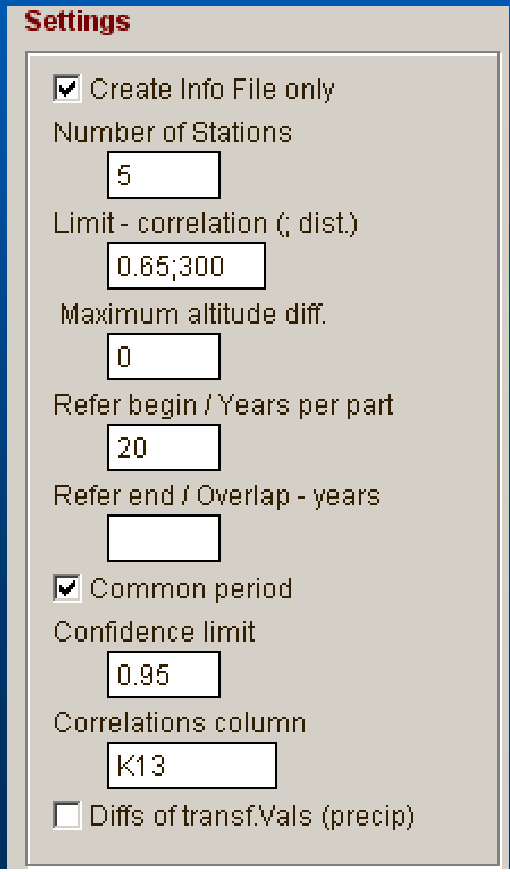
(way to automatize - objectify inhomogeneities detection phase)



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

- **Various outputs created for better inhomogeneities assessment**
- **Combining results with information from metadata whenever possible**
- **Decision about „undoubted“ inhomogeneities (without metadata) – coincidence of test results**

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:

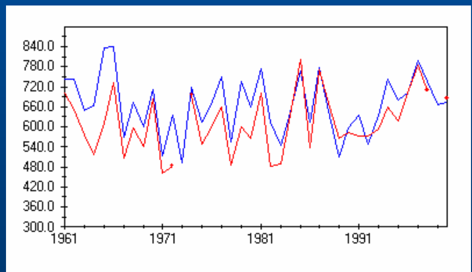
Begin	End	Length	InHomogeneity	Number	% detected inhom	% possible inhom	End	Missing
1911	1950	40		140	100	120		
			1927	60	43	51		
			1926	37	26	32		
			1928	9	6	8		4
			1937	7	5	6		
			1922	4	3	3		
			1935	4	3	3		
			1918	3	2	3		
			1930	3	2	3		
			1939	3	2	3		
			1940	3	2	3		2
			1938	2	1	2		
			1913	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		

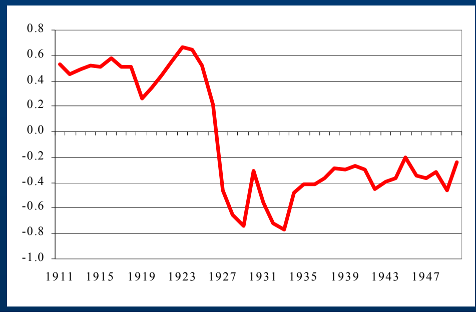
Summed numbers of detections for individual years

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	E	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





Inhomogeneities detection and correction

- **Detection – for months, seasons, year**
- **Correction – daily,** **for each months separately**

Adjusting daily values for inhomogeneities,

„delta“ method

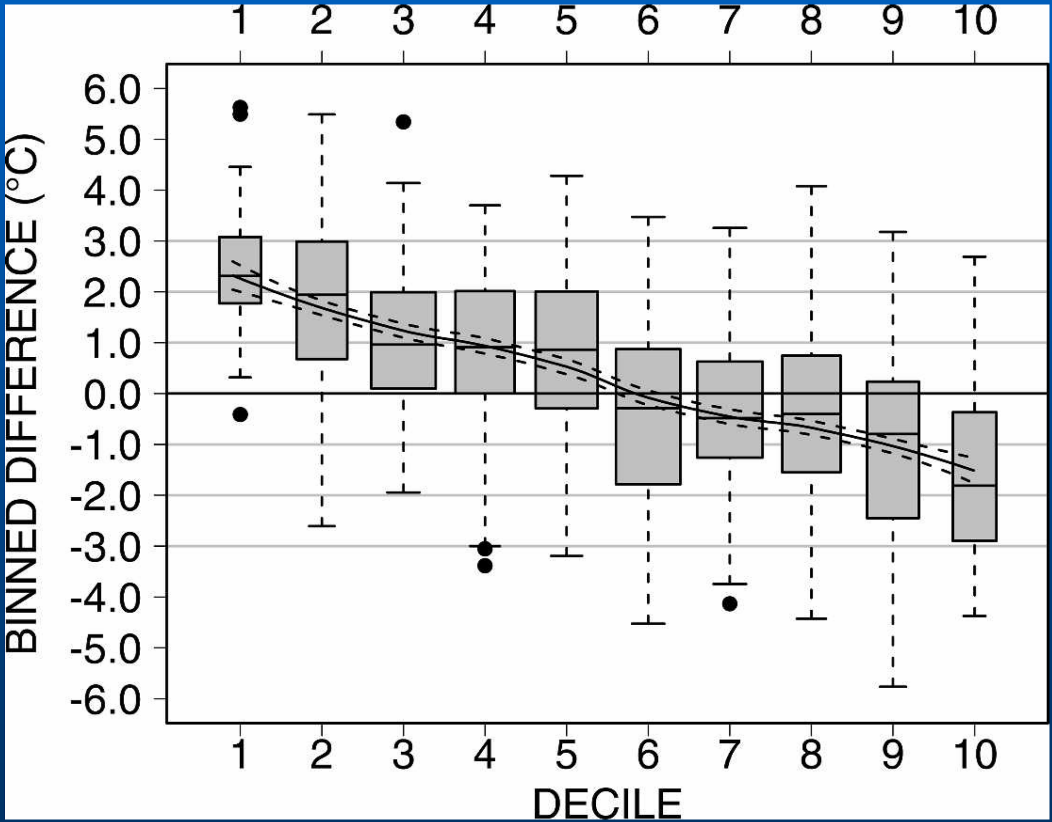
- **interpolation of monthly factors**
  - **MASH**
  - **Vincent *et al* (2002)**
- Is it natural that station changes has the same effect upon low and high extremes …?

