Quality Control and Homogenisation of the Belgian Historical Weather Data

C. Delvaux, R. Ingels, V. Vrabel, M. Journée, C. Bertrand

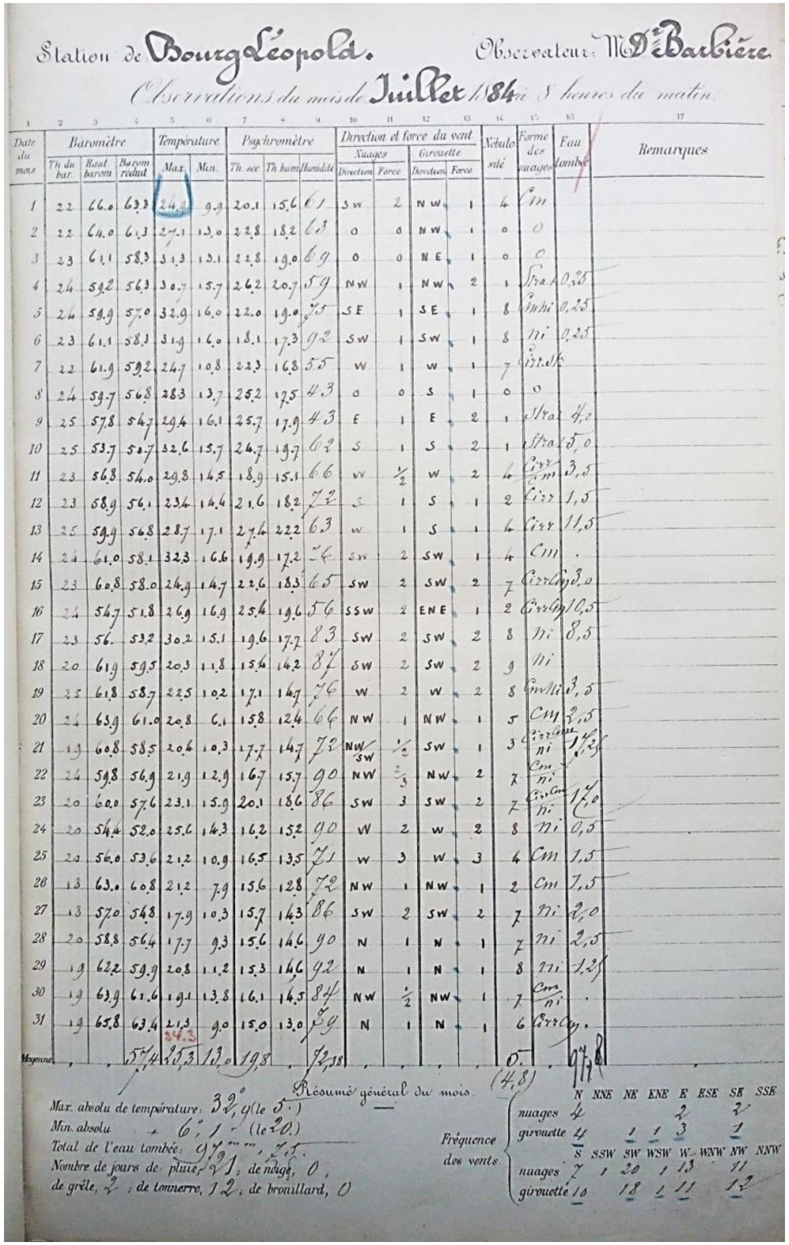
Royal Meteorological Institute of Belgium

Climatological and Meteorological Information Service

Introduction

- **Recent digitization project of belgian data**
  - Based on monthly climate bulletin
  - daily data (temperature and precipitation)
  - from 1880 to 1950

- Extend the Belgian daily data already available in our database from 1951 to nowadays



Station de Bourg Léopold. Observateur: M

Station de Bourg Léopold. Observateur: MD Barbière

Observations du mois de Juillet 1884 à 8 heures du matin.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Date du mois	Baromètre			Température		Psychromètre			Direction et force du vent				Nébulosité	Forme des nuages	Eau tombée	Remarques
									Nuages		Girouette					
	Th du bar.	Haut barom	Barom réduit	Max.	Min.	Th. sec	Th hum	Humidité	Direction	Force	Direction	Force				
1	22	66.0	63.3	24.0	9.9	20.1	15.6	61	SW	2	NW	1	4	Cm		
2	22	64.0	61.3	27.1	13.0	22.8	18.2	63	0	0	NW	1	0	0		
3	23	61.1	58.3	31.3	13.1	22.8	19.0	69	0	0	NE	1	0	0		
4	24	59.2	56.3	30.7	15.7	26.2	20.7	59	NW	1	NW	2	1	Stra	0,25	
5	24	59.9	57.0	32.9	16.0	22.0	19.0	75	SE	1	SE	1	8	Cmni	0,25	
6	23	61.1	58.3	31.9	16.0	18.1	17.3	92	SW	1	SW	1	8	ni	0,25	
7	22	61.9	59.2	24.7	10.8	22.3	16.8	55	W	1	W	1	7	Cirr.St		
8	24	59.7	56.8	28.3	13.7	25.2	17.5	43	0	0	S	1	0	0		
9	25	57.8	54.7	29.4	16.1	25.7	17.9	43	E	1	E	2	1	Stra	4,0	
10	25	53.7	50.7	32.6	15.7	24.7	19.7	62	S	1	S	2	1	Stra	5,0	
11	23	56.8	54.0	29.8	14.5	18.9	15.1	66	W	½	W	2	4	Cirr Cm	3,5	
12	23	58.9	56.1	23.4	14.4	21.6	18.2	72	S	1	S	1	2	Cirr	1,5	
13	25	59.9	56.8	28.7	17.1	27.4	22.2	63	W	1	S	1	4	Cirr	11,5	
14	24	61.0	58.1	32.3	16.6	19.9	17.2	76	SW	2	SW	1	4	Cm	.	
15	23	60.8	58.0	24.9	14.7	22.6	18.3	65	SW	2	SW	2	7	Cirr Cu	3,0	
16	24	54.7	51.8	26.9	16.9	25.4	19.6	56	SSW	2	ENE	1	2	Cirr Cu	10,5	
17	23	56.	53.2	30.2	15.1	19.6	17.7	83	SW	2	SW	2	8	ni	8,5	
18	20	61.9	59.5	20.3	11.8	15.4	14.2	87	SW	2	SW	2	9	ni		
19	25	61.8	58.7	22.5	10.2	17.1	14.7	76	W	2	W	2	8	Cm Ni	3,5	
20	24	63.9	61.0	20.8	6.1	15.8	12.4	64	NW	1	NW	1	5	Cm	2,5	
21	19	60.8	58.5	20.4	10.3	17.7	14.7	72	NW SW	½	SW	1	3	Cirr Cm ni	1,25	
22	24	59.8	56.9	21.9	12.9	16.7	15.7	90	NW	⅔	NW	2	7	Cm ni		
23	20	60.0	57.6	23.1	15.9	20.1	18.6	86	SW	3	SW	2	7	Cirr Cm ni	17,0	
24	20	54.4	52.0	25.6	14.3	16.2	15.2	90	W	2	W	2	8	ni	0,5	
25	20	56.0	53.6	21.2	10.9	16.5	13.5	71	W	3	W	3	4	Cm	1,5	
26	18	63.0	60.8	21.2	7.9	15.6	12.8	72	NW	1	NW	1	2	Cm	7,5	
27	18	57.0	54.8	17.9	10.3	15.7	14.3	86	SW	2	SW	2	7	ni	2,0	
28	20	58.8	56.4	17.7	9.3	15.6	14.6	90	N	1	N	1	7	ni	2,5	
29	19	62.2	59.9	20.8	11.2	15.3	14.6	92	N	1	N	1	8	ni	1,25	
30	19	63.9	61.6	19.1	13.8	16.1	14.5	84	NW	½	NW	1	7	Cm ni	.	
31	19	65.8	63.4	21.3 34.3	9.0	15.0	13.0	79	N	1	N	1	6	Cirr Cu	.	
Moyenne	.	.	57,4	25,3	13,0	19,8	.	72,38	.	.	.	.	0. (4,8)	.	97,8	

Résumé général du mois.

