Data Quality Control applied on ECA&D

Petr Štěpánek[1,2], Gerard van der Schrier[3], Pavel Zahradnícek[1,2]

Climate Change

[1] Global Change Research Institute, Czech Academy of Sciences, Brno, Czech Republic

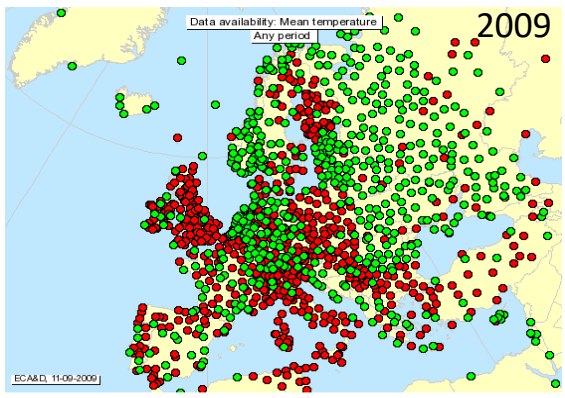
[2] Czech Hydrometeorological Institute, Brno, Czech Republic

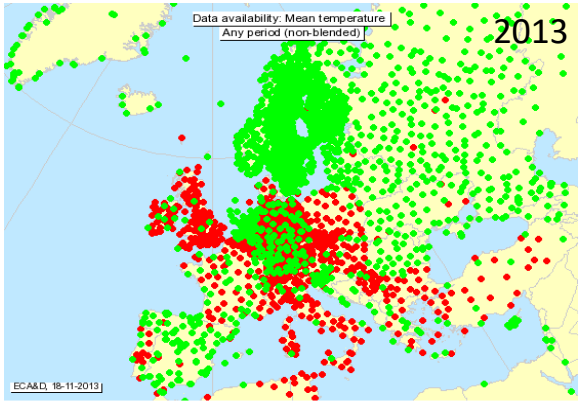
[3] Royal Netherlands Meteorological Institute, De Bilt, the Netherlands

Copernicus Europe's eyes on Earth | European Commission | ECMWF

Climate Change

Introduction





~200 stations

2.896 stations

7.848 stations

Status 2023: 23.317 stations

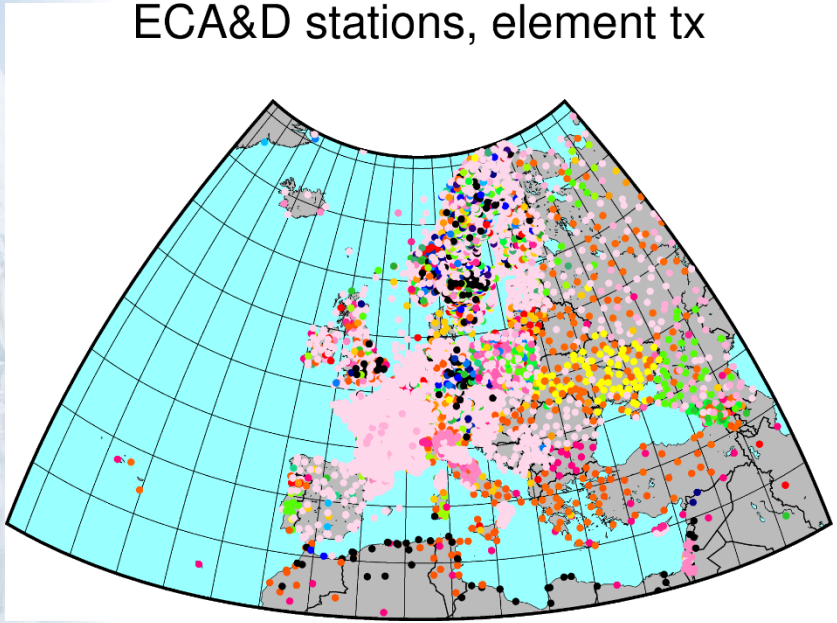
But….the QC method has remained the same in the past 20 years

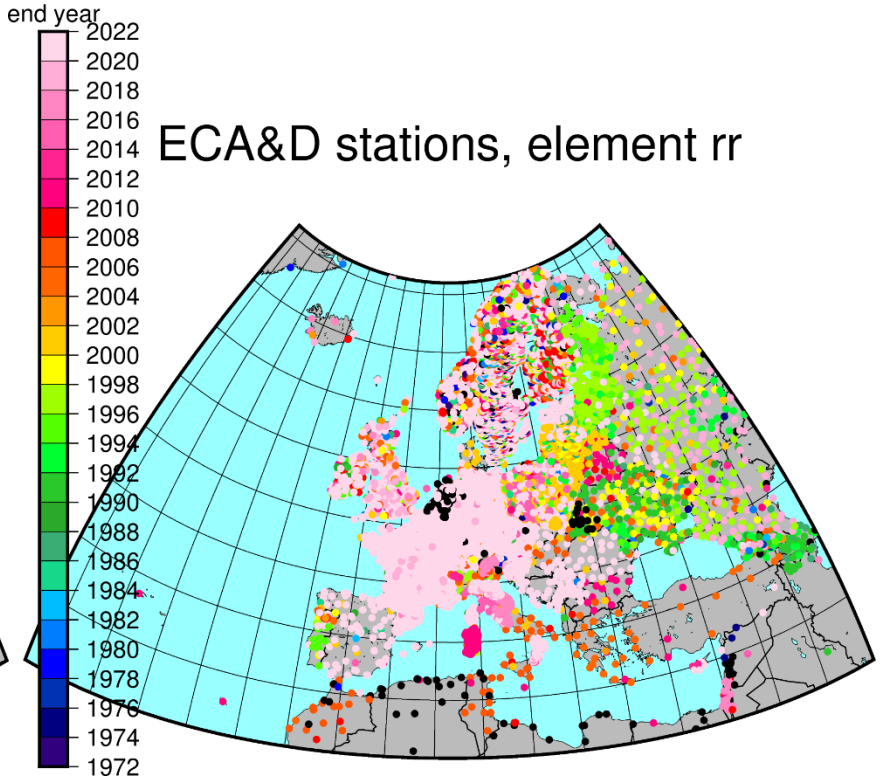
Can we benefit from the high station density by adding *inter-station comparisons* to QC?