Adjusting daily values for inhomogeneities,

Variable correction

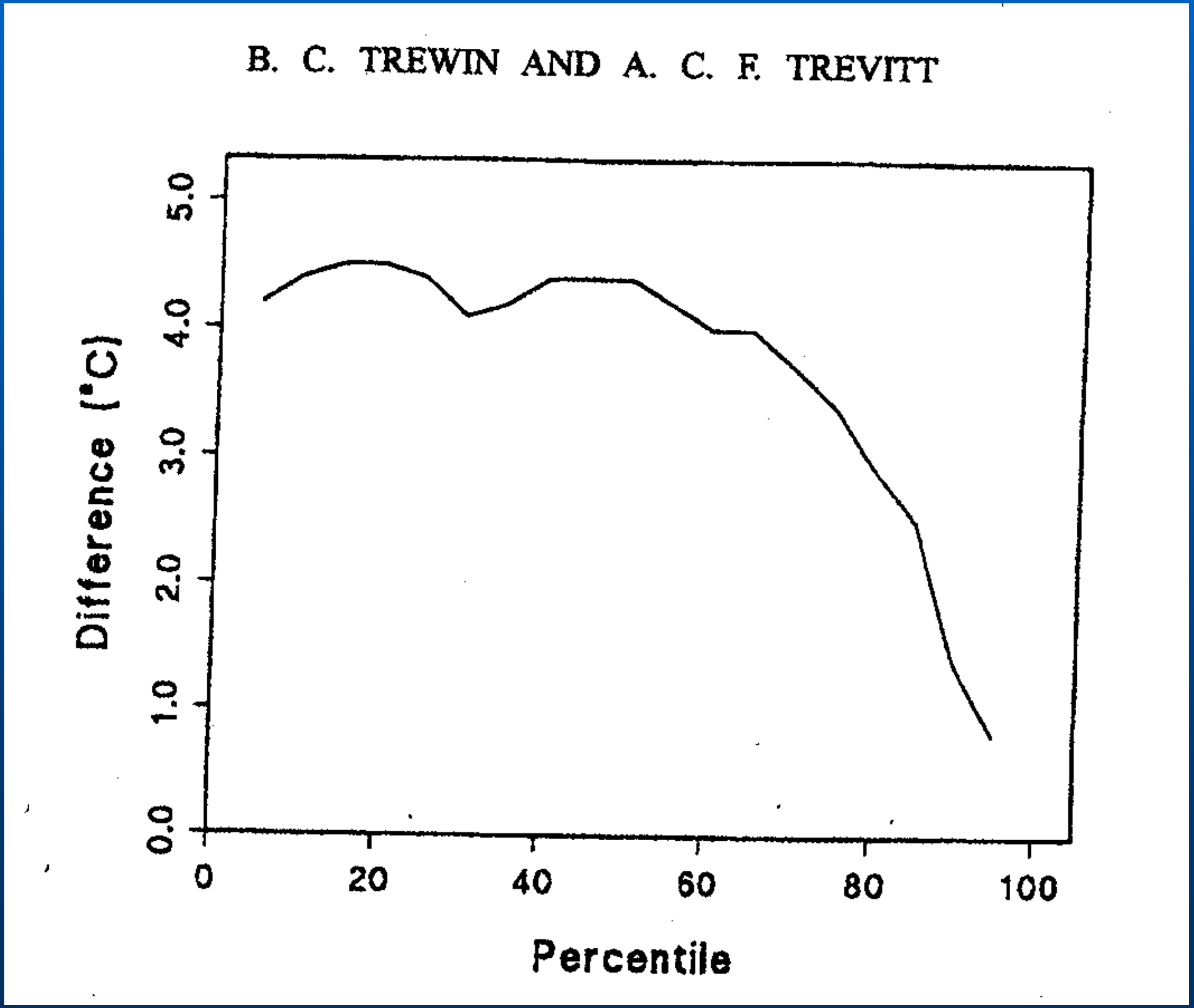
- **E.g.**
  - **Higher Order Moments (HOM), by Della Marta and Wanner (2006)**
  - **Two phase non-linear regression by Mestre (SPLIDHOM)**
  - our own percentile approach (similar to Déqué……)

Variable correction,
The higher-order moments method



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

Variable correction



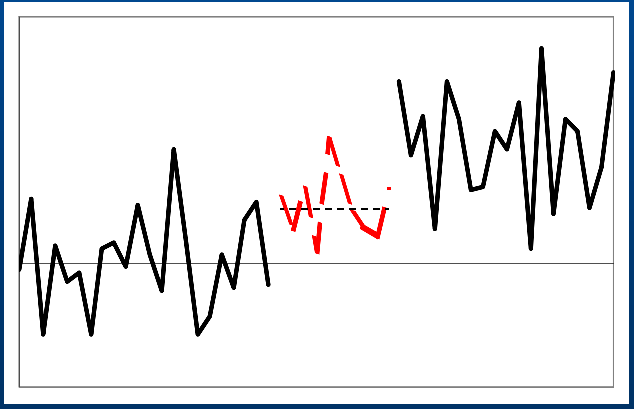
1996

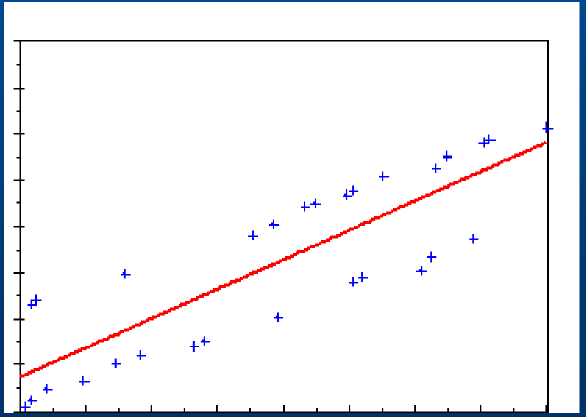
Iterative homogeneity testing

- **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: <u>in next iteration neighbours</u> should be already <u>homogenized</u>**

Filling missing values

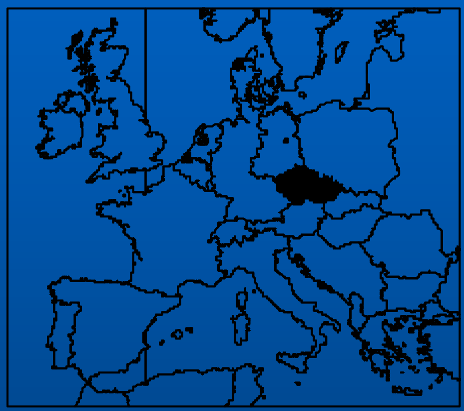
- **Before homogenization: influence on right inhomogeneity detection**
- **After homogenization: more precise - data are not influenced by possible shifts in the series**

