Quality control and homogenization procedure of the extreme temperature series of Murcia Region

Author: Elisa Mª Hernández García
Project: Climatological study of heatwaves

Tutor: Juan Andrés García Valero
R & D (D.T. AEMET Región de Murcia)

INDEX

1. Motivation
2. Objectives
3. Applications
4. General procedure
5. Daily database
6. Quality control
7. Homogenization
8. Climatol
9. Results and Conclusions
10. Bibliography

1. MOTIVATION

To study the natural variability of climatological series of temperature is necesary obtaining long and homogeneous reference series. However, climatological records often contain inhomogeneities (station relocations, equipment changes, equipment drifts, changes in the method of data collection and changes in the general surroundings of a station). Therefore, we want to obtain homogeneous daily reference series of extreme temperature for Murcia Region, which they can be used for future projects.

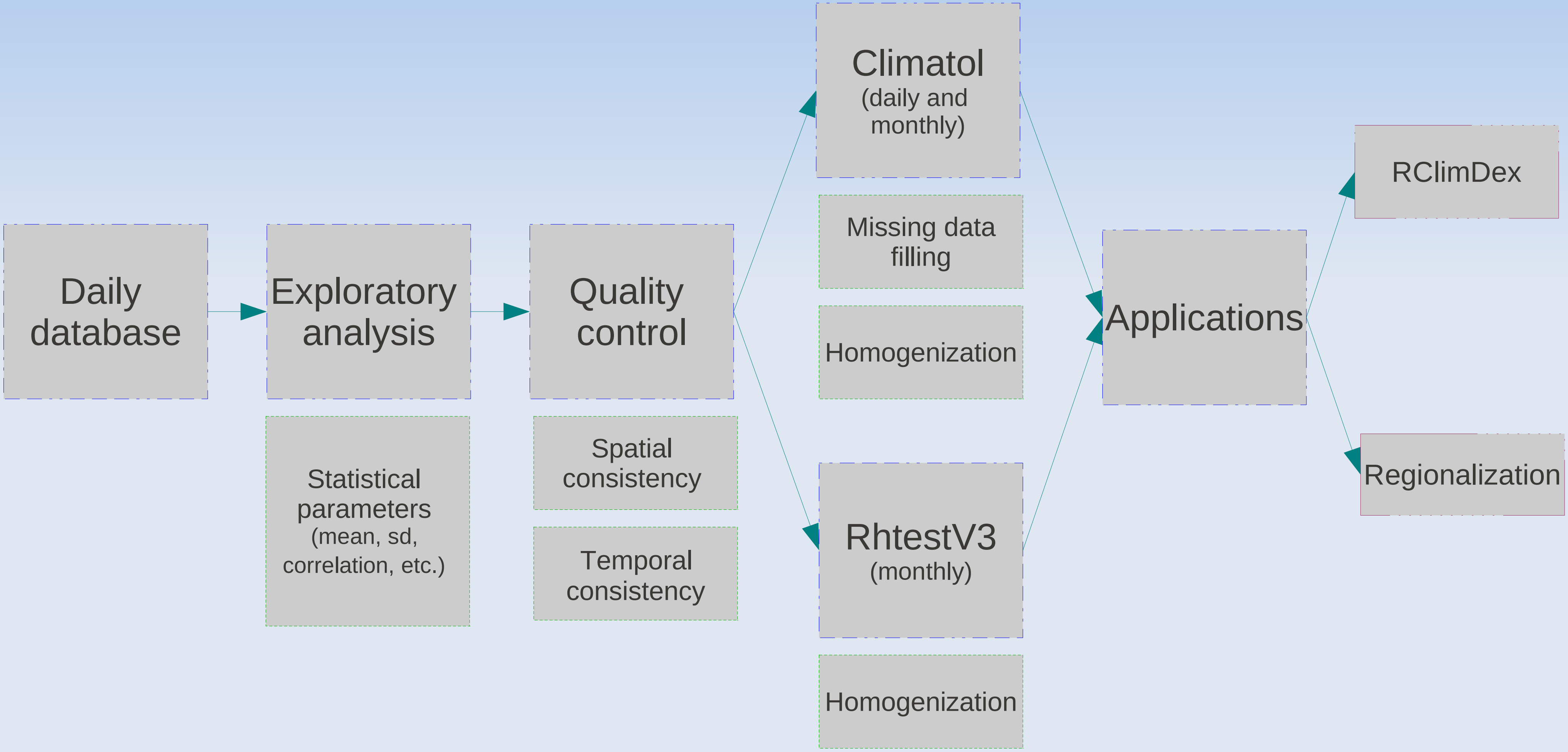
2. OBJECTIVES

- Development of some R programs for the quality control of daily series.
- Obtaining a complete daily database of homogeneous maximum temperature, in order to using it as a reference.

3. APPLICATIONS

- Spatial-temporal analysis for the daily regionalization of maximum temperatures in summer in Murcia Region
- Study of the temporal variability of homogeneous climatological series:
  - Trend analysis
  - Extreme indices (RClimDex)

4. GENERAL PROCEDURE



Daily database

Exploratory analysis

Statistical parameters (mean, sd, correlation, etc.)