ECMWF Copernicus Europe's eyes on Earth European Commission

Introduction

Climate Change





Issues to face

- Inhomogeneous station density
- Series are of variable length

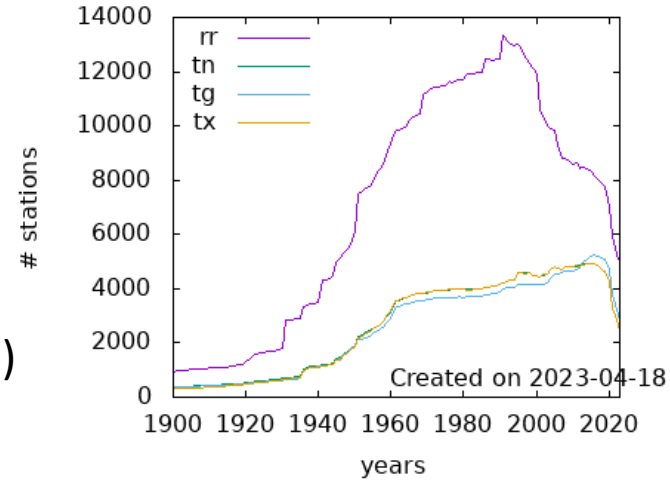
ECMWF Copernicus Europe's eyes on Earth European Commission

Climate Change

Introduction

Requirements

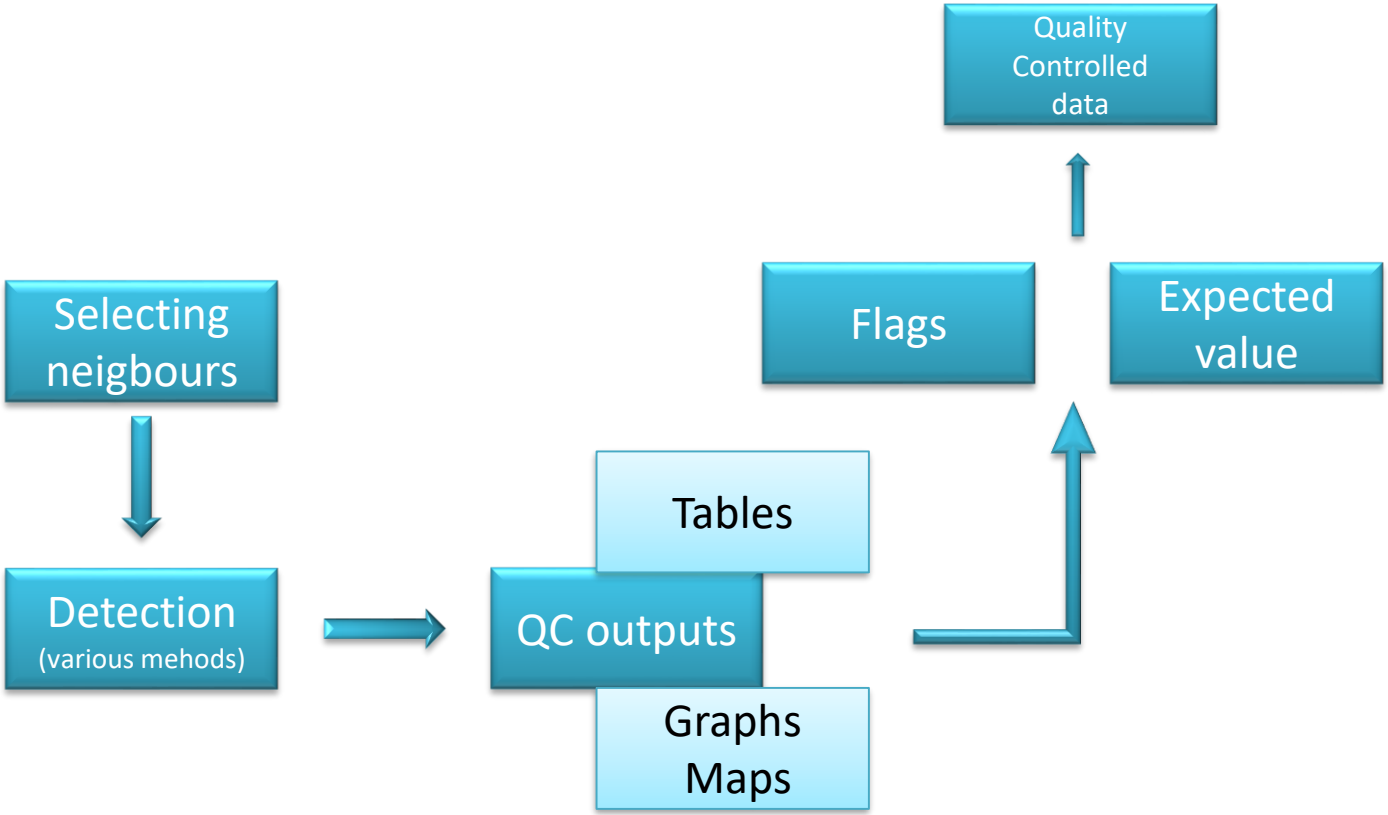
- Automated method
- Flexible to take-on all elements (temp., wind, etc.)
- Suggestion of alternative value
- Needs to be able to handle ‘messy’ data
  - combination of short and long series
  - gappy data
  - …..and we might have some duplicates
- Possibility to produce reports to feed-back to NMSs