Max. absolu de température: 32°,9 (le 5.)
Min. absolu + 6°,1 (le 20.)
Total de l'eau tombée: 97mm,75.
Nombre de jours de pluie, 21; de neige, 0;
de grêle, 2; de tonnerre, 12, de brouillard, 0

Fréquence des vents

	N	NNE	NE	ENE	E	ESE	SE	SSE
nuages	4				2		2	
girouette	4		1	1	3		1	

	S	SSW	SW	WSW	W	WNW	NW	NNW
nuages	7	1	20	1	13		11	
girouette	10		18	1	11		12	

Introduction

- **Project**

Create high quality climatological long series in Belgium (*period 1880-2015*)

- **Parameters**

Daily maximum temperature (TX)
Daily minimum temperature (TN)
Precipitation (RR)

- **Main steps to obtain good results**

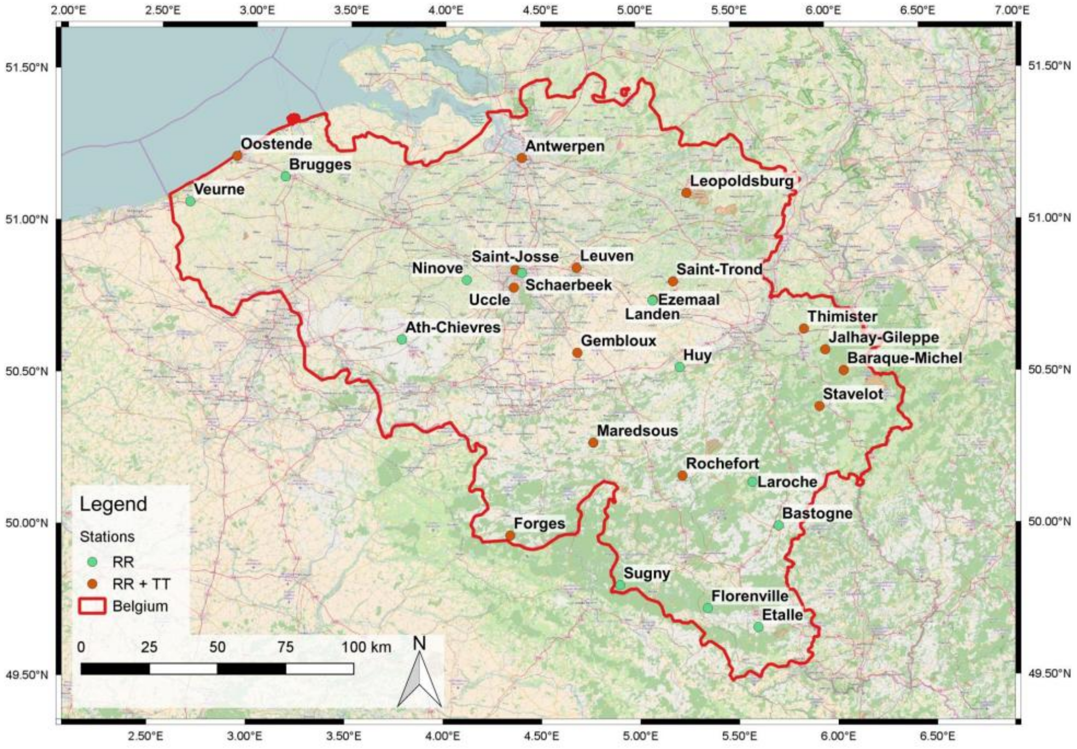
→ creation of long series
→ quality control of the data
→ monthly homogenization of the data (HOMER - Ongoing work)

Creation of long series

Almost none of the stations covers the entire period of time ->

The long series can be a combination of stations
maximum distance (10 km)
maximum elevation difference (50 m)

- **Long series** (1880 – 2015)
27 RR & 16 TT
- **Short series** (1951 – 2015)
162 RR & 66 TT



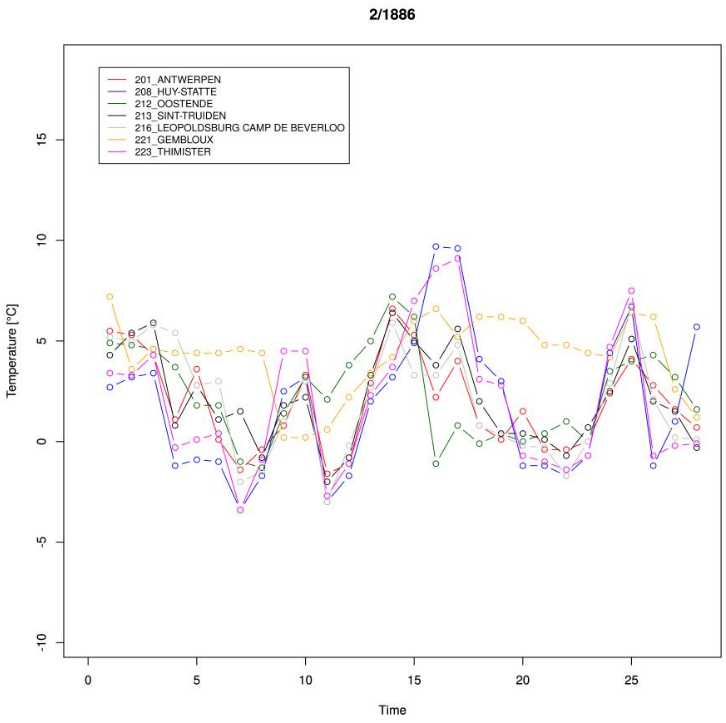
Quality control – Part 1

- Specific quality control needed for the new encoded data (1880 – 1950) of the long series (~ 1 million data) because of the bad quality of the data

Examples of the most frequent errors found

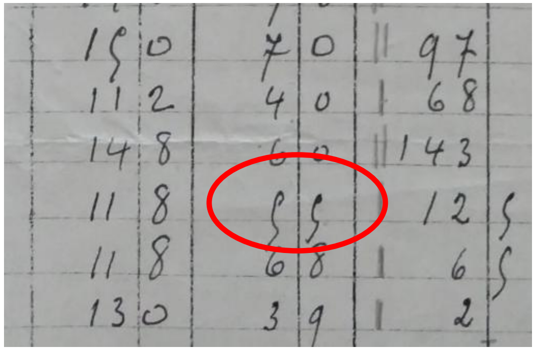
1) during encoding : wrong parameter encoded, duplicated data, confusion between missing/zero values, data attributed to a wrong station, classical typing error

➔ *Automatic and visual tests*



easy to detect and correct :
~~121 °C~~ → 12.1 °C

difficult to find, even with accurate test
9.9 °C -> 5.5 °C



15 0	7 0	97
11 2	4 0	68
14 8	6 0	143
11 8	5 5	125
11 8	6 8	65
13 0	3 9	2

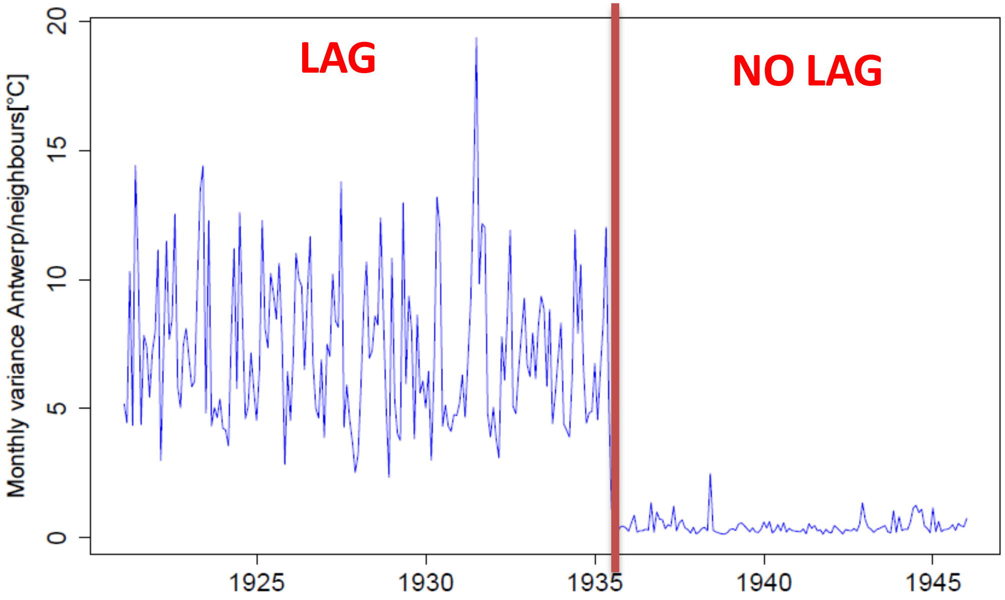
Quality control – Part 1

- Specific quality control needed for the new encoded data (1880 – 1949) of the long series (~ 1 million data) because of the bad quality of the data

Examples of the most frequent errors found

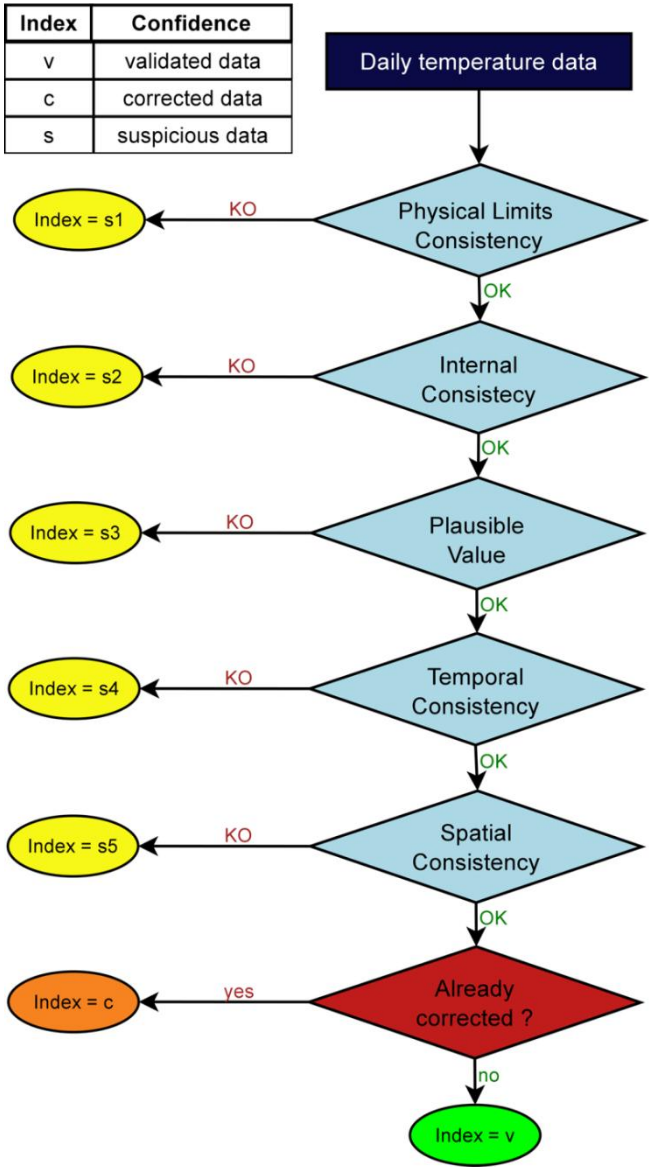
2) Observer error : precipitation not measured every day → accumulation

3) Transmission of data :
Bad communication between the institute and the observer



➔ About 20 % of data have been modified between 1880 – 1950 !