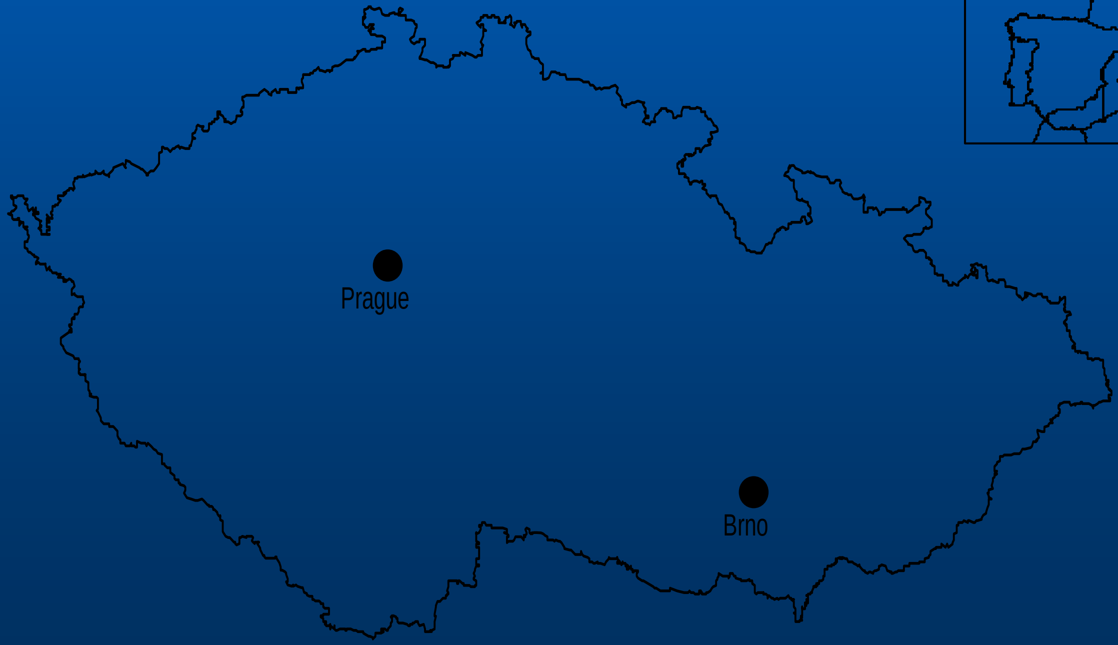




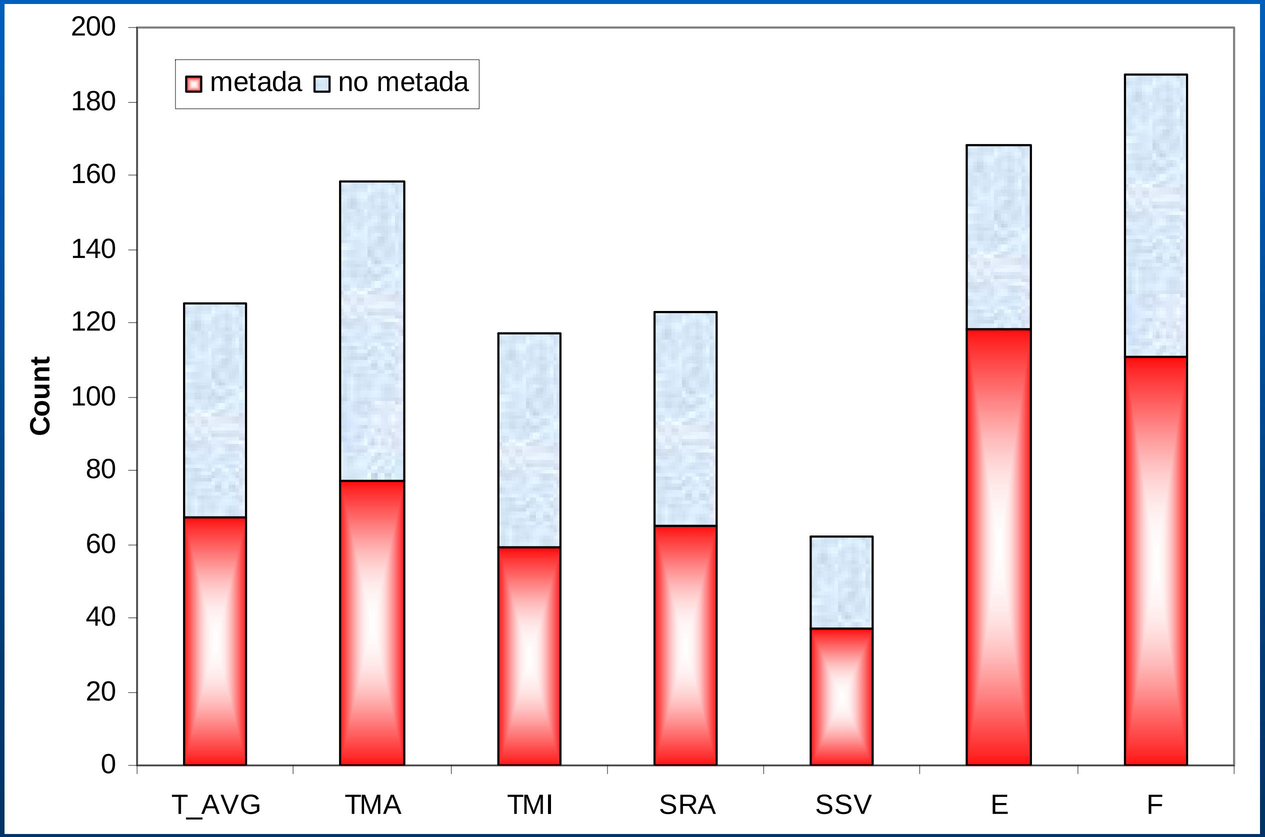
Dependence of tested series on reference series