ECMWF Copernicus Europe's eyes on Earth European Commission

Climate Change

Quality control- MetQC



ECMWF Copernicus Europe's eyes on Earth European Commission

Climate Change

QC outputs - flags

Inspired by other softwares

<u>All checked data are flagged:</u>

0 … valid
1 … error value (70/100% probability of error)
2 … suspect value (40/70% probability of error)
4 … repeated value (the same values repeated several times)
5 … duplicate value (same value found in neighbour station)
9 … missing value

ECMWF Copernicus Europe's eyes on Earth European Commission

Climate Change

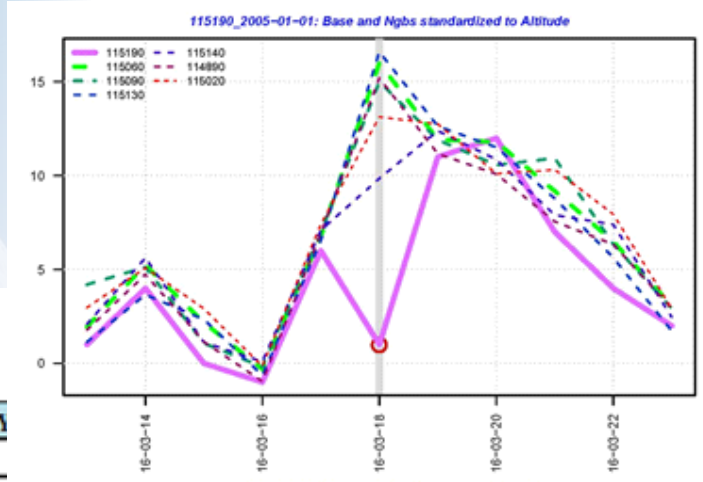
QC outputs – table, graphs, maps

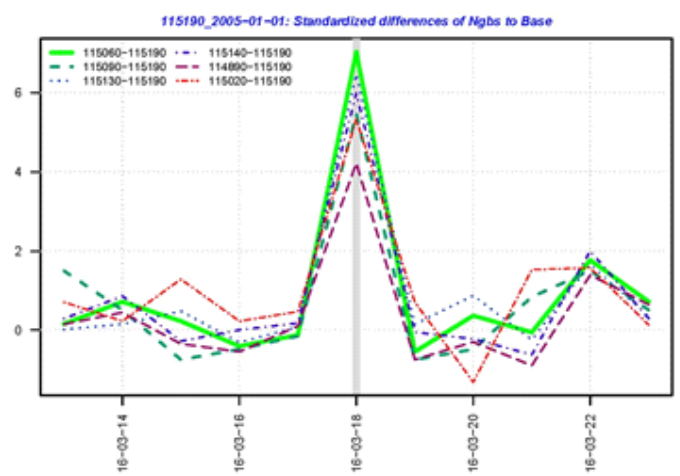
- Tables with errors
- Tables with suspisious values
- Tables with repeating values
- Tables with duplicity stations

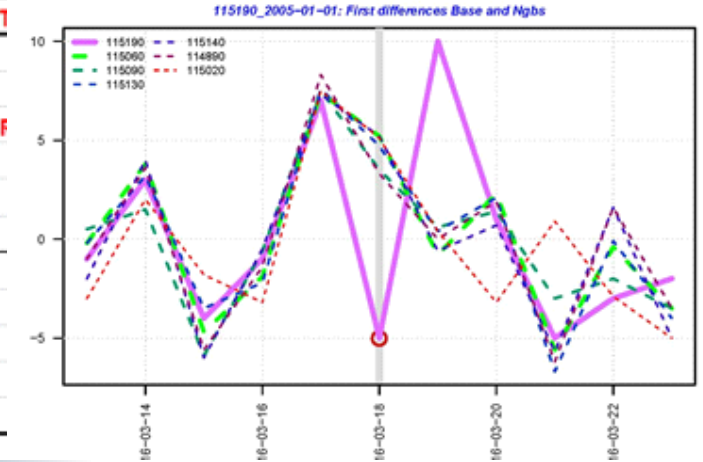
	Date			Test station	Calculate value	Difference		Reference stations					
ID	YEAR	MONTH	DAY	ST_BASE	EXPECT_VAL	DIFFS	REMARK	ST_1	ST_2	ST_3	ST_4	ST_5	ST_6
							Distances	64	110	116	128	142	136
128522	TEST STATION			120.0			Altitudes,limit	110	112	169	110	450	112
122702							st_1, Correl	0.8					
100149							st_2, Correl		0.8				
132674	REFERENCE STATIONS						st_3, Correl	Correlation coef.		0.8			
123506							st_4, Correl				0.8		
126122							st_5, Correl					0.7	
100183							st_6, Correl						0.7
128522	1950	4	10	19.9	9.4	-10.5		9.6	9.0	8.0	8.4	5.3	10.0
128522	1950	11	1	12.4	4.7	-7.7		5.3	4.3	3.5	4.6	4.2	4.6
128522	1951	12	24	9.4	0.6	-8.8		-0.4	0.5	0.0	2.3	2.0	-0.8
128522	1953	11	24	-0.6	6.8	7.4		7.5	7.2	6.2	7.9	5.1	6.0
128522	1959	11	26	10.5	3.6	-6.9		2.8	4.5	2.9	4.1	5.6	2.7