Quality control – Part 2



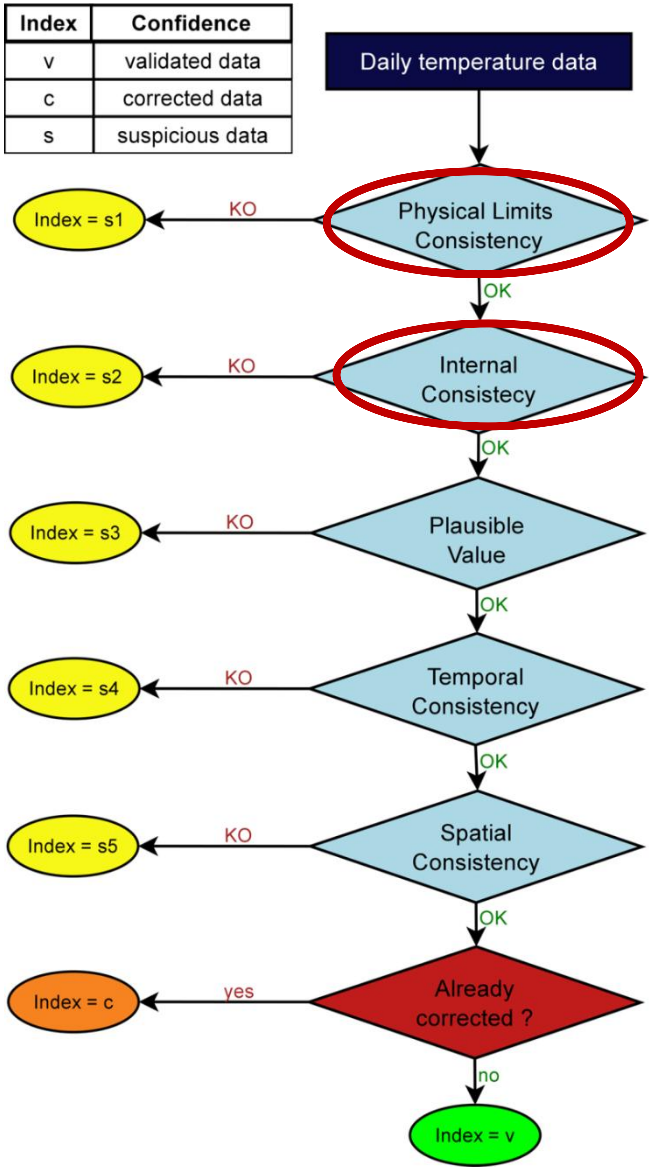
Index	Confidence
v	validated data
c	corrected data
s	suspicious data

- New quality control procedures

Minimum data quality required

- Applied to all the daily data (1880 – 2015)
- Basic tests to more specific tests
- Apply a quality index for each daily data
  - Validated data (*v)*
  - Suspicious data (*sX*) where *X* explain why the data is suspicious
  - Corrected data (*c*)
- Examples of some tests with TT (made for TN and TX)

Quality control – Part 2

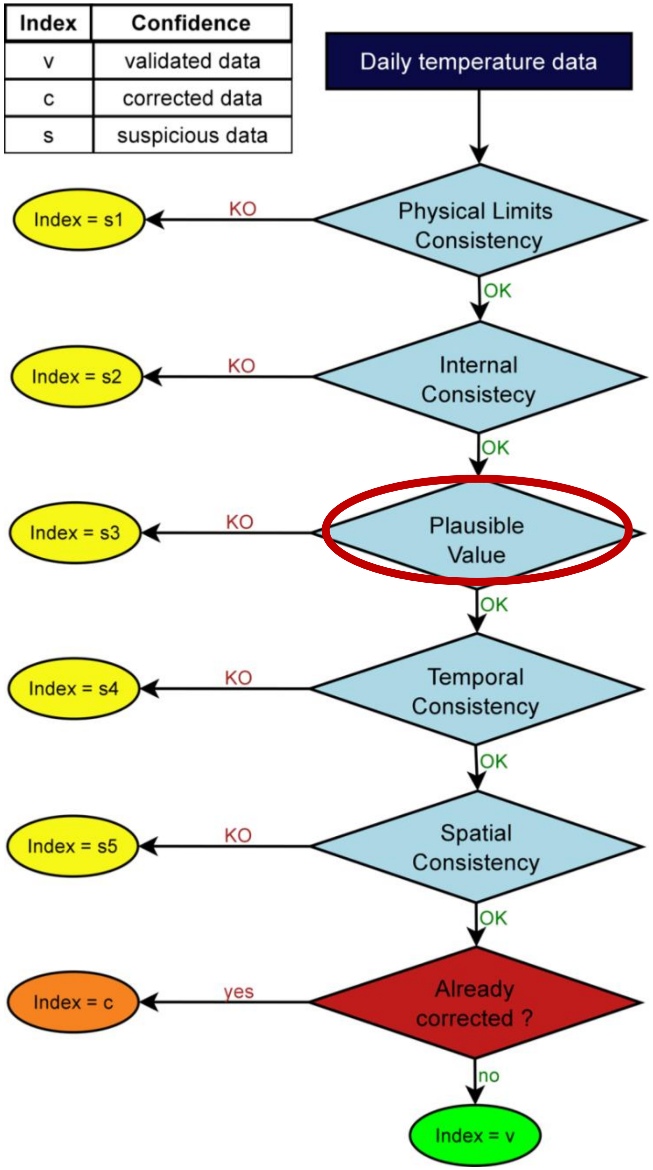


Index	Confidence
v	validated data
c	corrected data
s	suspicious data

- 50 °C < TT < 50 °C

TX > TN

Quality control – Part 2



Index	Confidence
v	validated data
c	corrected data
s	suspicious data

Daily temperature data

Physical Limits Consistency — KO → Index = s1; OK ↓

Internal Consistency — KO → Index = s2; OK ↓

Plausible Value — KO → Index = s3; OK ↓

Temporal Consistency — KO → Index = s4; OK ↓

Spatial Consistency — KO → Index = s5; OK ↓

Already corrected ? — yes → Index = c; no ↓

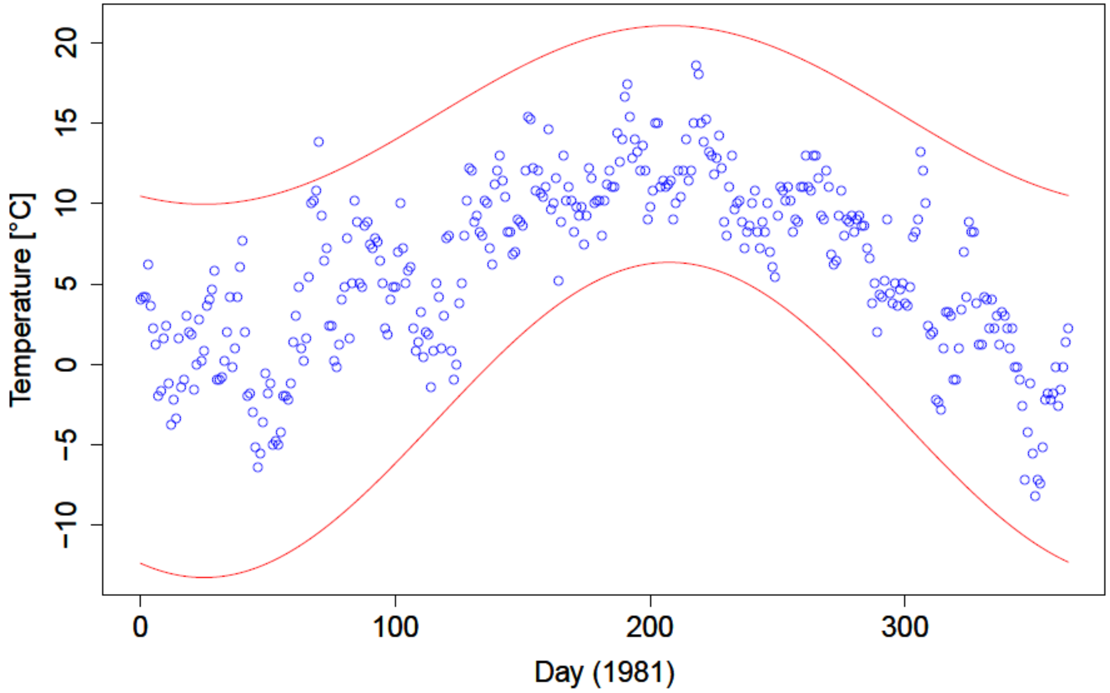
Index = v

T_{min} < TN < T_{max}

Envelope which assumes that the annual temperature variations follow a sinusoidal wave