Homogenization
of the series in the Czech Republic



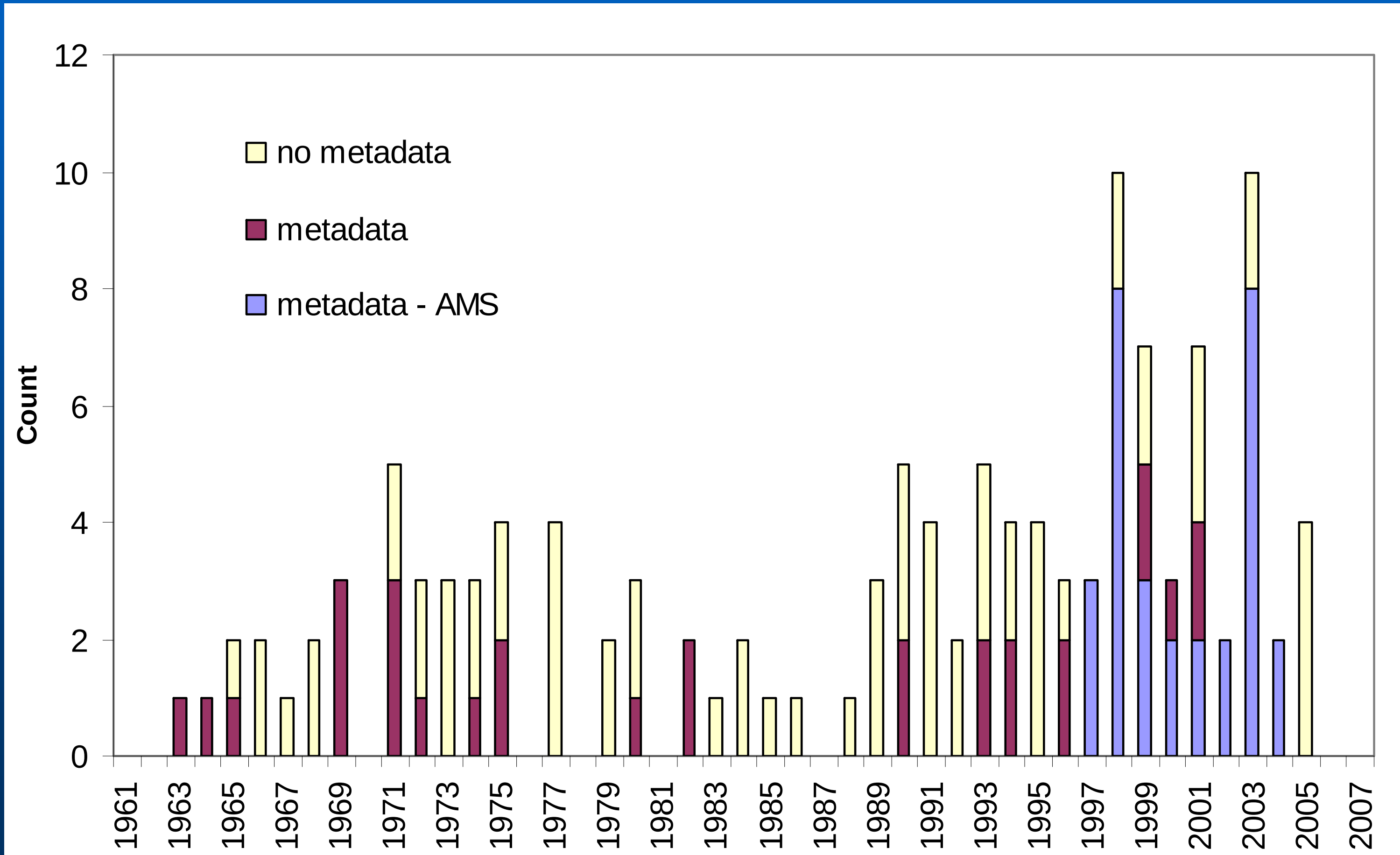


Number of inhomogeneities explained by metadata

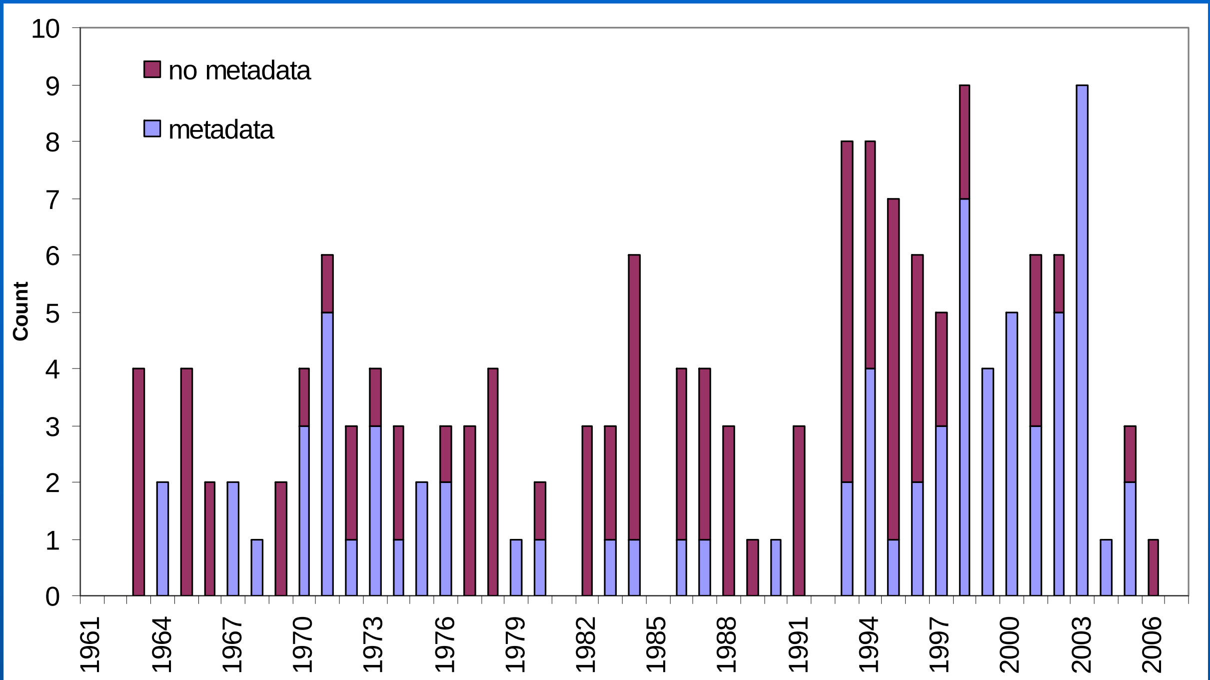


T – air temperature, TMA – maximum temperature, TMI – minimum temperature, SRA – precipitation, SSV – sunshine duration, E – water vapour pressure, F – wind speed

Number of inhomogeneities explained by metadata, T_AVG

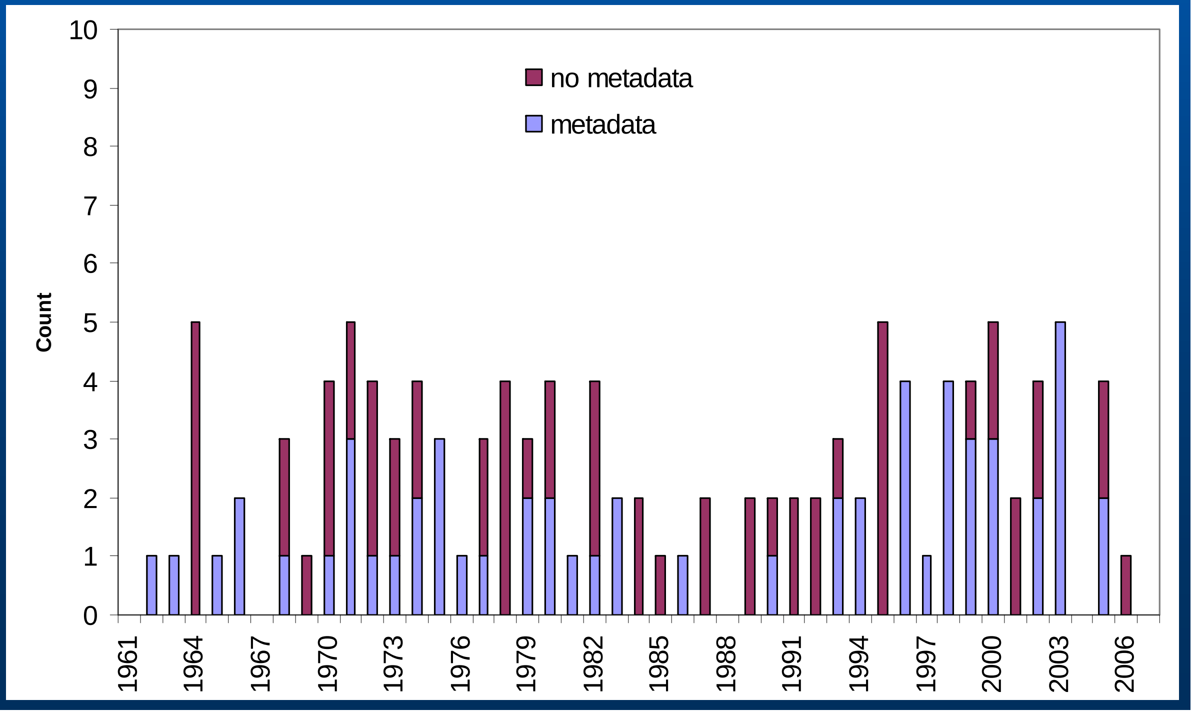


Tmax



no metadata
metadata
Count
0
1
2
3
4
5
6
7
8
9
10
1961
1964
1967
1970
1973
1976
1979
1982
1985
1988
1991
1994
1997
2000
2003
2006