Quality control

Spatial consistency

Temporal consistency

Climatol (daily and monthly)

Missing data filling

Homogenization

RhtestV3 (monthly)

Homogenization

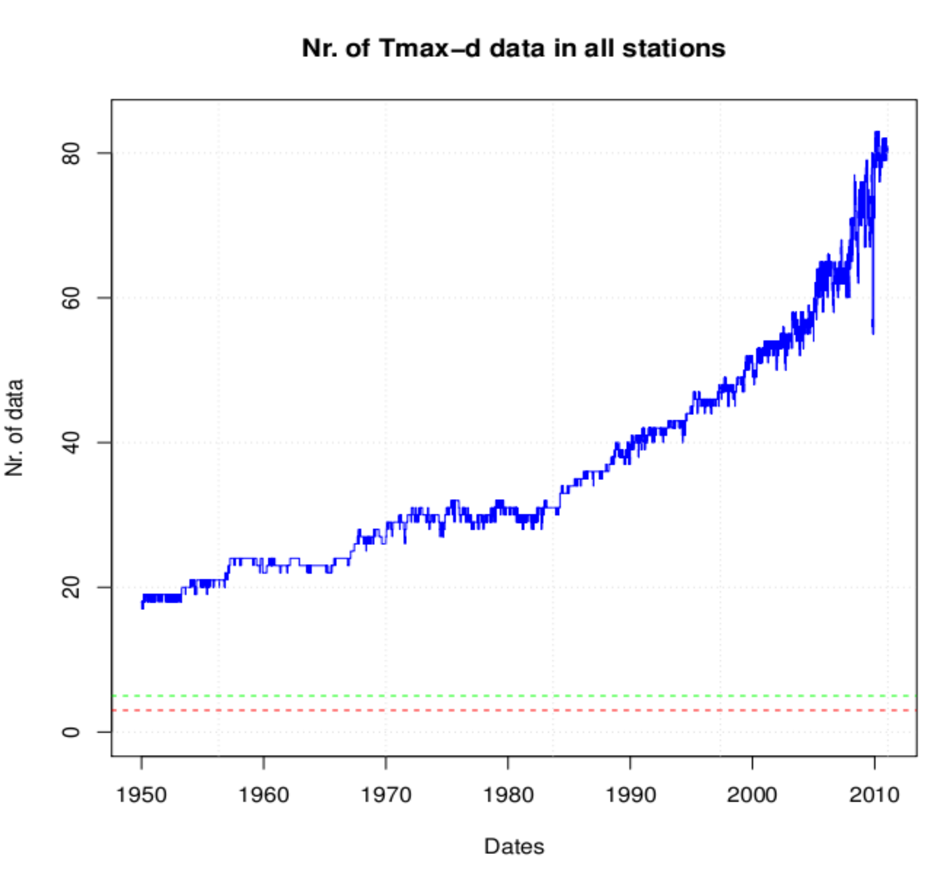
Applications

RClimDex

Regionalization

5. DAILY DATABASE

- Characteristics:
  - 84 stations
  - Period: 1950-2010
  - Variables: extreme temperatures



- prepara-matriz.R (batch):
  - It imports the data from the AEMET database and it changes the format (date, $station_1$, $station_2$, ..., $station_n$)

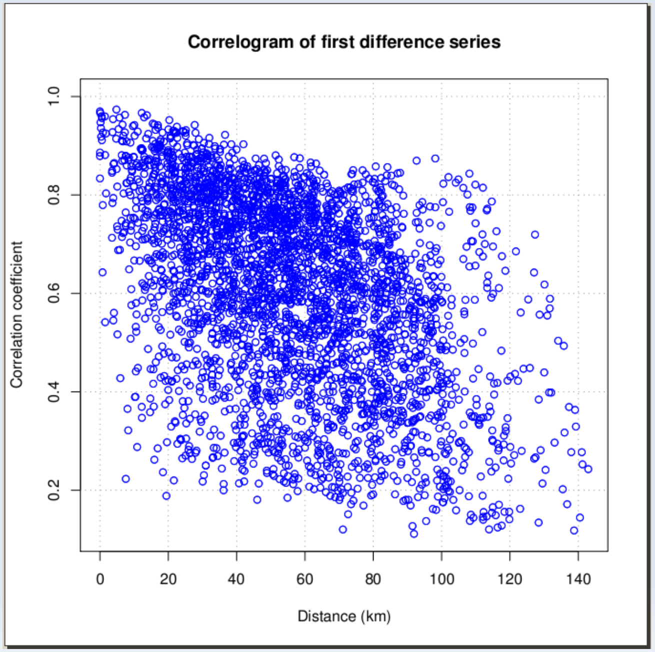
5.1. Exploratory analysis

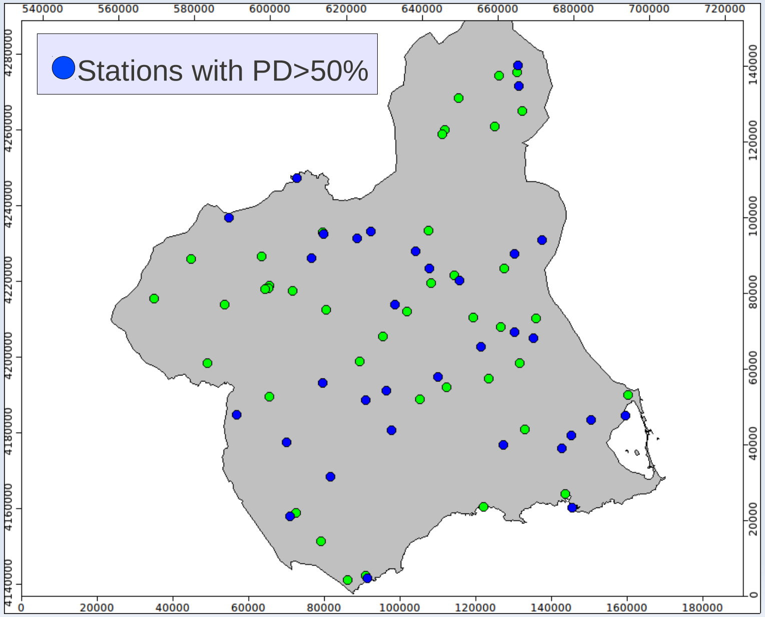
- Exploratory analysis of data series from the database for extreme temperatures.