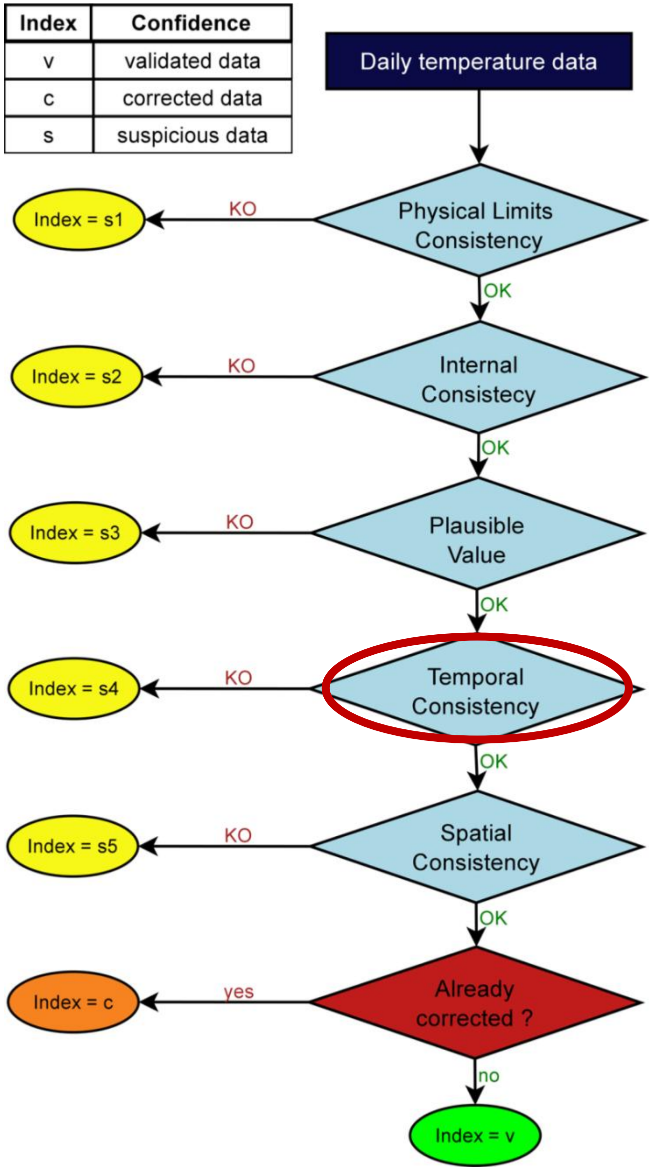
-> Upper and lower bounds by regions

-> Based on validated extreme temperature data observed each day



Example for TN – Lemberge (1981)

Quality control – Part 2



Index	Confidence
v	validated data
c	corrected data
s	suspicious data

Daily temperature data

Physical Limits Consistency — KO → Index = s1; OK

Internal Consistecy — KO → Index = s2; OK

Plausible Value — KO → Index = s3; OK

Temporal Consistency — KO → Index = s4; OK

Spatial Consistency — KO → Index = s5; OK

Already corrected ? — yes → Index = c; no

Index = v

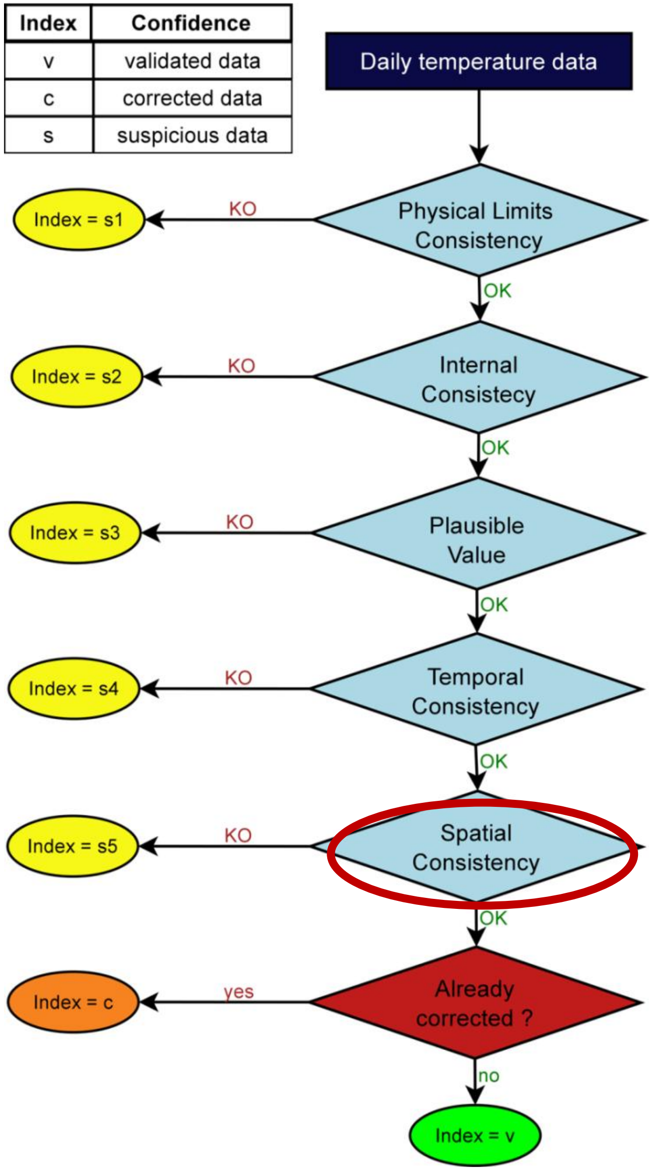
$|TN(day)\text{-}TN(day\text{-}1)| < \varepsilon$

- *ε based on the extreme temperature difference observed between two consecutive validated data*

- by month & by regions

Month	ε (TN)
Jan	14.3
Feb	12.5
Mar	12
Apr	10.1
May	10.4
Jun	10.3
Jul	9.6
Aug	9.4
Sep	10.6
Oct	11.7
Nov	13.2
Dec	12.6

Quality control – Part 2



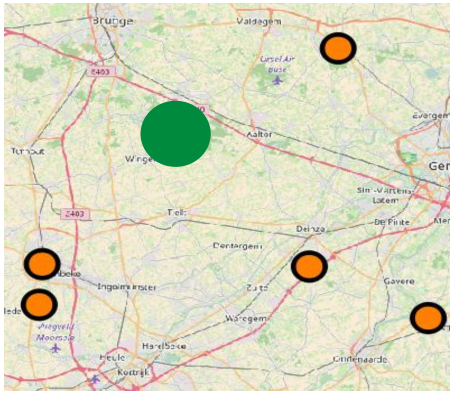
Index	Confidence
v	validated data
c	corrected data
s	suspicious data

1) Classic spatial test

5 closest neighboring values

*Based on “distance + 100 * altitude”*

- Inverse Distance Weighting
- Standard Deviation



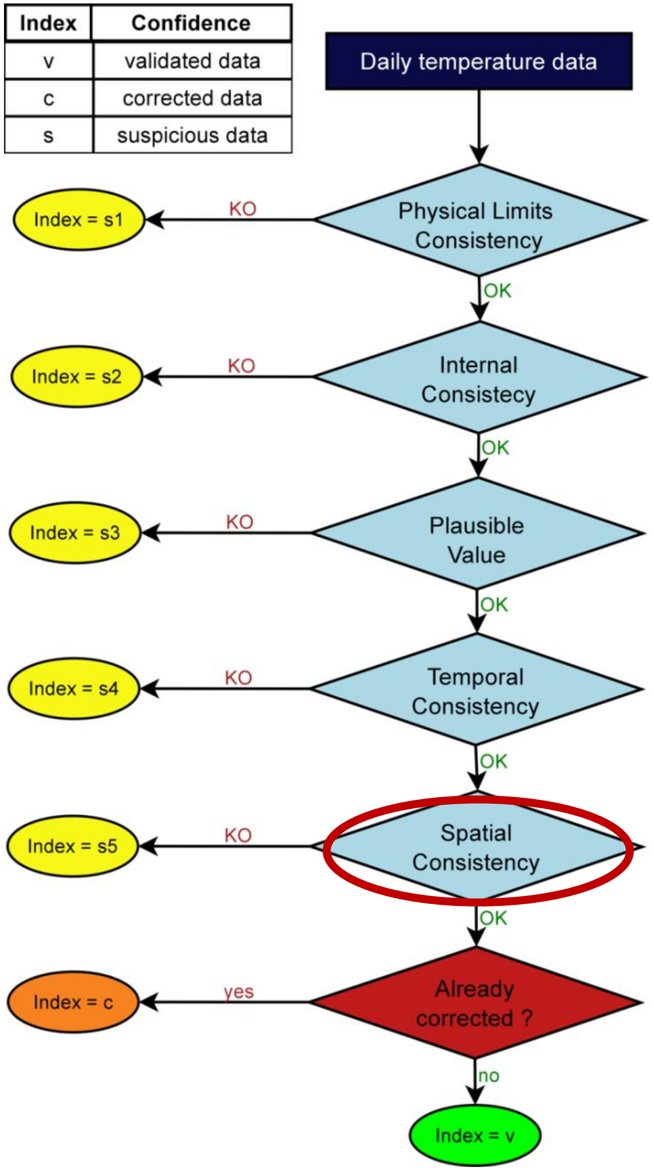
TX values suspicious (too warm) if :