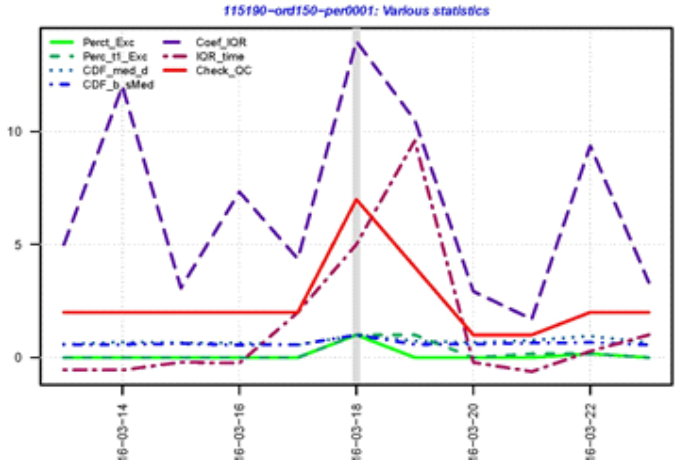
QC outputs – table, graphs, maps

Climate Change





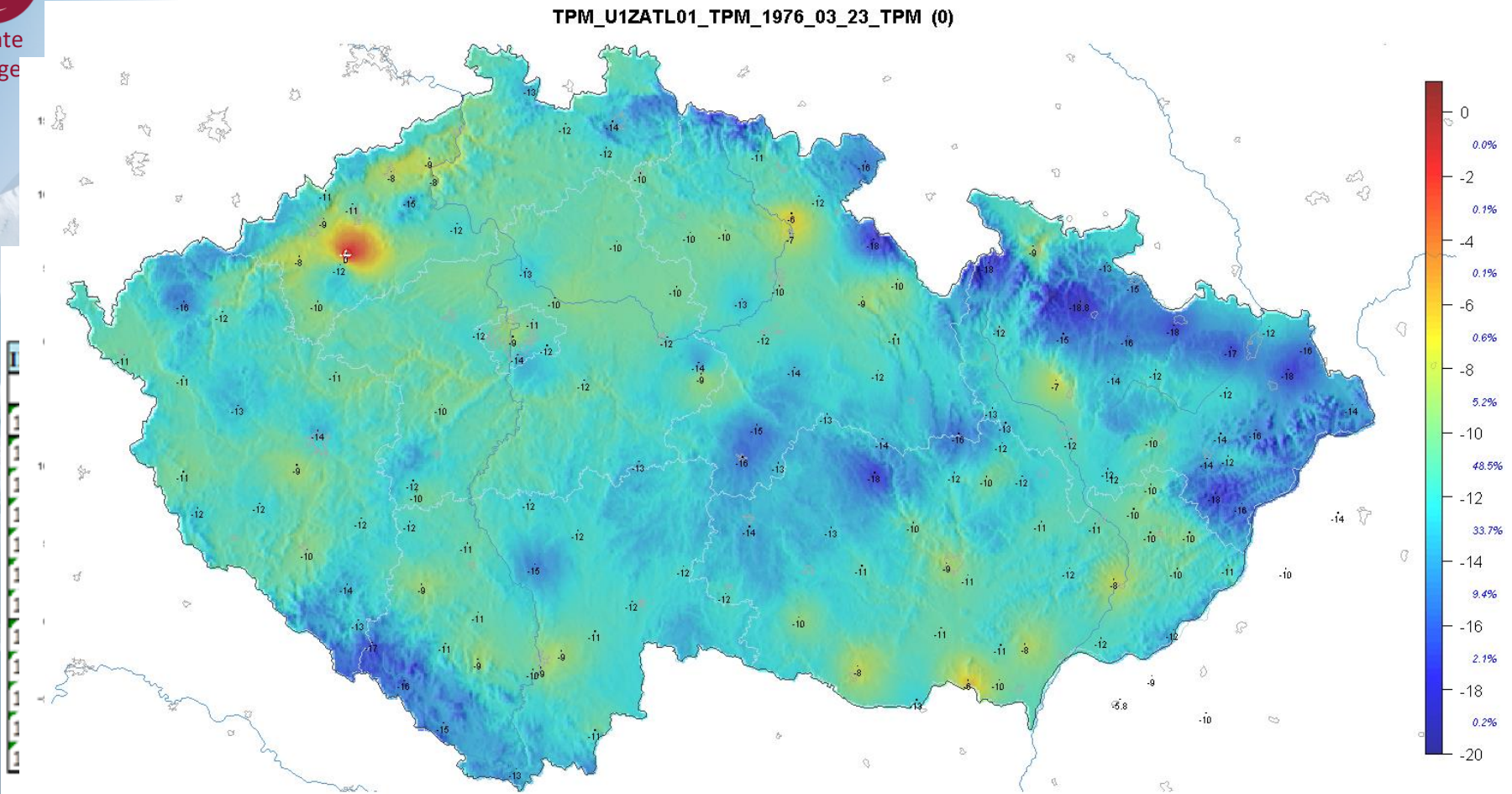




ID	ST_5	ST_6
128522	142	136
122702	450	112
100149		
132674		
123506		
126122	0.7	
100183		0.7
128522	5.3	10.0
128522	4.2	4.6
128522	2.0	-0.8
128522	5.1	6.0
128522	5.6	2.7

Climate Change

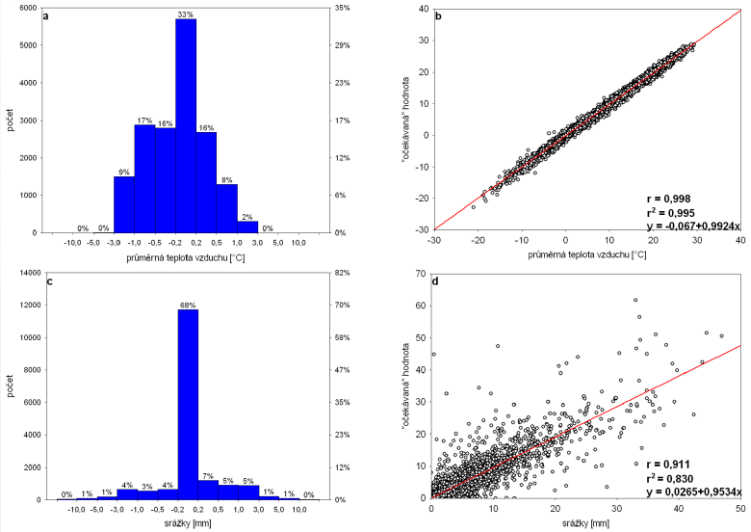
QC outputs – table, graphs, maps



Climate Change

Expected value

- "Expected" value is very important tool in QC
- calculated solely from neighboring stations
- used for a comparison with candidate station value
- the "expected" value serves for QC evaluation, but it can be used also for filling missing values, or to replace wrong measurement, if needed