BATCH	DESCRIPTION
analysis1.R	Calculation of some statistics (mean, sd, %NA, %NA5) from the climatological series (extreme temperatures) of the stations
analysis2.R	Analysis of spatial consistency, by means of the correlation between stations
analysis3.R	Analysis of spatial consistency, by means of the covariance between stations

5.2. Exploratory Analysis. Results

VARIABLE	MIN.	1st QU.	MEDIAN	MEAN	3rd QU.	MAX.
Mean (ºC)	16.18	21.80	23.01	22.56	23.83	24.96
Standard deviation (ºC)	5.39	7.06	7.66	7.47	8.19	8.79
Missing data (%)	0.01	17.64	68.98	56.68	90.21	99.60
Missing data last 5 years (%)	0.00	1.95	5.36	18.02	20.18	95.89

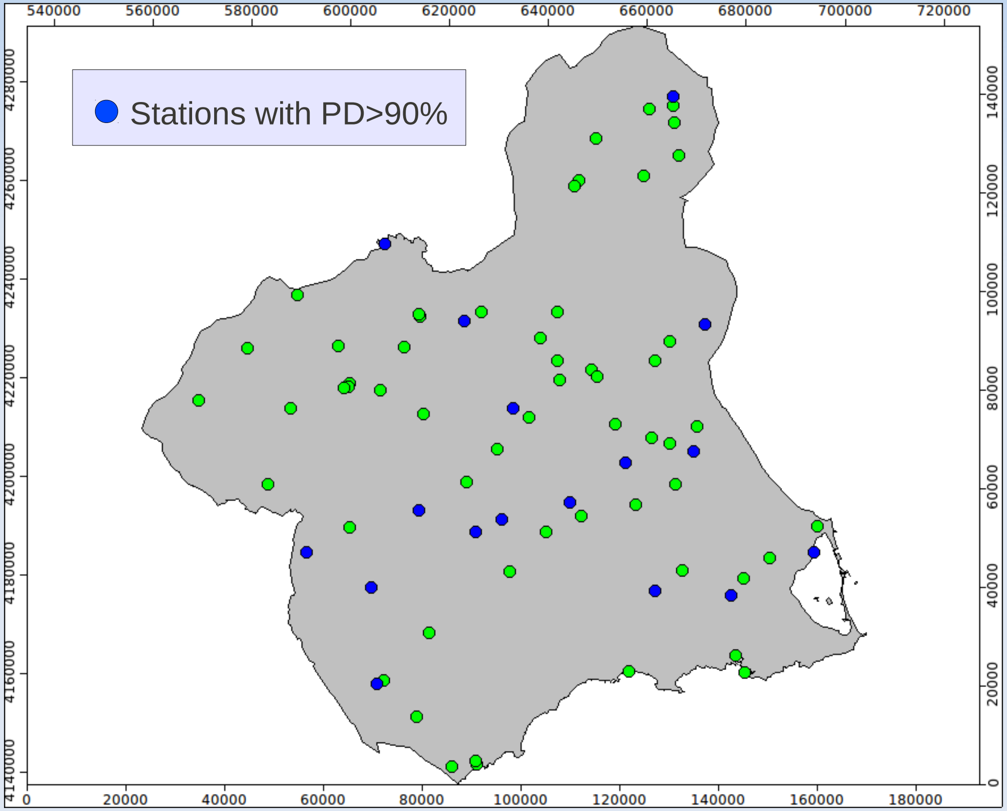




The stations in blue have more than 50% of data in the period 1950-2010.

5.2. Exploratory Analysis. Results

CODE	STATION	% NA
7228	Murcia/Alcantarilla	0.01
7026	Cartagena (Pozo Estrecho)	0.24
7083	Moratalla (Emb. Cenajo)	0.31
7198	Lorca (Emb. Valdeinfierno)	0.41
7031	Murcia/San Javier	0.55
7211	Puerto Lumbreras C.H.S.	0.79
7205	Lorca (Emb. Puentes)	0.82
7250	Abanilla C.H.S.	1.04
7168	Mula (Emb. Cierva)	1.43
7129	Calasparra (Emb. Alfonso XIII)	1.63
7219	Alhama (Huerta Espuña)	1.63
7206	Lorca (Zarzadilla de Totana)	1.68
7275	Yecla C.H.S.	2.06
7226	Librilla C.H.S.	2.70
7231	Murcia (Beniaján)	3.39
7214	Totana (Alquerías)	4.41
7023	Fuente Álamo C.H.	6.75

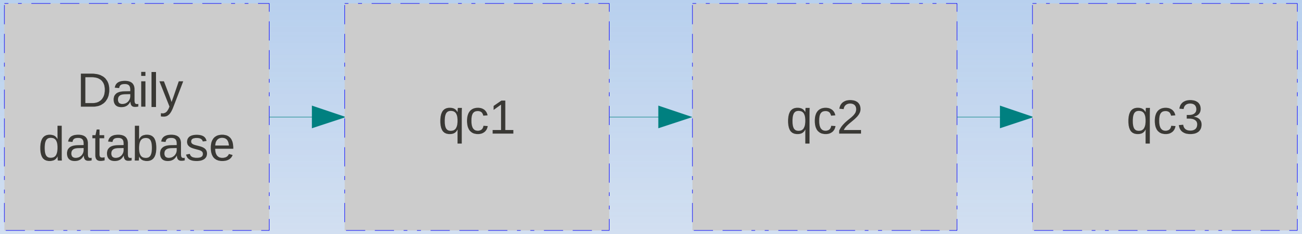


6. QUALITY CONTROL

- Methods *(Feng et al., 2004)*

METHODS	**AUTHOR**
High-low extreme check for daily values	*Kubecka, P., 2001* *Gleason, E., 2002*
Internal consistency check	*Reek et al., 1992*
Temporal outliers check	Lanzante, J.R., 1996 *Gleason, E., 2002*
Spatial outliers check	*Hubbard, K.G., 2001*
Missing data	*Stooksbury et al., 1999*

6.1. Quality controls performed



BATCH	DESCRIPTION
qc1.R	Analysis of persistence in extreme temperatures for each station. This function removes identical data, in maximum and minimum temperatures, for several consecutive days.
qc2.R	Analysis of spatial consistency between stations. Inconsistency data (>3sd) are removed in each climatological serie.
qc3.R	Analysis of temporal consistency: outliers in extreme temperatures, daily increase of temperature, internal consistency check, daily amplitude of temperature and interquartile range of temperatures.