TX > IDW + Standard Deviation + 6°C

AND

TX > TX *(of the 5 neighbors!)* + 4°C

Quality control – Part 2



2) Trend test

- Neighbor comparison of daily rise/drop of temperature

Date	TX (Leuven)	Trend	5 closest neighbors	Trend Test
9/07/42	26.1	..	..	..
10/07/42	16.3	↓↓↓	ALL ↓↓↓	OK
11/07/42	19.2	↑↑↑	ALL ↓↓↓	KO

↑↑↑ = *temperature increase of more than 2 degrees*

↓↓↓ = *temperature decrease of more than 2 degrees*

Quality control – Part 2

- Quality Control procedures is realized in **two times**

-> First run allows to assign a first quality index to all the data
-> Second run takes only validated data for spatial tests

Date	TX (Leuven)	Trend	5 closest neighbors	Quality Index
9/07/1942	26.1	..		..
10/07/1942	16.3 (**s51!**)	↓↓↓	ALL ↓↓↓	s51
11/07/1942	19.2	~~↑↑↑~~	ALL ↓↓↓	~~**S52**~~ **v**

Quality control – Part 2

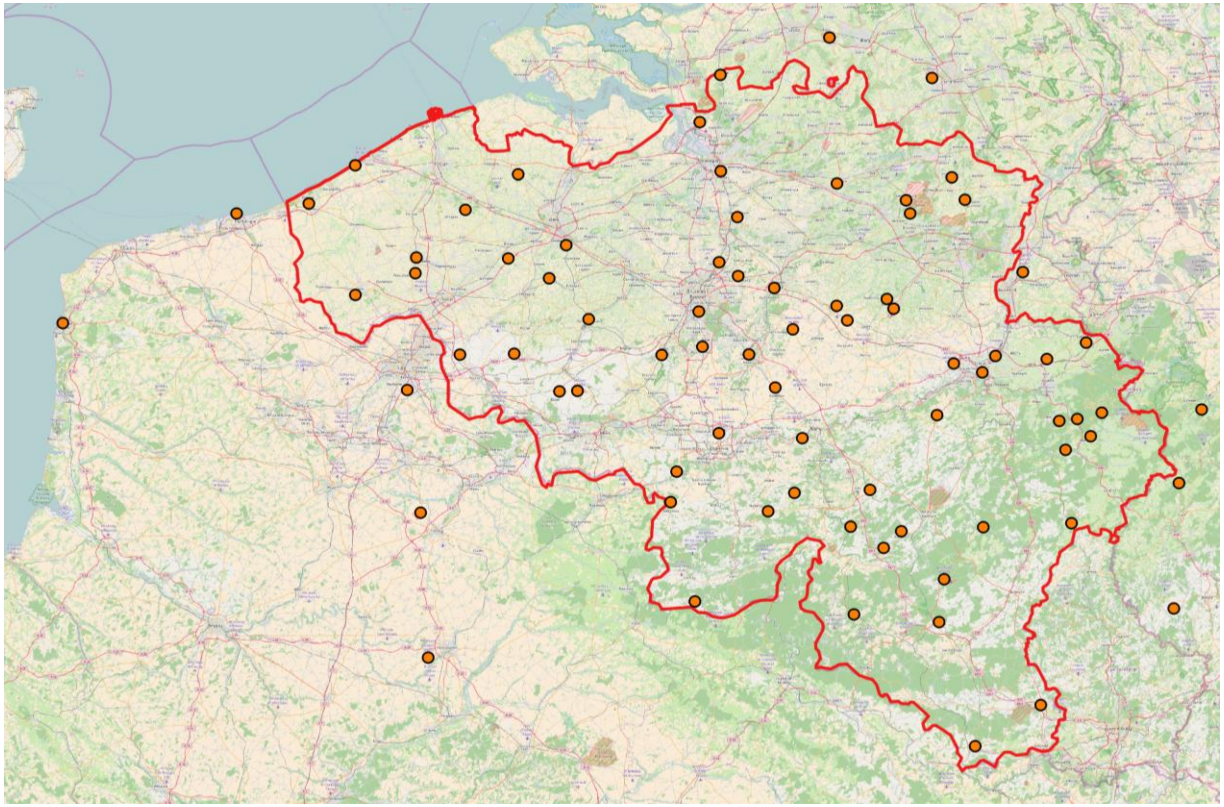
Results :
~ 99.5 % of validated temperature data (about 10000 values)
Can be explained by the basic QC already made for data from 1951
About 80 % of the suspicious values are detected by spatial tests

Some corrections when it was possible (especially for *s1* and *s2*)
Only when no doubt

Homogenisation of temperature short series

Station locations

- 66 stations (short series TT)
- From 1951 to 2015
- At least 90 % of daily data
- 11 foreign stations
  - 5 FR
  - 3 GE
  - 3 DE

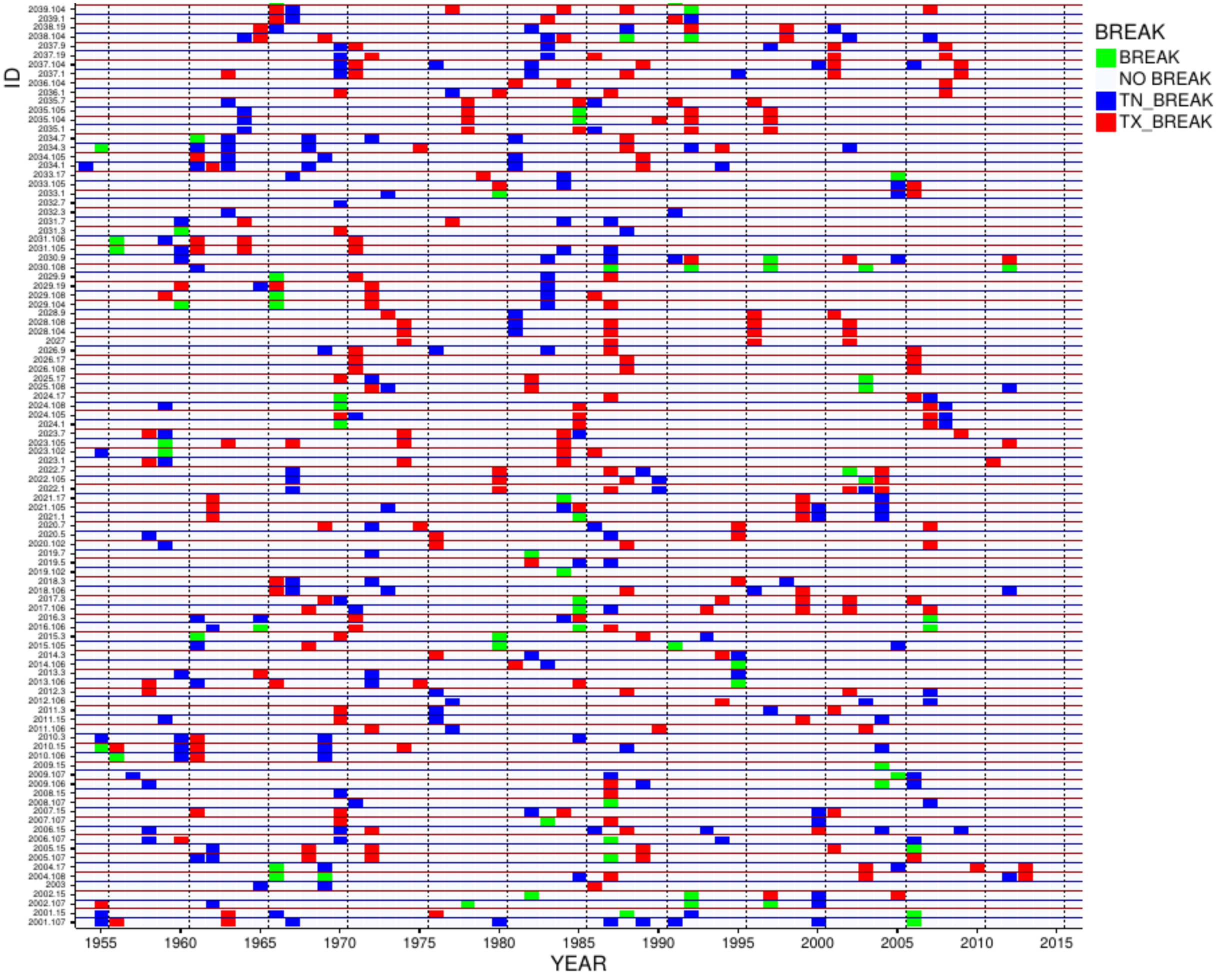


Metadata

- Station catenations
- Shelter relocation
- Change of shelter type
- Change of instrument
- Automatisation
- Change of observer
- Other things like information on the shelter site