Validation
air temperatue,
0.998 correlation coeficient between calculated and original values

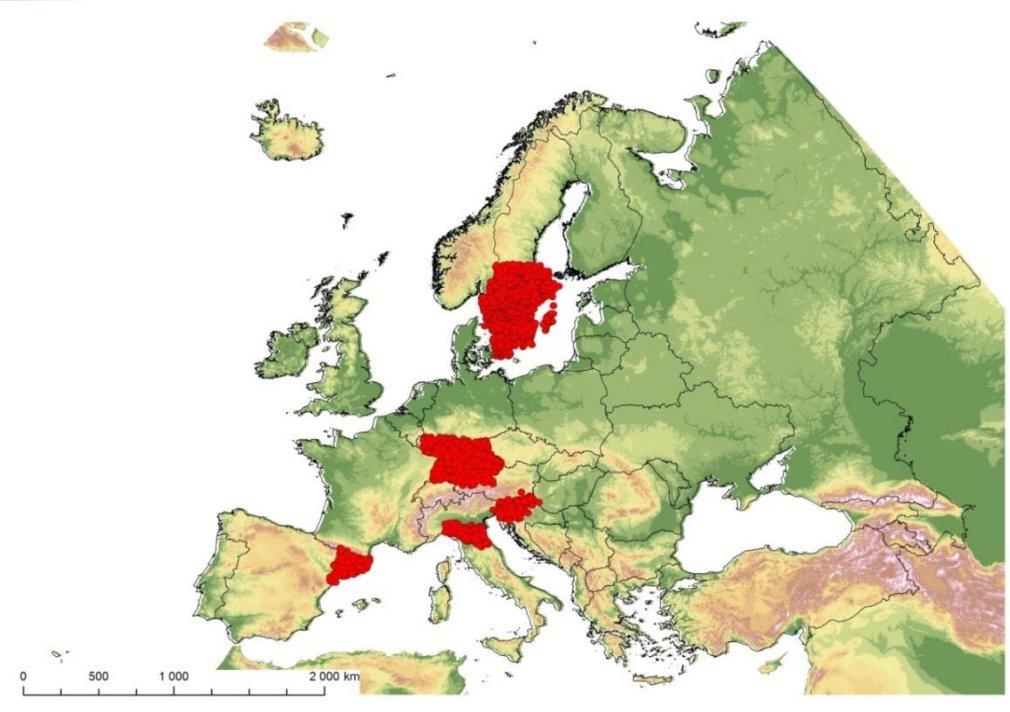
precipitation

ECMWF Copernicus Europe's eyes on Earth European Commission

Climate Change

Datasets for methods evaluation

- To evaluate the data quality control, "*real*" benchmark dataset based on 4 selected European regions was created, consisting of 1042 stations in total



- Catalonya was removed, short series with many problems
- Time series contain gaps and errors

ECMWF Copernicus Europe's eyes on Earth European Commission

Climate Change

Datasets for methods evaluation

- „Real“ benchmark dataset has been created in the way, that any error detection, by several methods, has been taken into account and the value has been replace with missing value (four regions – countries from Europe)
- Besides real dataset, surrogate data have been used for methods evaluation (to satisfy both “climatological” and “mathematical” point of view)
- Surrogate dataset: "clear worlds" from ISTI initiative (selection of 2 of 4 USA regions)
- Known errors were introduces both into real and surrogate datasets

ECMWF Copernicus Europe's eyes on Earth European Commission

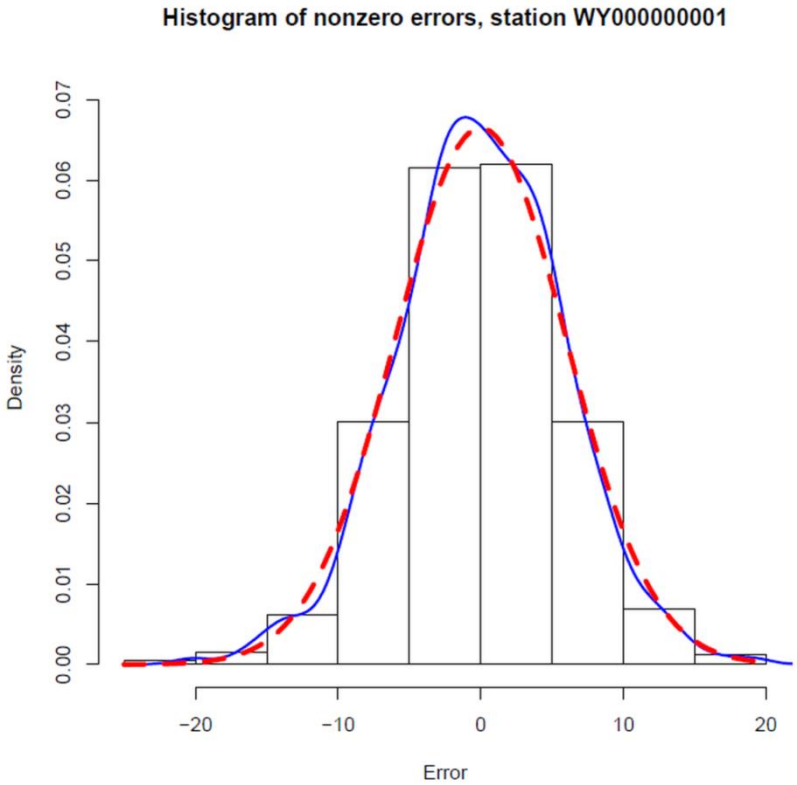
Climate Change

Introducing known errors into **real** and **surrogate** datasets

- For mean daily temperature we randomly input errors into each station of "clean world" **surrogate** datasets for Wyoming and South East region