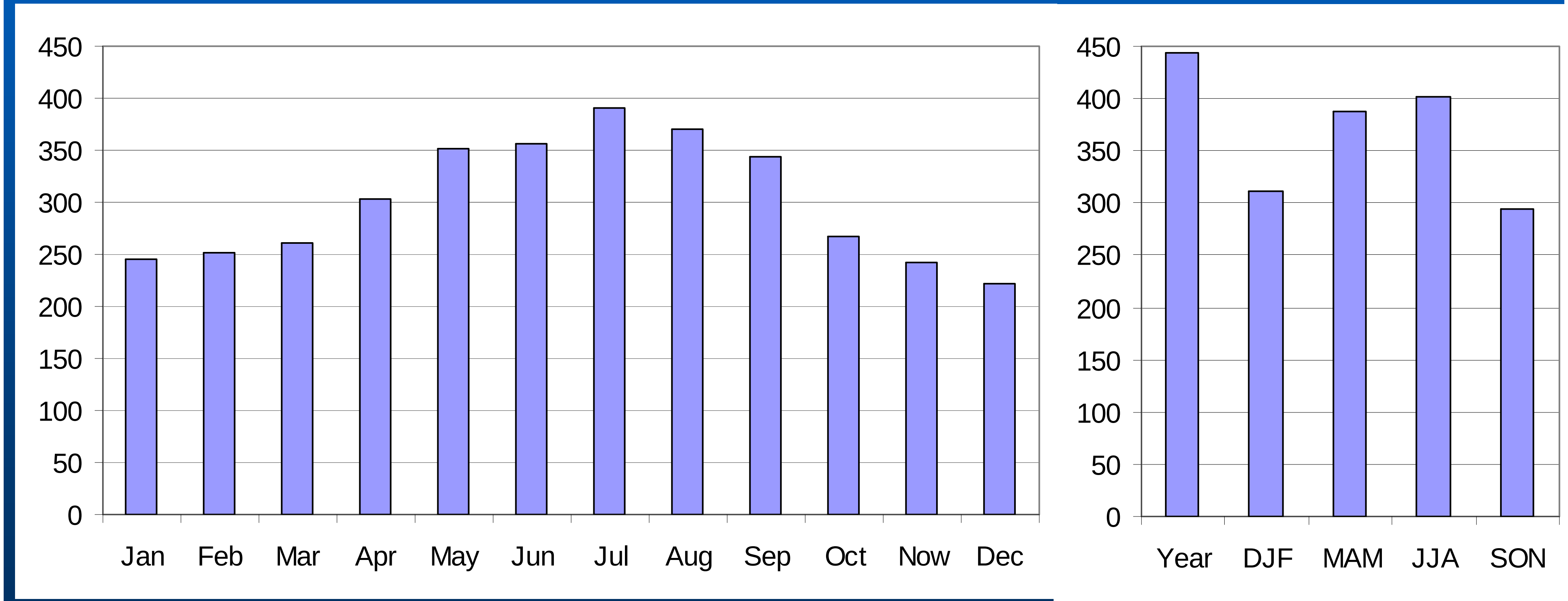
Tmin



no metadata
metadata
Count
0
1
2
3
4
5
6
7
8
9
10
1961
1964
1967
1970
1973
1976
1979
1982
1985
1988
1991
1994
1997
2000
2003
2006

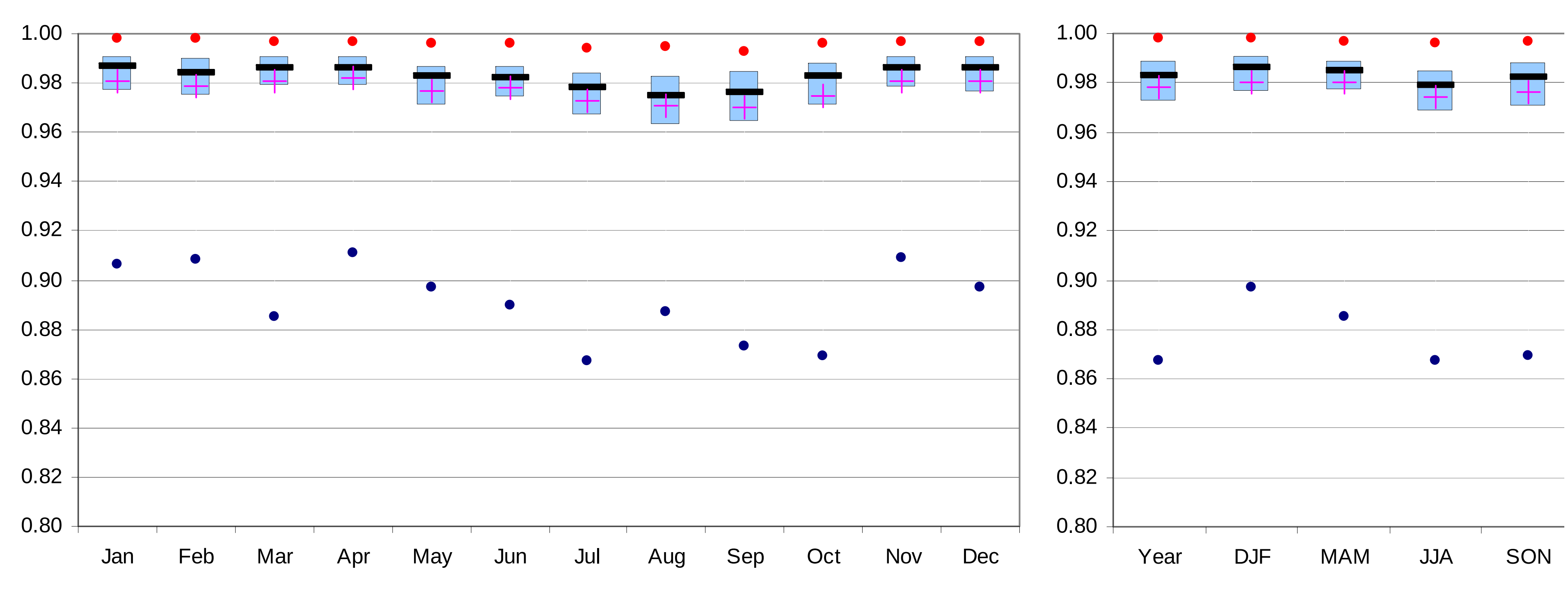
Homogeneity testing results
Air temperature