6.2. Results QC

CODE	STATION	NUM. DAYS	DATE	% PERSISTENCE
72375	Molina de Segura (Los Valientes)	5 5 5	19970731 19980807 19980822	0.69
71192	Caravaca, Fuentes del Marqués	4	20081021	0.58
71183	Caravaca (Archivel)	3 3 3	20030826 20051215 20100714	0.53
72031	Zarcilla de Ramos	5	20090823	0.46
71372	Jumilla (Fuente del Pino)	5	20021120	0.15
72212	Alhama (Cena Guerrero)	10	19870920	0.10
7113	Moratalla (Campo de San Juan)	4	19620803	0.10
7195	Caravaca (Los Royos)	5	20100626	0.08
71274	Bullas (Depuradora)	4	20020929	0.08
71235	Moratalla (Inazares)	2	20090821	0.08
70015	Águilas (Montagro)	5	20030901	0.05

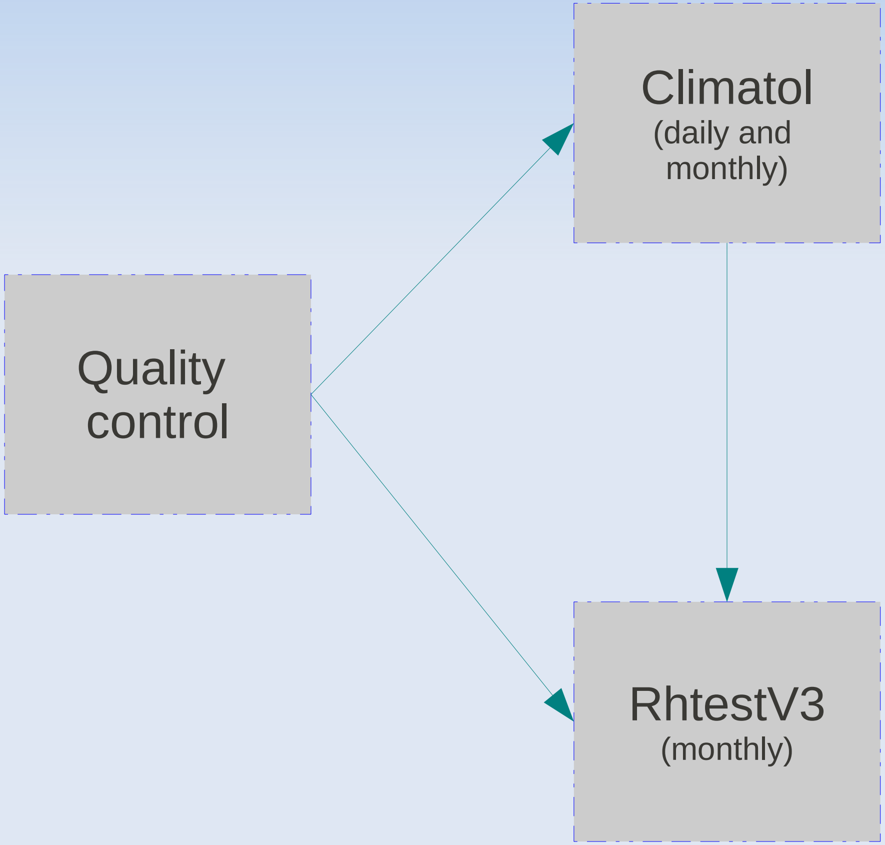
6.2. Results QC

CODE	STATION	% TEMPORAL OUTLIERS
7031	Murcia/San Javier	0.031
70015	Águilas (Montagro)	0.026
7231	Murcia (Beniaján)	0.017
7026	Cartagena (Pozo Estrecho)	0.013
70289	Torre Pacheco (Torre Blanca)	0.013
7228	Murcia/Alcantarilla	0.013
7002	Águilas Diputación	0.013
7032	San Pedro del Pinatar Ayto.	0.008
71611	Archena H.E.	0.008
7182	Murcia Alfonso X	0.008
7218	Totana I.L.	0.008
63719	Lorca (La Escarihuela)	0.004
7013	Cartagena Puerto	0.004
7168	Mula (Emb. Cierva)	0.004
7209	Lorca	0.004
7219	Alhama (Huerta Espuña)	0.004
72212	Alhama (Cena Guerrero)	0.004
7227	Alhama (Comarza)	0.004
7237	Fortuna C.H.S.	0.004

7. HOMOGENIZATION

Standard normal homogeneity test (SNHT) without trend *(Alexandersson, 1986)*	SNHT with trend *(Alexandersson and Moberg, 1997)*	Multiple linear regression (MLR) *(Vincent, 1998)*	Two-phase regression (TPR) *(Easterling and Peterson, 1995)*	Wilcoxon rank-sum (WRS) *(Karl and Williams, 1987)*
Sequential testing for equality of means (ST) *(Gullett et al., 1990)*	Bayesian approach without reference series *(Ouarda et al., 1999; Perreault et al., 1999, 2000)*	Bayesian approach with reference series *(Ouarda et al., 1999; Perreault et al., 1999, 2000)*	Higher-order moments (HOM) (A. Toreti, 2010)	Higher-order moments for autocorrelated data (HOMAD) (A. Toreti, 2010)
Pettit Test *(Pettit, A.N., 1979)*	Buishand range test *(Buishand, T.A., 1982)*	Von Neumann ratio test (VNRT) *(Von Neumann, 1941)*	Penalized maximal *t* test (PMT) *(Wang, X.L., 2008)*	Penalized maximal *F* test (PMF) *(Wang, X.L., 2008)*

7.1. Procedure



Quality control

Climatol
(daily and monthly)

Missing data filling

Homogenization (SNHT)

R contributed package for homogenization of climatological series, which it was developed by J. A. Guijarro (Guijarro, 2006)

RhtestV3
(monthly)

Homogenization (PMF)

Software package used to detect, and adjust for, multiple changepoints that could exist in data series
(Wang and Feng, 2010)

8. CLIMATOL (version 2.1)

- **Homogen function**: automatic homogenization of climatological series, including missing data filling and detection/correction of outliers and shifts in the mean of the series.
- Steps:
  1) Calculation of a reference serie for each station, by means of standardized data and distance criteria (iterative process).
  2) Analysis of outliers and homogenization in each serie.
  3) Missing data filling.