- For maximum and minimum temperature of the **real** datasets

- We defined the error frequency equal to 5 % with randomly selected places in the dataset of each station, errors from the normal distribution with mean equal to 0 and standard deviation equal to 6



ECMWF Copernicus Europe's eyes on Earth European Commission

Climate Change

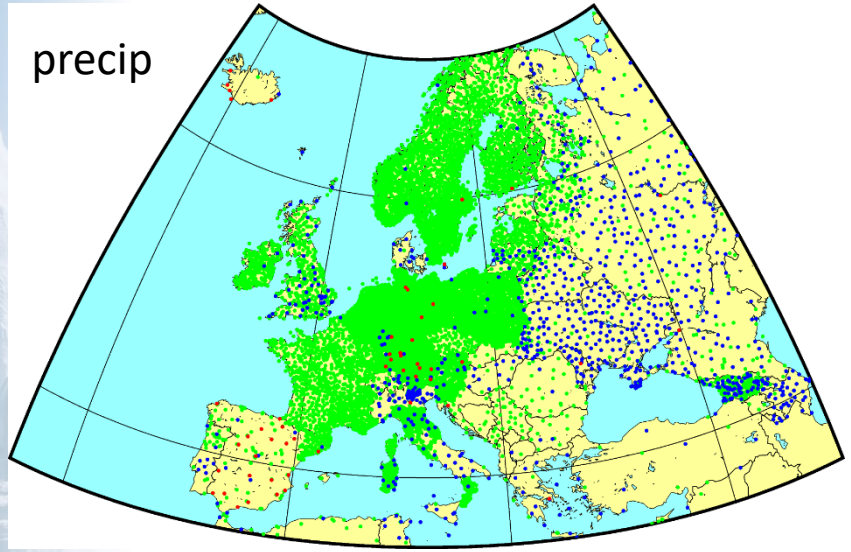
Errors detections

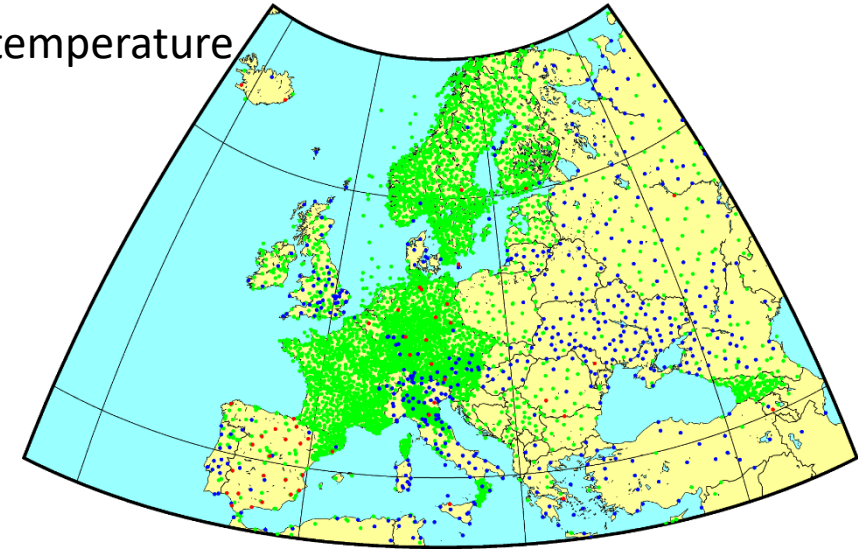
	Surrogate data		Absolute numbers			Totals		Percentages of errors		
SW Version	**REGION**	**Detection**	**HITS**	**FALSE ALARMS**	**MISSES**	**number of errors**	**total**	**HITS**	**FALSE ALARMS**	**MISSES**
Original 2018	SE	errors only	37761	**24**	79051	116 812	*2 347 020*	32.33	**0.02**	67.67
Original 2018	SE	errors and suspicious	56521	**428**	60291	116 812	*2 347 020*	48.39	**0.37**	51.61
Original 2018	Wyoming	errors only	8644	**34**	48416	57 060	*1 150 500*	15.15	**0.06**	84.85
Original 2018	Wyoming	errors and suspicious	16918	**732**	40142	57 060	*1 150 500*	29.65	**1.28**	70.35
update 2023	SE	errors only	17017	**1**	99795	116 812	*2 347 020*	14.57	**0.00**	85.43
update 2023	SE	errors and suspicious	31320	**44**	85492	116 812	*2 347 020*	26.81	**0.04**	73.19
update 2023	Wyoming	errors only	3277	**8**	53783	57 060	*1 150 500*	5.74	**0.01**	94.26
update 2023	Wyoming	errors and suspicious	7690	**74**	49370	57 060	*1 150 500*	13.48	**0.13**	86.52