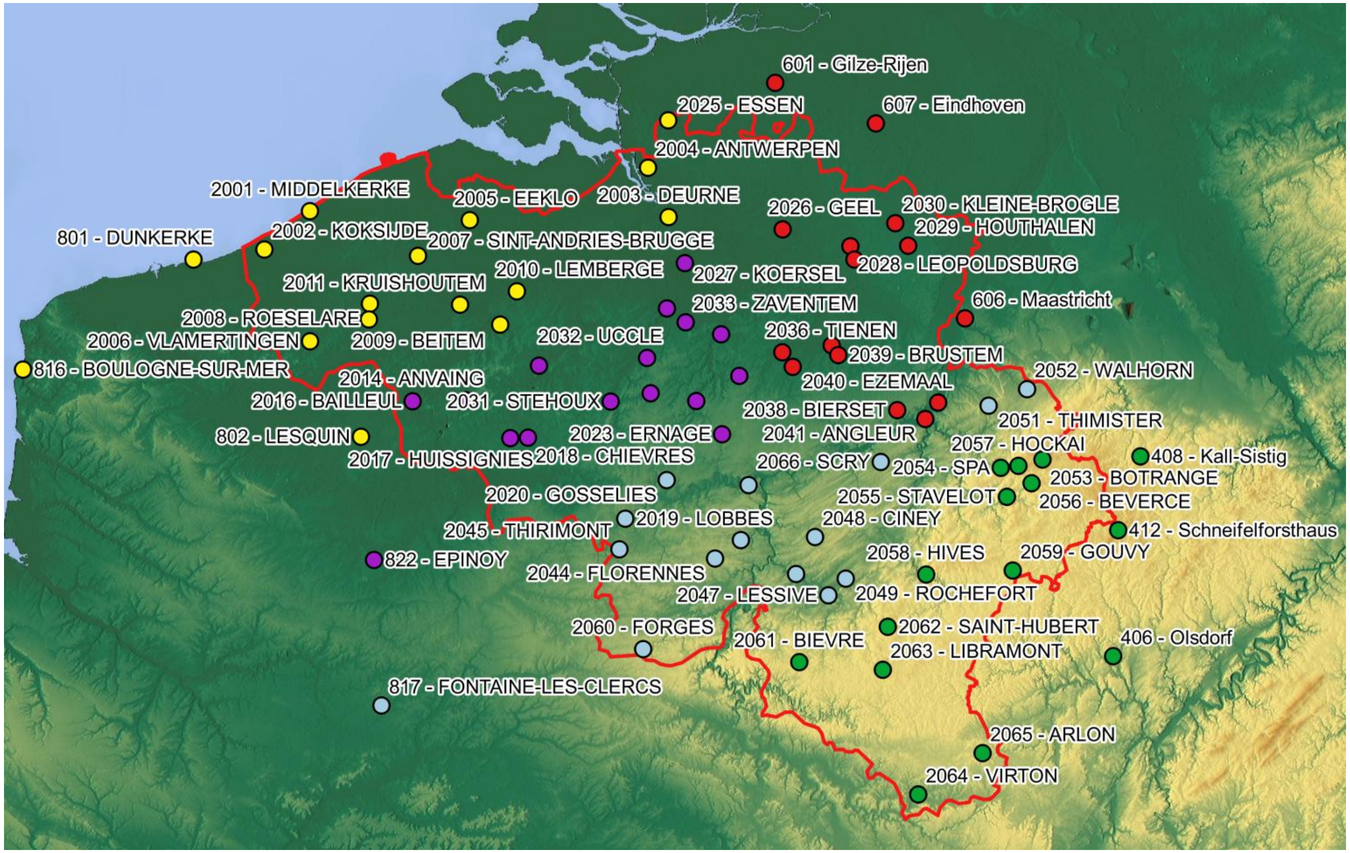
Methodology

- HOMER
- Trainings
- 3 people working separatly on different cluster composition
- Common breaks (usually big ones)
- Improving breaks list
- Re-do homogenisation with final clusters



Clusters

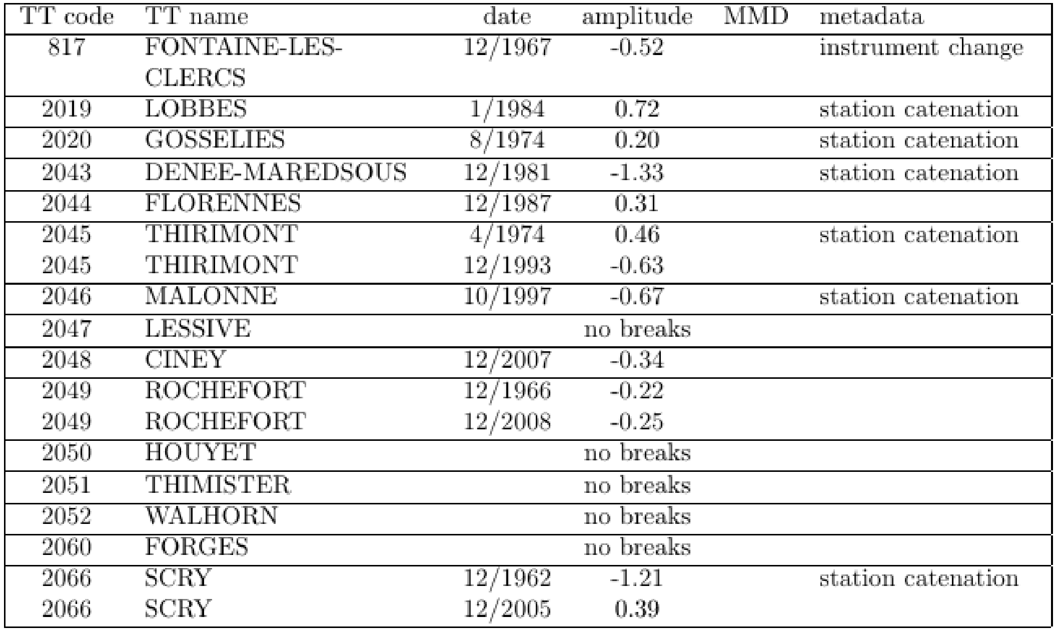
Creation based on :
- proximity
- correlation
- climatic area



- 5 clusters
- Around 15 stations

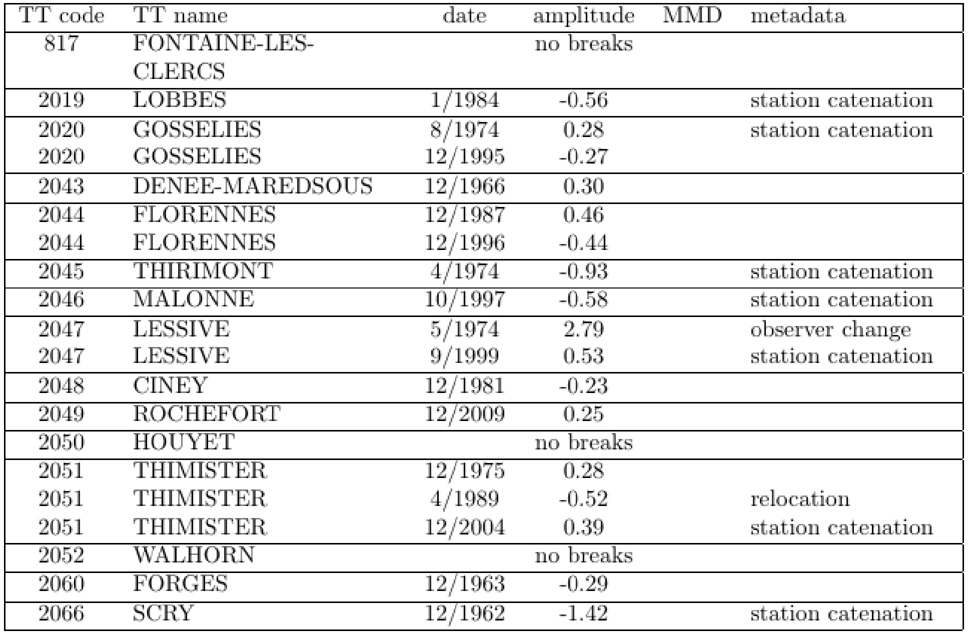
First results with HOMER

TN



TT code	TT name	date	amplitude	MMD	metadata
817	FONTAINE-LES-CLERCS	12/1967	-0.52		instrument change
2019	LOBBES	1/1984	0.72		station catenation
2020	GOSSELIES	8/1974	0.20		station catenation
2043	DENEE-MAREDSOUS	12/1981	-1.33		station catenation
2044	FLORENNES	12/1987	0.31		
2045	THIRIMONT	4/1974	0.46		station catenation
2045	THIRIMONT	12/1993	-0.63		
2046	MALONNE	10/1997	-0.67		station catenation
2047	LESSIVE		no breaks		
2048	CINEY	12/2007	-0.34		
2049	ROCHEFORT	12/1966	-0.22		
2049	ROCHEFORT	12/2008	-0.25		
2050	HOUYET		no breaks		
2051	THIMISTER		no breaks		
2052	WALHORN		no breaks		
2060	FORGES		no breaks		
2066	SCRY	12/1962	-1.21		station catenation
2066	SCRY	12/2005	0.39		