8.1. Homogen function: Arguments

varcli	Name of the studied climatic variable
anyi	Initial year of the data
anyf	Final year of the data
nm	Number of data per year in each station (nm=0 for daily data)
nref	Maximum number of references for data estimation
dz.max	Threshold of outlier tolerance, in standard deviations
wd	Distance (in km) at which reference data will weigh half that of another located at the same site of the series been estimated
tVt	Threshold value of the stepped SNHT window test
tVf	Tolerance factor to split several series at time
swa	Size of the step forward to be applied to the windowed application of SNHT
snhtt	Threshold value for the SNHT test when applied to the complete series
mxdif	Maximum difference of any data item in consecutive iterations
force	Force break even when only one reference is available
a, b	Parameters of the optional transformation a+b*dat to be applied to data when read from the files

wz	Scale parameter of the vertical coordinate Z
deg	Set to TRUE if the input coordinates are in geographical degrees instead of km
rtrans	Root transformation to apply to the data
std	Type of normalization
ndec	Number of decimal digits to which the homogenized data must be rounded
mndat	Minimum number of data for a plit fragment to become a new series
leer	Set to FALSE if you read your data with your own R routines
gp	Graphic parameter
na.strings	Character string to be treated as a missing value
nclust	Maximum number of stations for the cluster analysis
maxite	Maximum number of iterations when computing the means of the series
ini	Initial date (with format 'YYYY-MM-DD')
vmin	Minimum possible value of the studied variable
vmax	Maximum possible value of the studied variable
verb	Verbosity

8.2. Homogen function: Daily tests

```
homogen(”Max”, 1950, 2010, nm=0, b=0.1, deg=TRUE,
ini=”1950-01-01”).
```

- All stations

TEST	tVt	wd	swa	Mean SNHT	Mean RMSE	Mean PD
0	0	(50,50,25)	Default value (60)	899.7	1.79	42.74
1	0	(0,0,100)	Default value	889.6	1.81	42.74
2	0	(0,0,25)	Default value	899.7	1.78	42.74
3	0	(30,30,15)	Default value	898.8	1.77	42.74
5	260	(25,25,25)	1080	117.3	1.75	19.26
6	450	(30,30,15)	1080	258.6	1.63	26.10

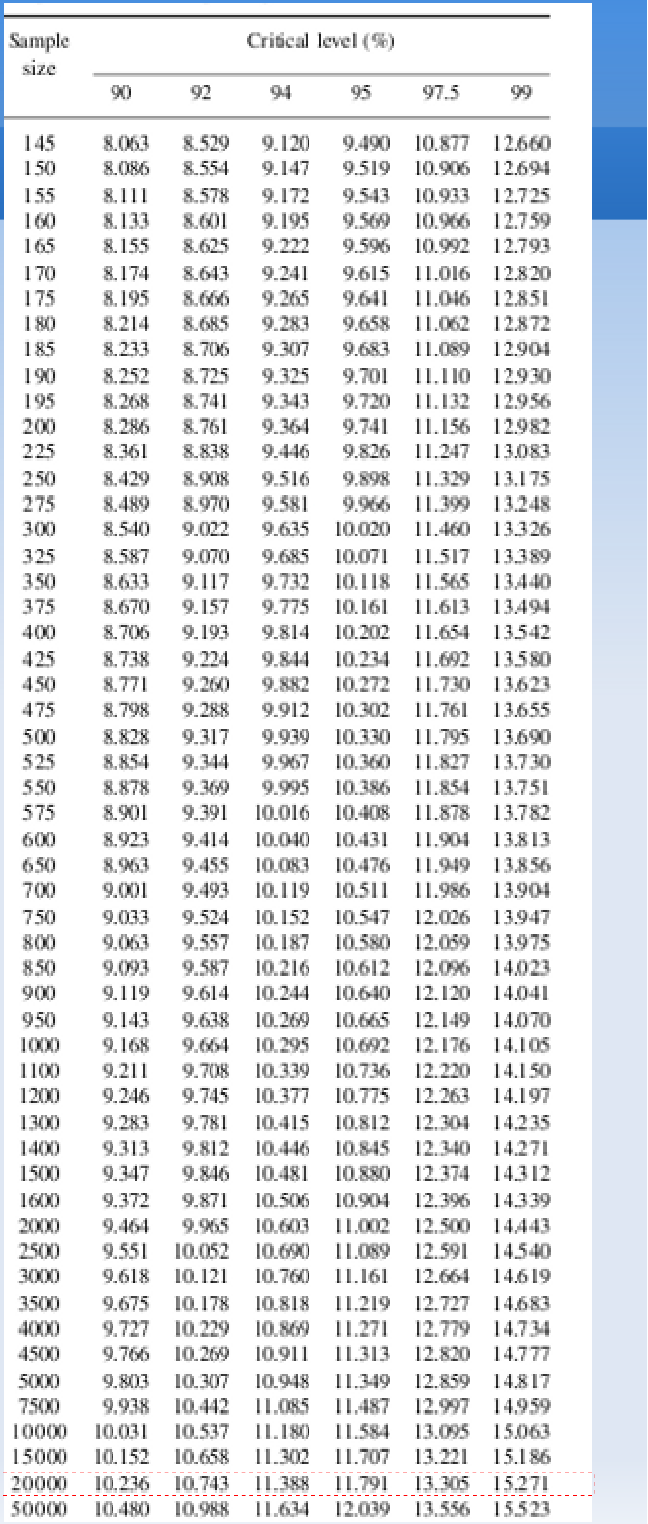
8.2. Homogen function: Daily tests

homogen(”Max”, 1950, 2010, nm=0, b=0.1, deg=TRUE, ini=”1950-01-01”)