- **Number of detected inhomogeneities** **(significant, 0.05)**



Correlations between tested and reference series, daily values

Air temperature



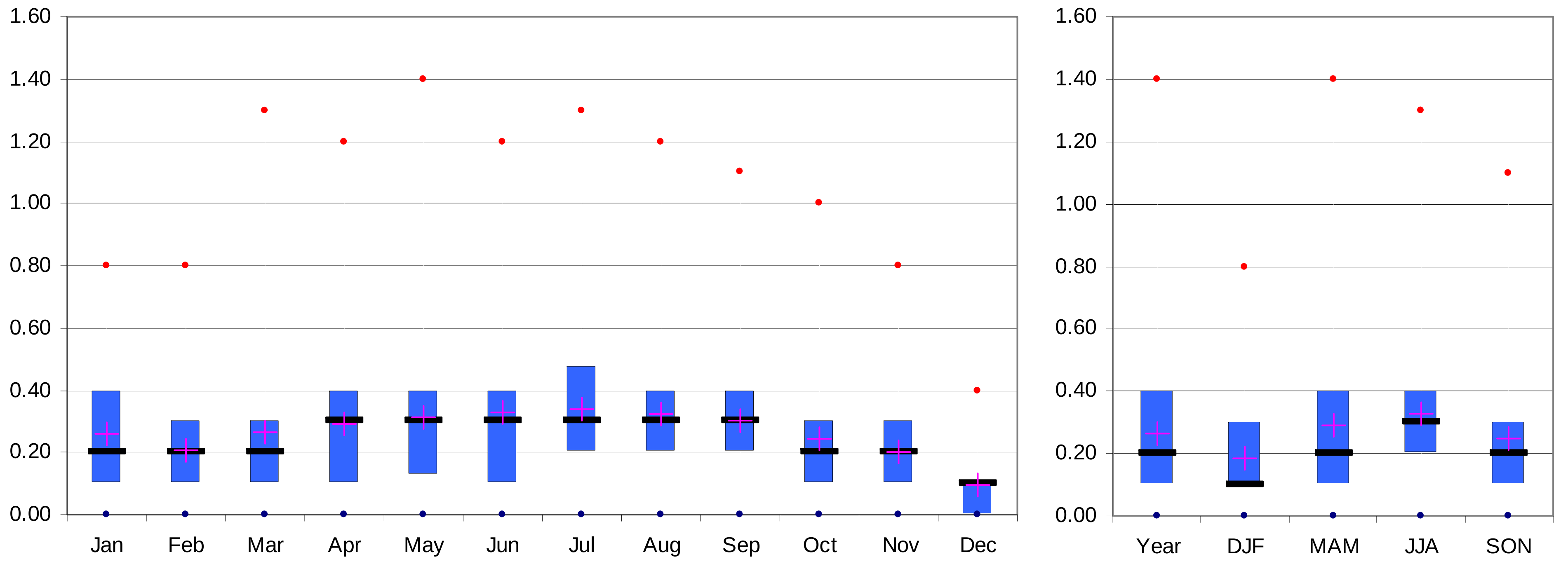
Boxplots:

- Median, average
- Upper and lower quartiles
- minimum and maximum value

(for 115 stations)

Adjustments, monthly averages of abs. values

Air temperature



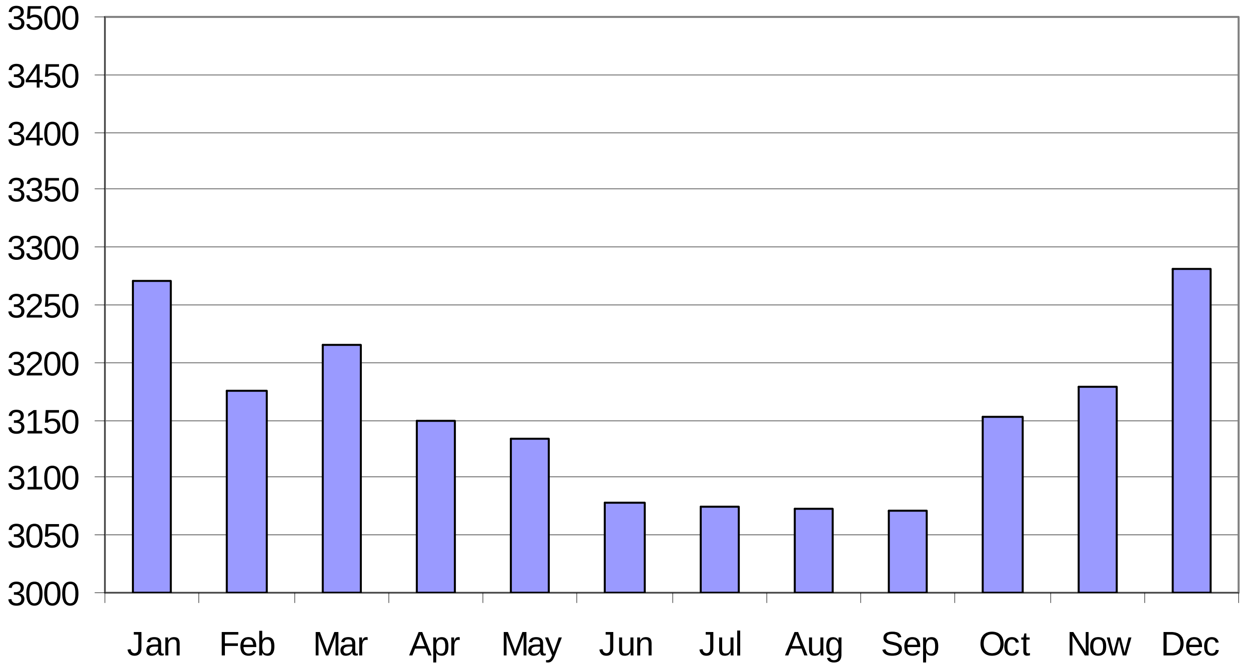
Boxplots:

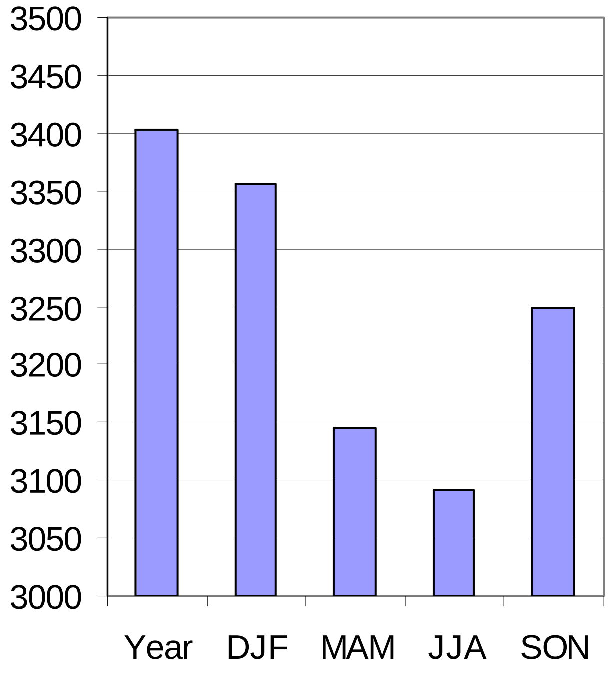
- Median, average
- Upper and lower quartiles
- minimum and maximum value

(for 115 stations)