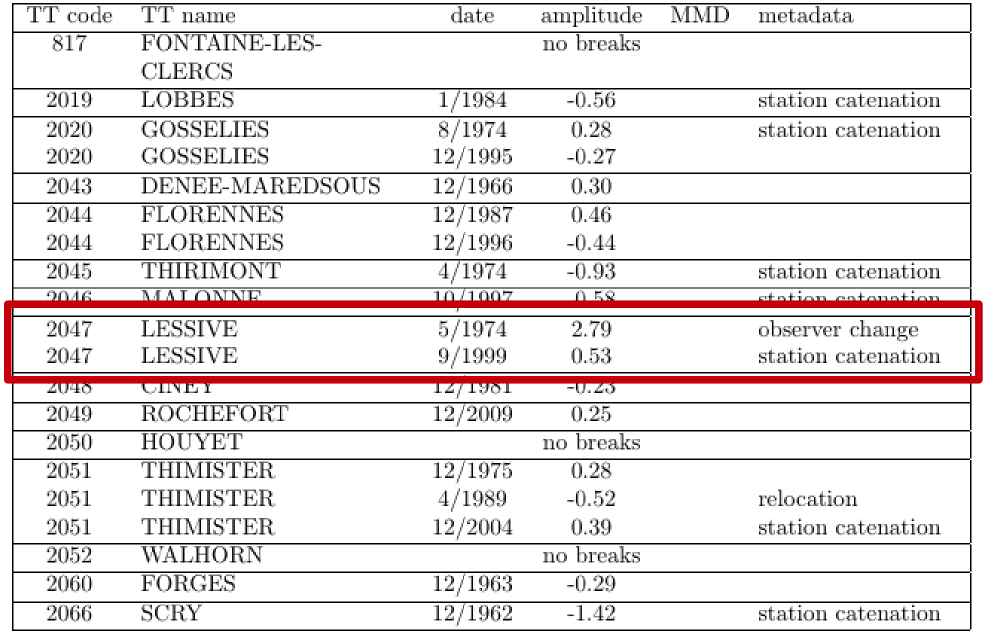
TX



TT code	TT name	date	amplitude	MMD	metadata
817	FONTAINE-LES-CLERCS		no breaks		
2019	LOBBES	1/1984	-0.56		station catenation
2020	GOSSELIES	8/1974	0.28		station catenation
2020	GOSSELIES	12/1995	-0.27		
2043	DENEE-MAREDSOUS	12/1966	0.30		
2044	FLORENNES	12/1987	0.46		
2044	FLORENNES	12/1996	-0.44		
2045	THIRIMONT	4/1974	-0.93		station catenation
2046	MALONNE	10/1997	-0.58		station catenation
2047	LESSIVE	5/1974	2.79		observer change
2047	LESSIVE	9/1999	0.53		station catenation
2048	CINEY	12/1981	-0.23		
2049	ROCHEFORT	12/2009	0.25		
2050	HOUYET		no breaks		
2051	THIMISTER	12/1975	0.28		
2051	THIMISTER	4/1989	-0.52		relocation
2051	THIMISTER	12/2004	0.39		station catenation
2052	WALHORN		no breaks		
2060	FORGES	12/1963	-0.29		
2066	SCRY	12/1962	-1.42		station catenation

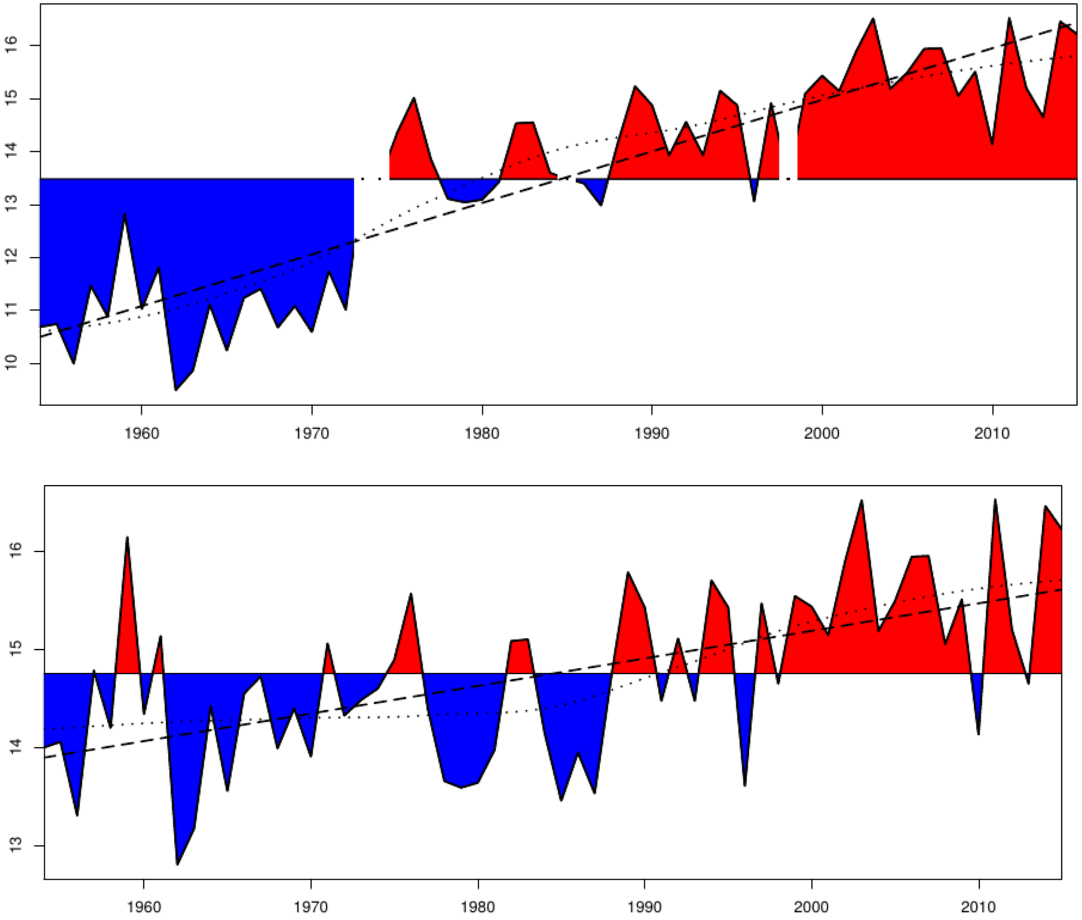
First results with HOMER

TX



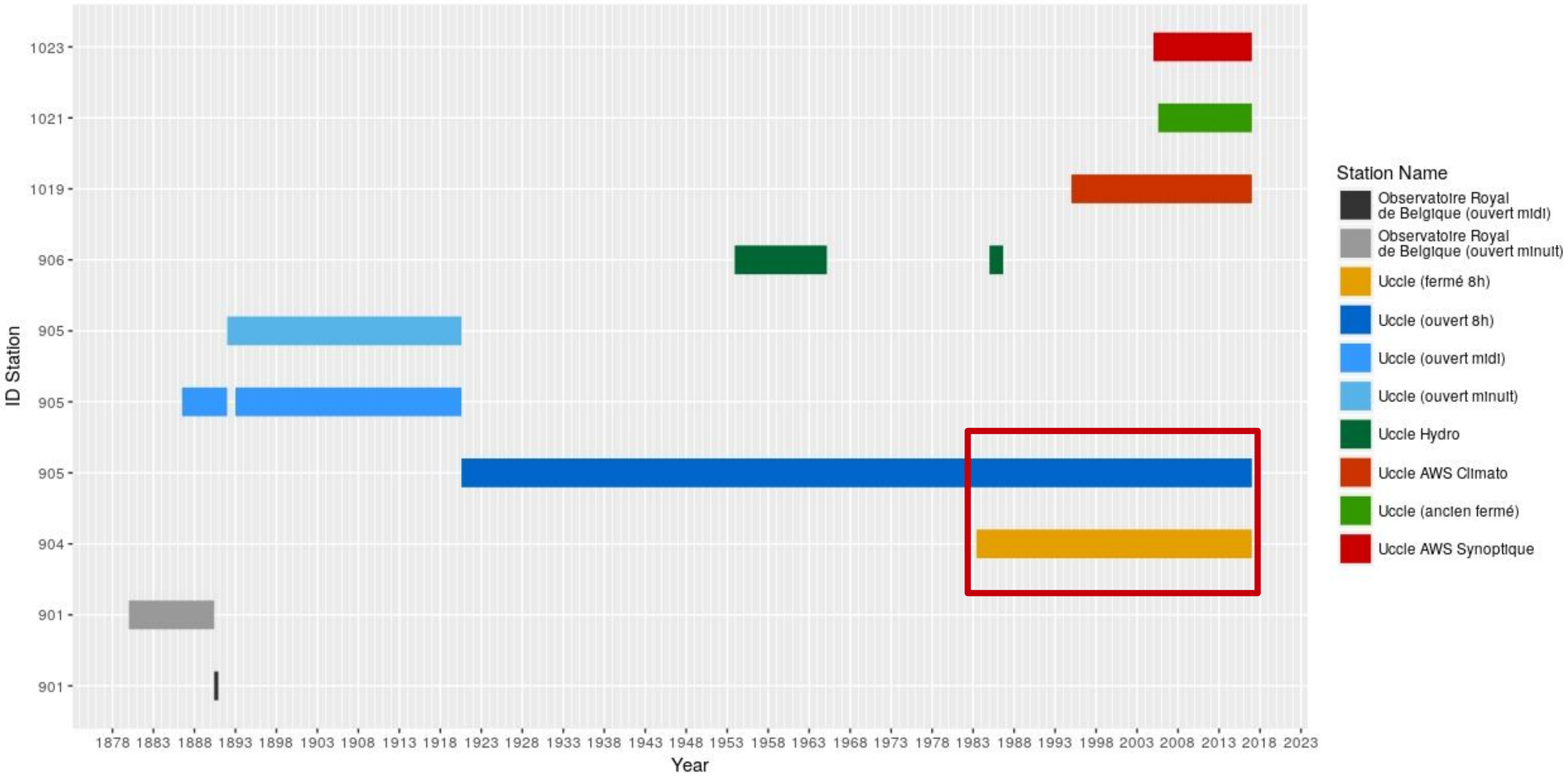
TT code	TT name	date	amplitude	MMD	metadata
817	FONTAINE-LES-CLERCS		no breaks		
2019	LOBBES	1/1984	-0.56		station catenation
2020	GOSSELIES	8/1974	0.28		station catenation
2020	GOSSELIES	12/1995	-0.27		
2043	DENEE-MAREDSOUS	12/1966	0.30		
2044	FLORENNES	12/1987	0.46		
2044	FLORENNES	12/1996	-0.44		
2045	THIRIMONT	4/1974	-0.93		station catenation
2046	MALONNE	10/1997	0.58		station catenation
2047	LESSIVE	5/1974	2.79		observer change
2047	LESSIVE	9/1999	0.53		station catenation
2048	CINEY	12/1981	-0.23		
2049	ROCHEFORT	12/2009	0.25		
2050	HOUYET		no breaks		
2051	THIMISTER	12/1975	0.28		
2051	THIMISTER	4/1989	-0.52		relocation
2051	THIMISTER	12/2004	0.39		station catenation
2052	WALHORN		no breaks		
2060	FORGES	12/1963	-0.29		
2066	SCRY	12/1962	-1.42		station catenation

Lessive station



Compare HOMER with something else ?