- Stations PD 90

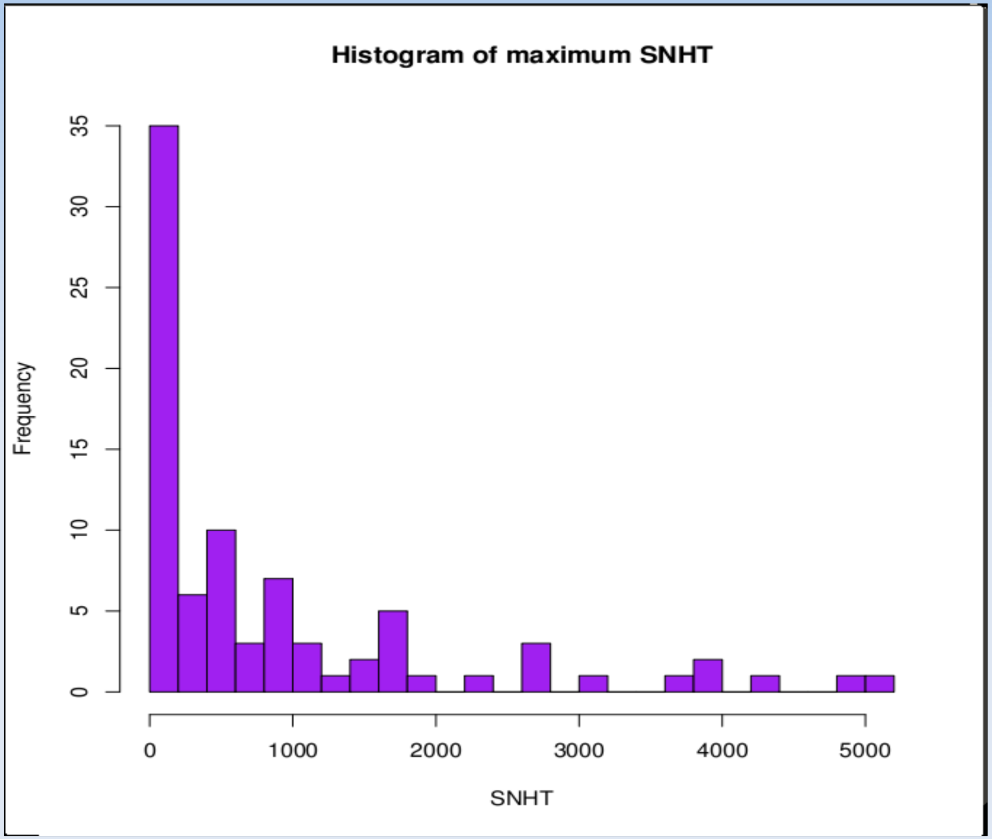
TEST	tVt	wd	swa	Mean SNHT (PD90)	Mean RMSE (PD90)	Mean PD (PD90)
0	0	(50,50,25)	Default value (60)	2271	2.28	96.74
1	0	(0,0,100)	Default value	2237	2.28	96.74
2	0	(0,0,25)	Default value	2271	2.28	96.74
3	0	(30,30,15)	Default value	2277	2.27	96.74
5	260	(25,25,25)	1080	128.2	2.47	27.89
6	450	(30,30,15)	1080	447.1	2.28	40.89

8.2.1. Daily tests: Results



Sample size	Critical level (%)					
	90	92	94	95	97.5	99
145	8.063	8.529	9.120	9.490	10.877	12.660
150	8.086	8.554	9.147	9.519	10.906	12.694
155	8.111	8.578	9.172	9.543	10.933	12.725
160	8.133	8.601	9.195	9.569	10.966	12.759
165	8.155	8.625	9.222	9.596	10.992	12.793
170	8.174	8.643	9.241	9.615	11.016	12.820
175	8.195	8.666	9.265	9.641	11.046	12.851
180	8.214	8.685	9.283	9.658	11.062	12.872
185	8.233	8.706	9.307	9.683	11.089	12.904
190	8.252	8.725	9.325	9.701	11.110	12.930
195	8.268	8.741	9.343	9.720	11.132	12.956
200	8.286	8.761	9.364	9.741	11.156	12.982
225	8.361	8.838	9.446	9.826	11.247	13.083
250	8.429	8.908	9.516	9.898	11.329	13.175
275	8.489	8.970	9.581	9.966	11.399	13.248
300	8.540	9.022	9.635	10.020	11.460	13.326
325	8.587	9.070	9.685	10.071	11.517	13.389
350	8.633	9.117	9.732	10.118	11.565	13.440
375	8.670	9.157	9.775	10.161	11.613	13.494
400	8.706	9.193	9.814	10.202	11.654	13.542
425	8.738	9.224	9.844	10.234	11.692	13.580
450	8.771	9.260	9.882	10.272	11.730	13.623
475	8.798	9.288	9.912	10.302	11.761	13.655
500	8.828	9.317	9.939	10.330	11.795	13.690
525	8.854	9.344	9.967	10.360	11.827	13.730
550	8.878	9.369	9.995	10.386	11.854	13.751
575	8.901	9.391	10.016	10.408	11.878	13.782
600	8.923	9.414	10.040	10.431	11.904	13.813
650	8.963	9.455	10.083	10.476	11.949	13.856
700	9.001	9.493	10.119	10.511	11.986	13.904
750	9.033	9.524	10.152	10.547	12.026	13.947
800	9.063	9.557	10.187	10.580	12.059	13.975
850	9.093	9.587	10.216	10.612	12.096	14.023
900	9.119	9.614	10.244	10.640	12.120	14.041
950	9.143	9.638	10.269	10.665	12.149	14.070
1000	9.168	9.664	10.295	10.692	12.176	14.105
1100	9.211	9.708	10.339	10.736	12.220	14.150
1200	9.246	9.745	10.377	10.775	12.263	14.197
1300	9.283	9.781	10.415	10.812	12.304	14.235
1400	9.313	9.812	10.446	10.845	12.340	14.271
1500	9.347	9.846	10.481	10.880	12.374	14.312
1600	9.372	9.871	10.506	10.904	12.396	14.339
2000	9.464	9.965	10.603	11.002	12.500	14.443
2500	9.551	10.052	10.690	11.089	12.591	14.540
3000	9.618	10.121	10.760	11.161	12.664	14.619
3500	9.675	10.178	10.818	11.219	12.727	14.683
4000	9.727	10.229	10.869	11.271	12.779	14.734
4500	9.766	10.269	10.911	11.313	12.820	14.777
5000	9.803	10.307	10.948	11.349	12.859	14.817
7500	9.938	10.442	11.085	11.487	12.997	14.959
10000	10.031	10.537	11.180	11.584	13.095	15.063
15000	10.152	10.658	11.302	11.707	13.221	15.186
20000	10.236	10.743	11.388	11.791	13.305	15.271
50000	10.480	10.988	11.634	12.039	13.556	15.523