ECMWF Copernicus Europe's eyes on Earth European Commission

Climate Change

Duplicate series check





Inter-station comparison identifies the duplicate series

'project' data have been added to ECA&D to rapidly increase coverage

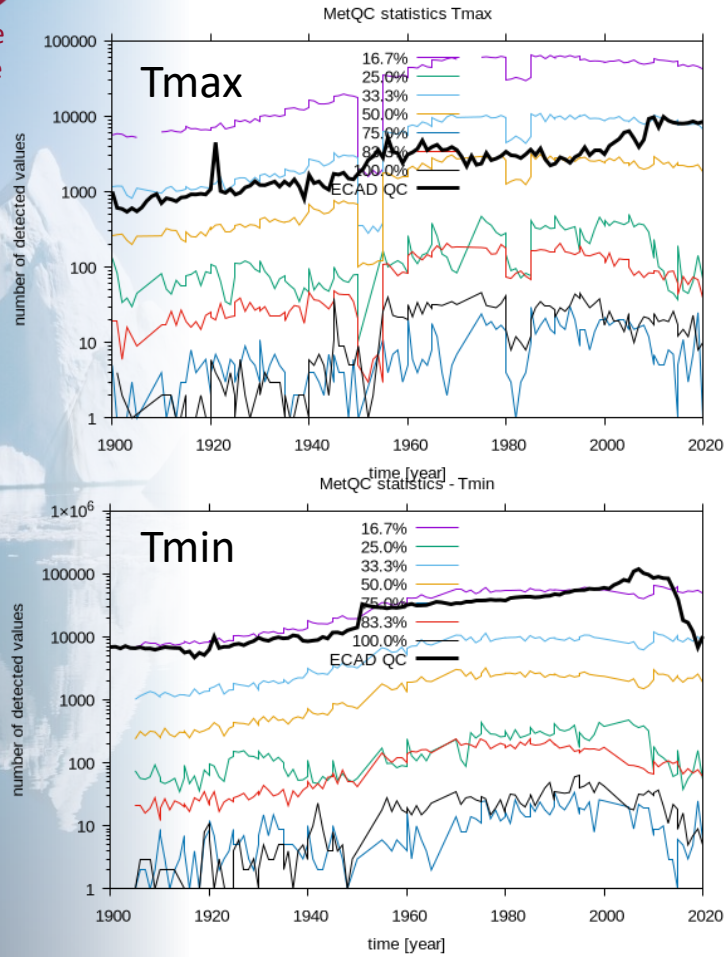
Data sources: NMHS, projects, Emulate

- 268 duplicate 'project' series deleted
- Overlap with 'Emulate' deleted

ECMWF Copernicus Europe's eyes on Earth European Commission

Implementation in ECA&D: Temperature

Climate Change



Black line:

flagged values using standard ECA&D

Coloured lines:

MetQC tests

Rule of thumb:

40% - 70% failed tests: -> ‘suspect’

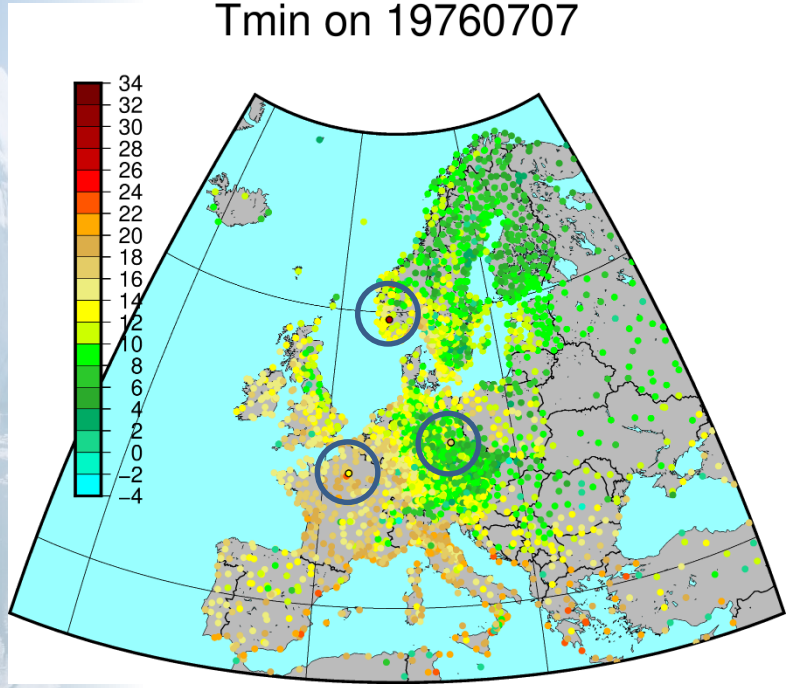
> 70% failed tests -> ‘error’

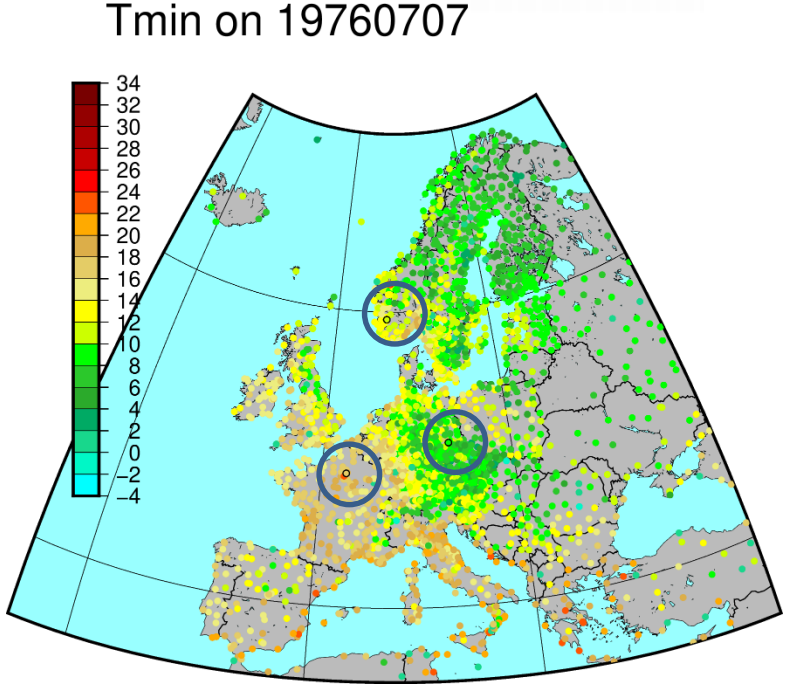
With MetQC

- ~200 - ~2000 additional ‘suspect’ values/yr
- ~20 - ~200 additional ‘error’ vaues/yr