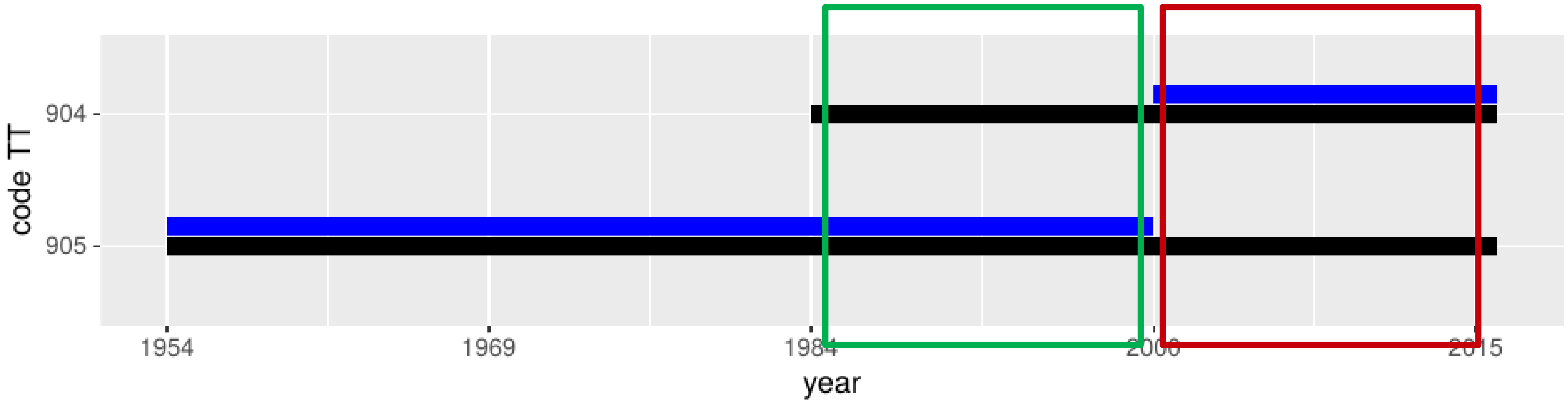
Parallel data on Uccle site→ extrapolation by linear regression

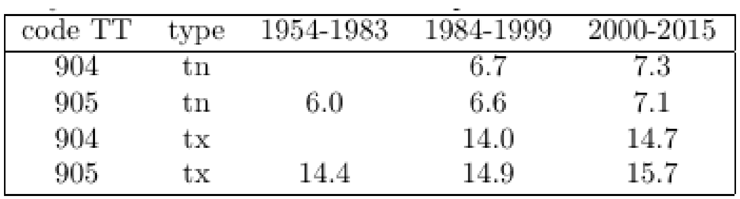


Linear regression

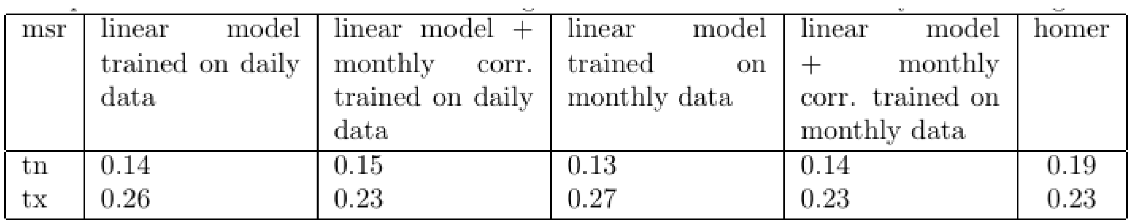
Sensitivity study between HOMER and linear regression

Two regression models where computed for TN and TX



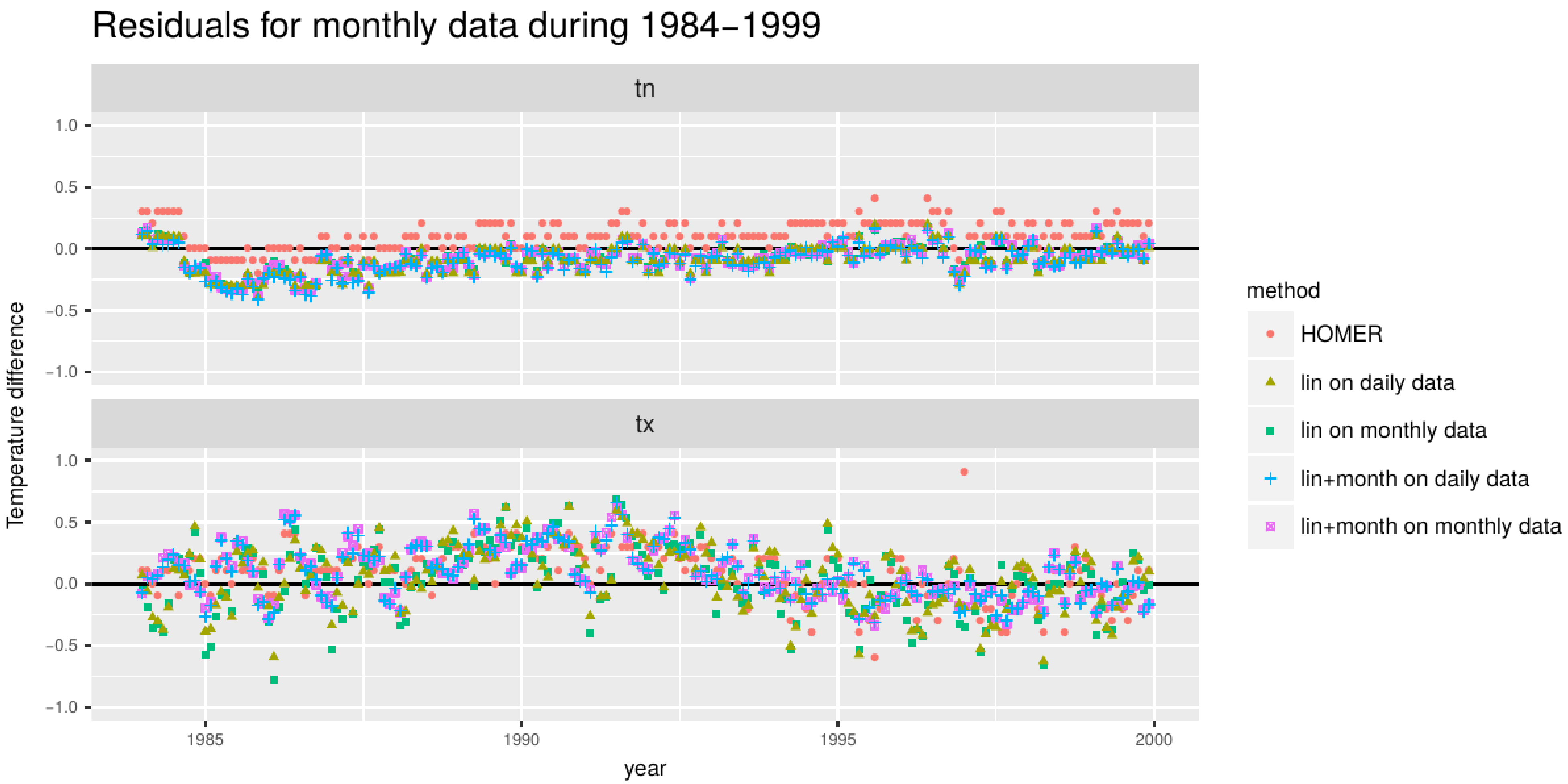


code TT	type	1954-1983	1984-1999	2000-2015
904	tn		6.7	7.3
905	tn	6.0	6.6	7.1
904	tx		14.0	14.7
905	tx	14.4	14.9	15.7



msr	linear model trained on daily data	linear model + monthly corr. trained on daily data	linear model trained on monthly data	linear model + monthly corr. trained on monthly data	homer
tn	0.14	0.15	0.13	0.14	0.19
tx	0.26	0.23	0.27	0.23	0.23

Linear regression



HOMER oddities

- Begining and ending of series (interpolation and homogenisation)
- Climatic events
- ACMANT ? Didn’t use
- Order of break implantation consequence
- Difference TX/TN ?

Conclusion & perspectives

Sometimes very difficult to get all metadata, especially for older stations before 50’

Homer → Human factor very important, be very careful
→ Sensitivity test ok

What’s next ? → Spatial interpolation of homogenised temperature, long series and precipitation, daily homogenisation