(Khaliq and Ouarda, 2007)



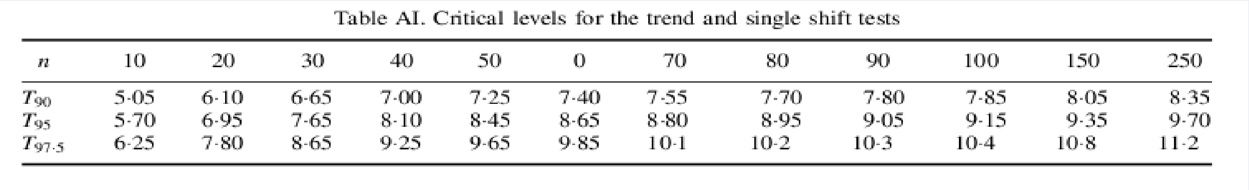


Table AI. Critical levels for the trend and single shift tests

n	10	20	30	40	50	0	70	80	90	100	150	250
T_{90}	5·05	6·10	6·65	7·00	7·25	7·40	7·55	7·70	7·80	7·85	8·05	8·35
T_{95}	5·70	6·95	7·65	8·10	8·45	8·65	8·80	8·95	9·05	9·15	9·35	9·70
$T_{97\cdot5}$	6·25	7·80	8·65	9·25	9·65	9·85	10·1	10·2	10·3	10·4	10·8	11·2

(Alexandersson and Moberg, 1997)

8.3. Homogen function: Monthly tests

homogen(”Max”, 1950, 2010, nm=12, deg=TRUE)

- All stations

TEST	tVt	wd	swa	Mean SNHT	Mean RMSE	Mean PD
Monthly 0	0	(50,50,25)	Default value (60)	85.39	0.91	43.6
Monthly 1	25	(30,30,15)	Default value	13.82	0.66	14.86
Monthly 2	33	(50,50,25)	Default value	19.44	0.71	18.07

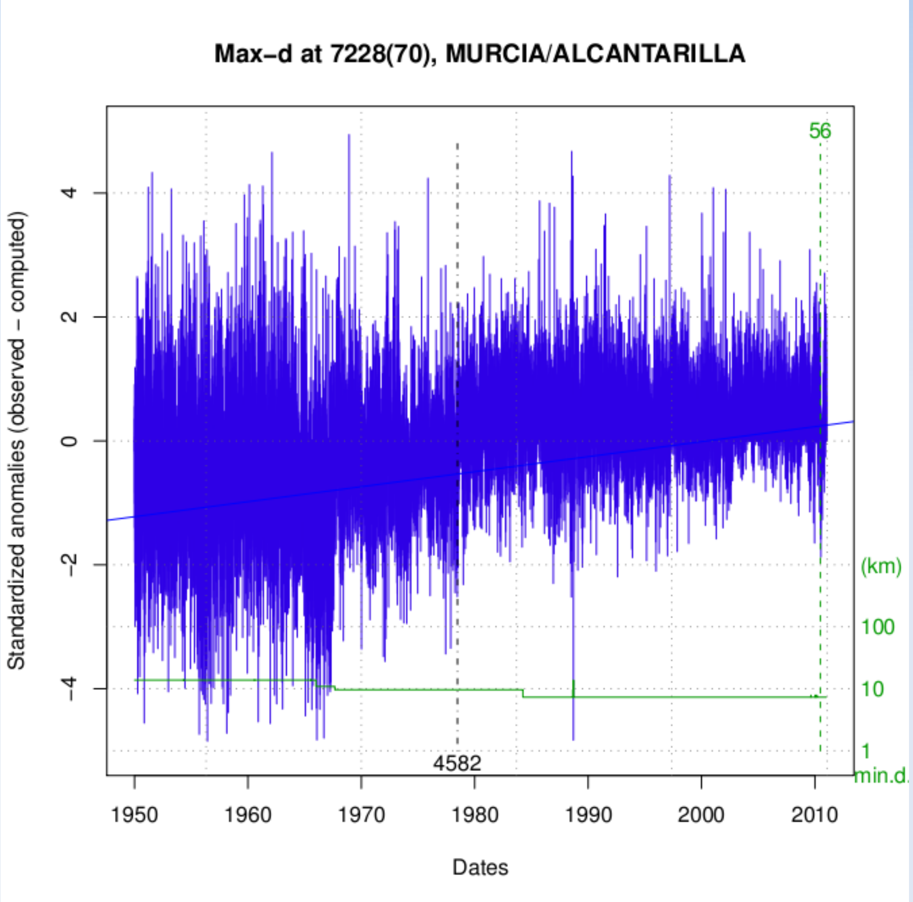
8.3. Homogen function: Monthly tests

homogen(”Max”, 1950, 2010, nm=12, deg=TRUE)