Homogeneity testing results Precipitation

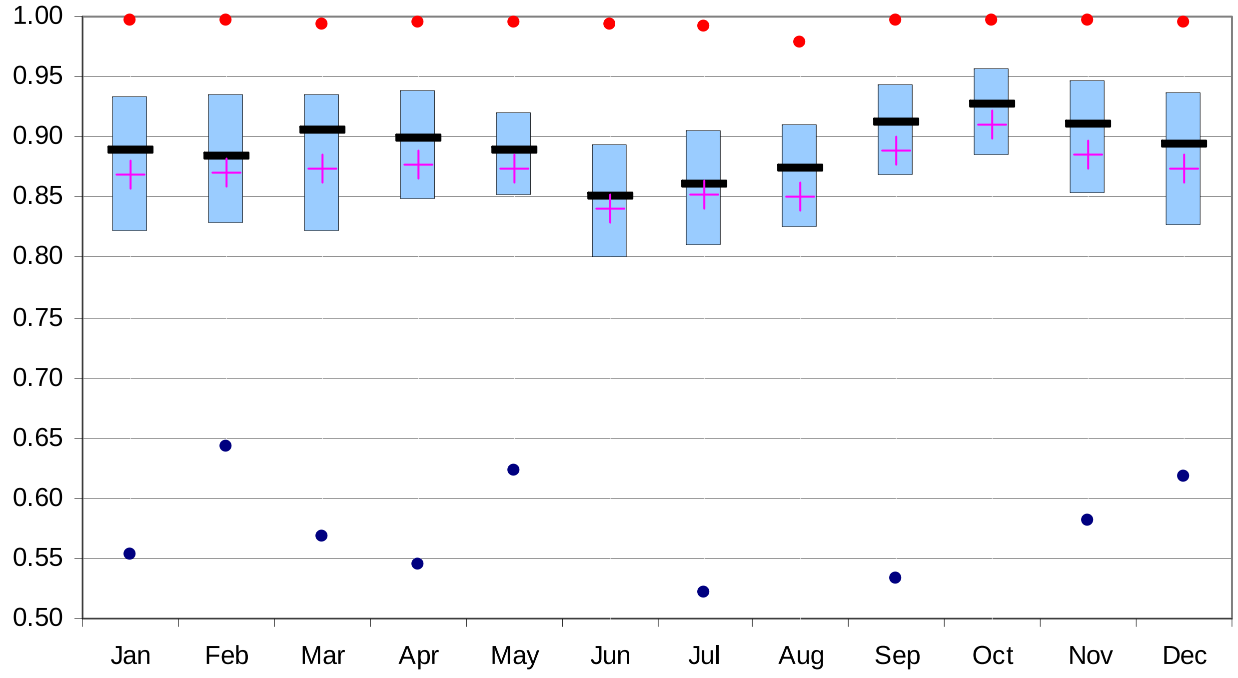
- **Number of detected inhomogeneities (significant, 0.05)**

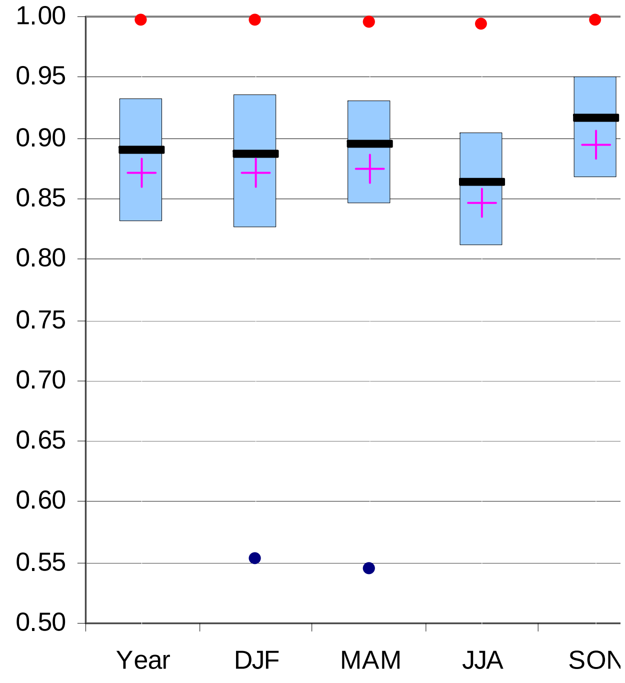




Correlations between tested and reference series, daily values

Precipitation



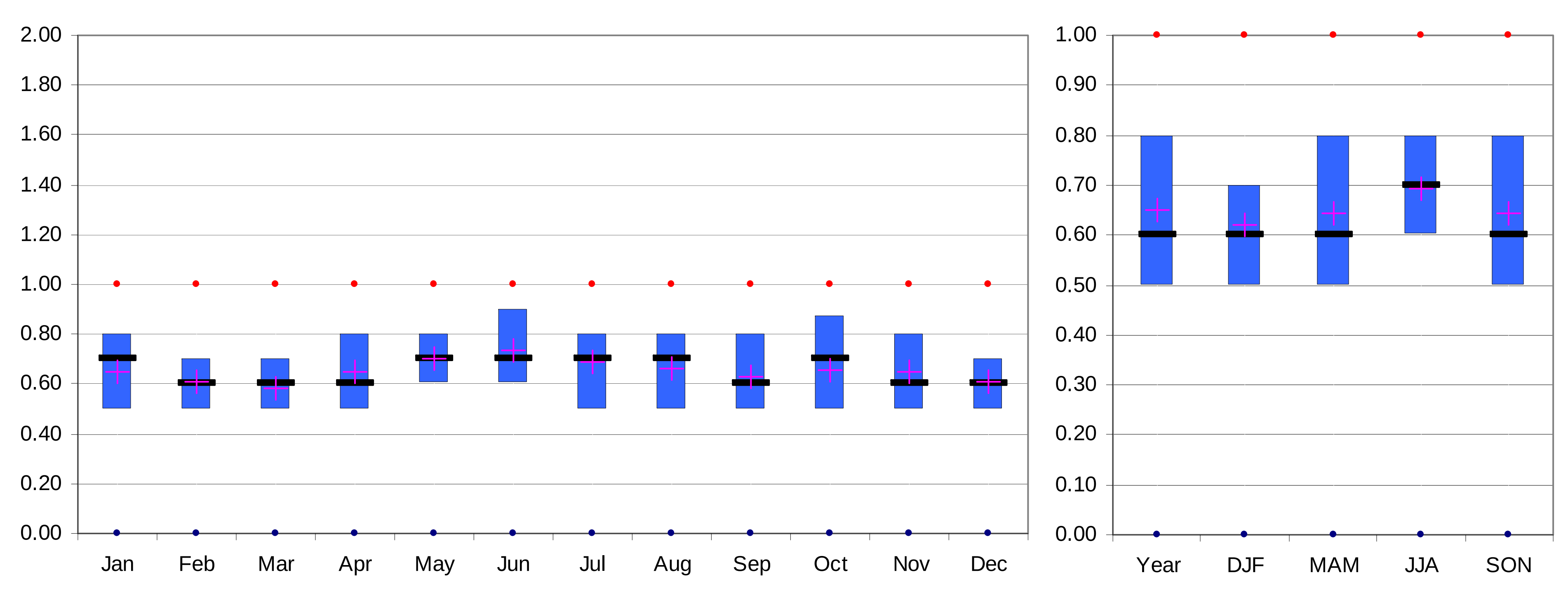


Boxplots:
- Median, average
- Upper and lower quartiles
- minimum and maximum value

(for 121 stations)

Adjustments, monthly averages of quotines < 1

Precipitation



Boxplots:

- Median, average
- Upper and lower quartiles
- minimum and maximum value

(for 115 stations)

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, prevailing 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
Conlusions

- - „Ensemble“ approach to homogenization (combining information from different statistical tests, time frames, overlapping periods, reference series, meteorological elements, …)

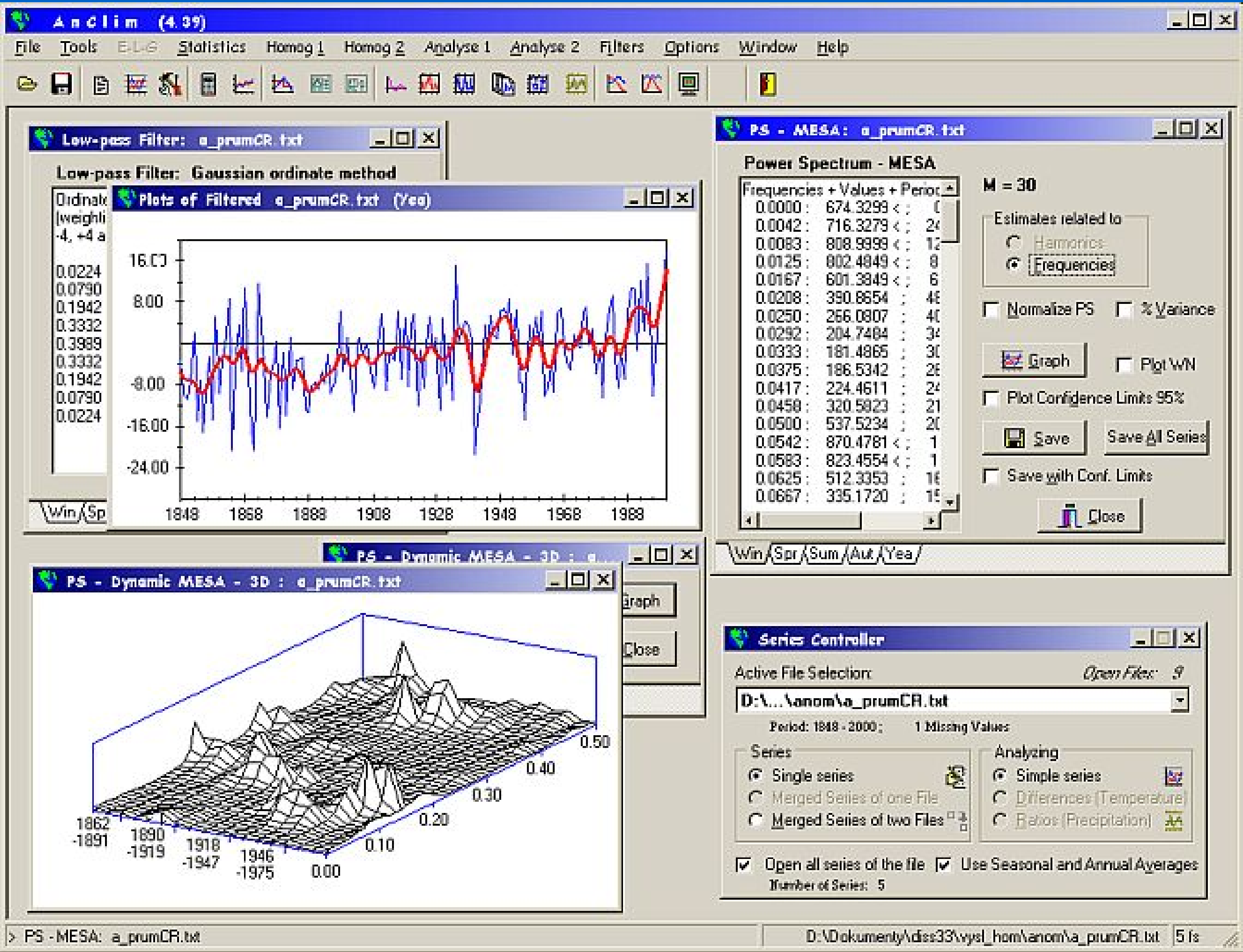
  - more information for inhomogeneities assessment – higher quality of homogenization in case metadata are incomplete
- annual cycle of inhomogeneities, adjustments, …

Software used for data processing

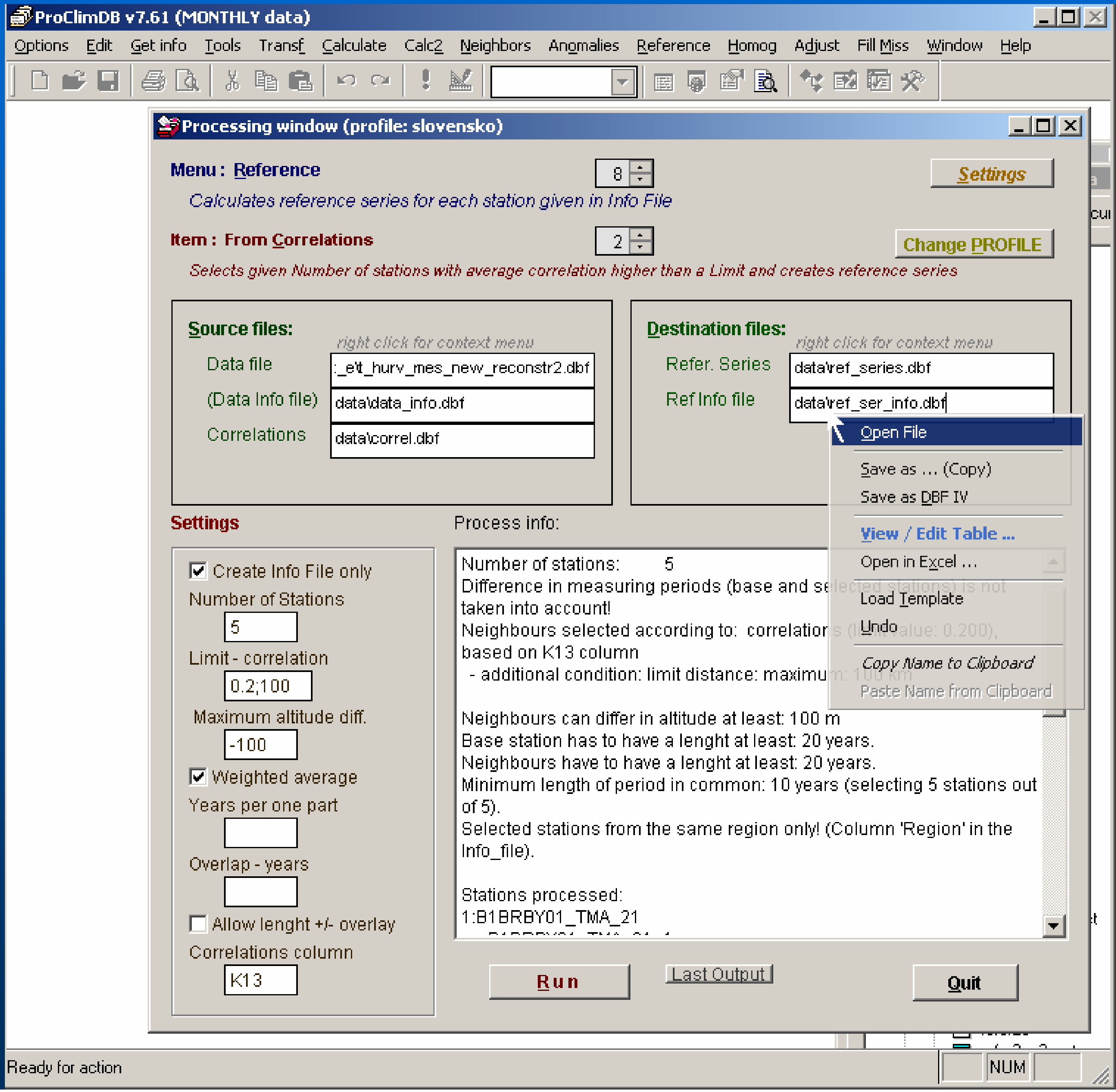
- **LoadData - application for downloading data from central database (e.g. Oracle)**
- **ProClimDB software for processing whole dataset (finding outliers, combining series, creating reference series, preparing data for homogeneity testing, extreme value analysis, RCM outputs validation, correction,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

http://www.climahom.eu

AnClim software



ProClimDB software



http://www.climahom.eu