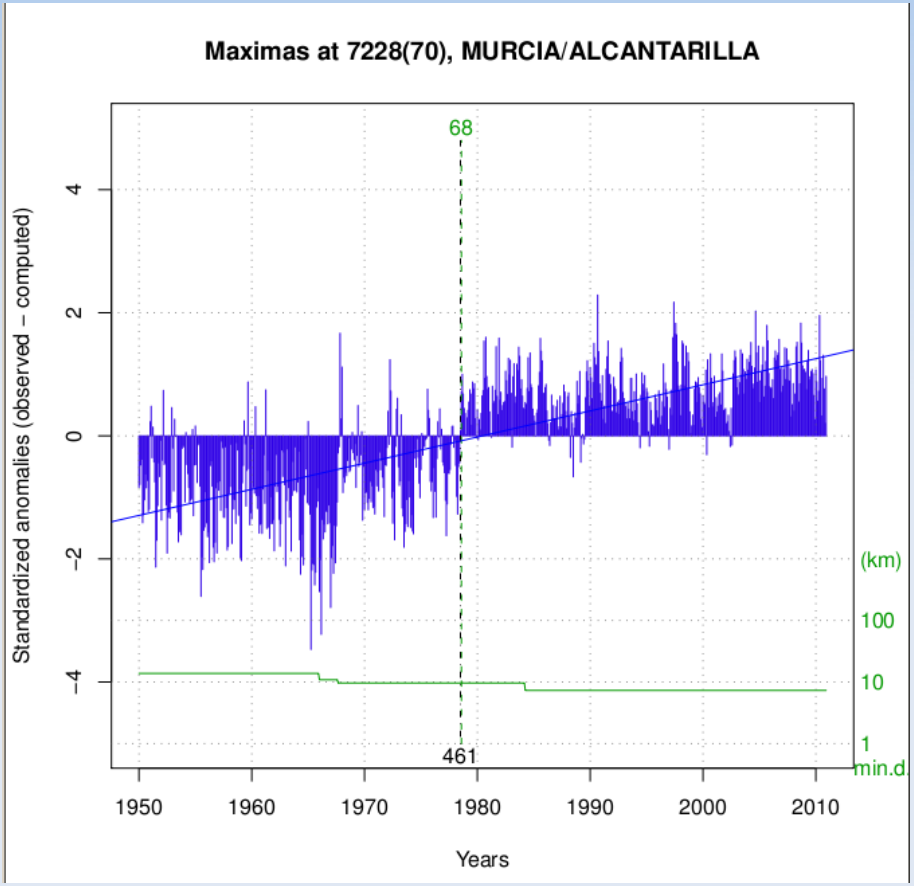
- Stations PD90

TEST	tVt	wd	swa	Mean SNHT (PD90)	Mean RMSE (PD90)	Mean PD (PD90)
Monthly 0	0	(50,50,25)	Default value (60)	201.6	1.33	97.11
Monthly 1	25	(30,30,15)	Default value	18.26	1.14	19.74
Monthly 2	33	(50,50,25)	Default value	26.16	1.29	21.32

8.3. Homogen function: Outputs



Daily



Monthly

8.4. Climatol Vs. RHtestV3

Breaks in Murcia/Alcantarilla	
CLIMATOL	***RHtestV3***
August 1967	August 1967
July 1978	July 1978
March 1986	
May 1997	
September 2002	

9. RESULTS AND CONCLUSIONS

- The choice of shorter distances to create reference series improves the values of RMSE, but the values of SNHT get worse.
- The choice of low tVt in daily series cut them into several series, decreasing PD values and RMSE improves slightly. Therefore we have to continue the research of those values.
- Monthly series have similar results than daily series.
- The combination of results, which are obtained by means of different homogenization methods (Climatol, RhtestV3), can help to detect the most important inhomogeneities and to choose the tVt thresolds.
- In the daily scale, the choice of higher values of tVt is advisable, due to it will let to leak the higher inhomogeneities.
- The highest values of SNHT belong to the longest climatological series and they exceed the critical values, which are given by the literature.

↓

Is there any sense to homogenize daily scale?

10. BIBLIOGRAPHY

- Aguilar et al., 2003. ”Guidelines on climate metadata and homogenization”. WCDMP 53 – WMO/TD 1186.
- Barrera, A. 2004. ”Técnicas de completado de series mensuales y aplicación al estudio de la influencia de la NAO en la distribución de la precipitación en España”. Univ. Barcelona
- Ducre-Robitaille et al., 2003. ”Comparison of techniques for detection of discontinuities in temperature series”. Int. J. Climatol. 23: 1087-1101.
- Feng et al., 2004. ”Quality control of daily meteorological data in China, 1951-2000: a new dataset”. Int. J. Climatol. 24:853-870.
- Guijarro, J.A., 2011. ”User's guide to climatol”
- Martinez et al., 2010. ”Time trends of daily maximum and minimum temperatures in Catalonia for the period 1975-2004”. Int. J. Climatol. 30: 267–290.
- Toreti et al., 2010. ”A novel method for the homogenization of daily temperature series and its relevance for climate change analysis”. J. Climate. Vol. 23, 19: 5325-5331.
- Wang, X.L., 2008. ”Accounting for autocorrelation in detection mean shifts in climate data series using the penalized maximal *t* or *F* test”. Appl. Meteorol. Climatol., 47: 2423-2444.
- Wang, X.L. and Feng, Y., 2010. ”RhtestV3 User Manual”