Implementation in ECA&D: Temperature

Climate Change



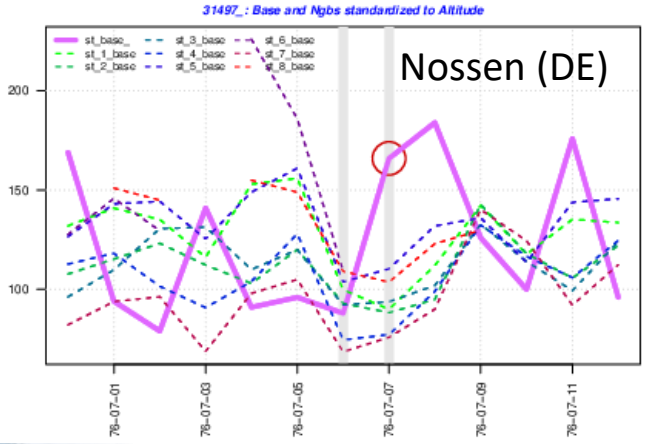


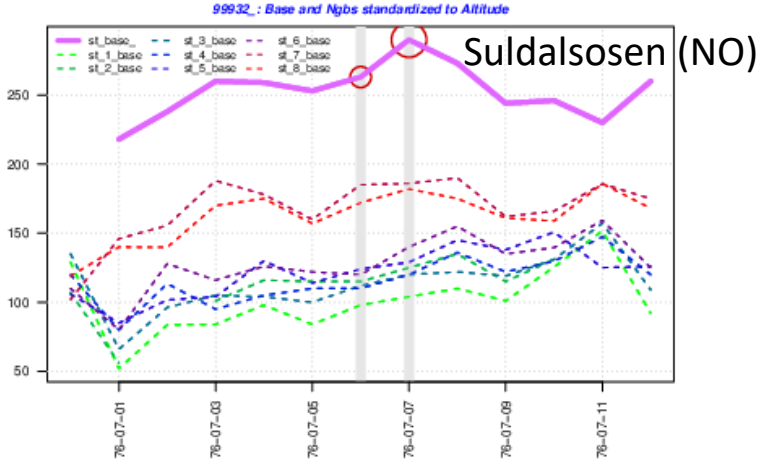
3 ‘error’ values found

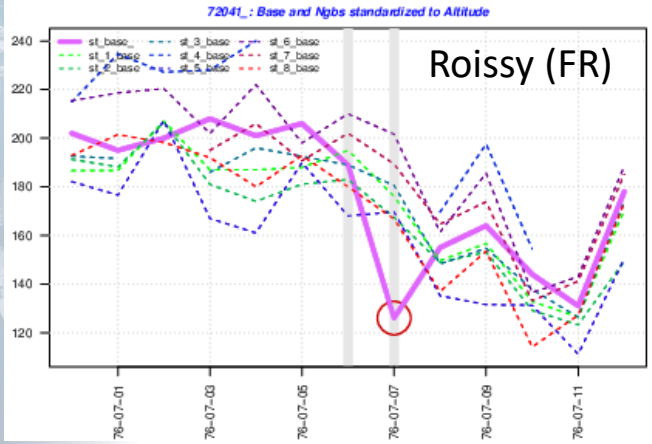
ECMWF Copernicus Europe's eyes on Earth European Commission

Implementation in ECA&D: Temperature

Climate Change





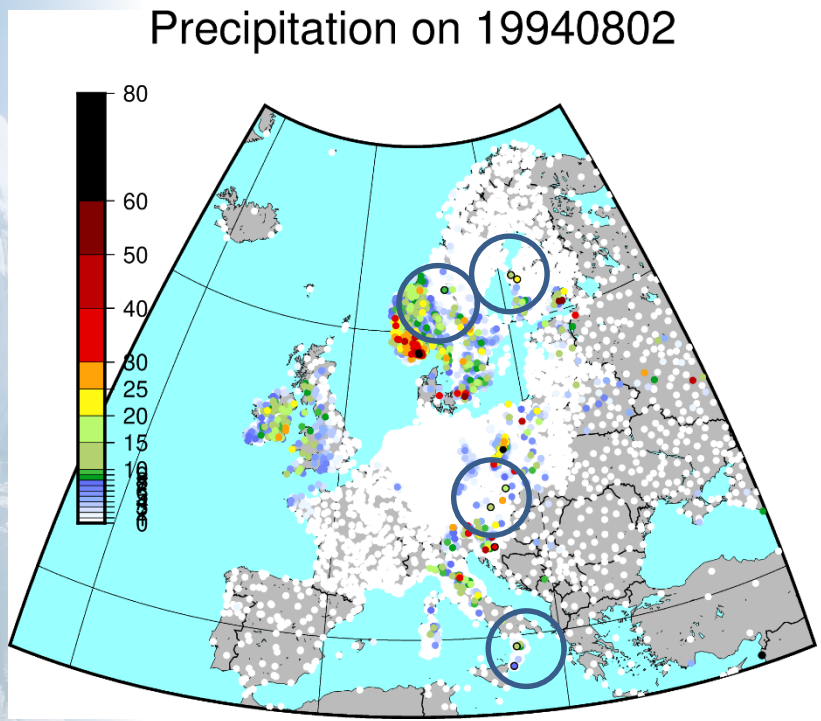


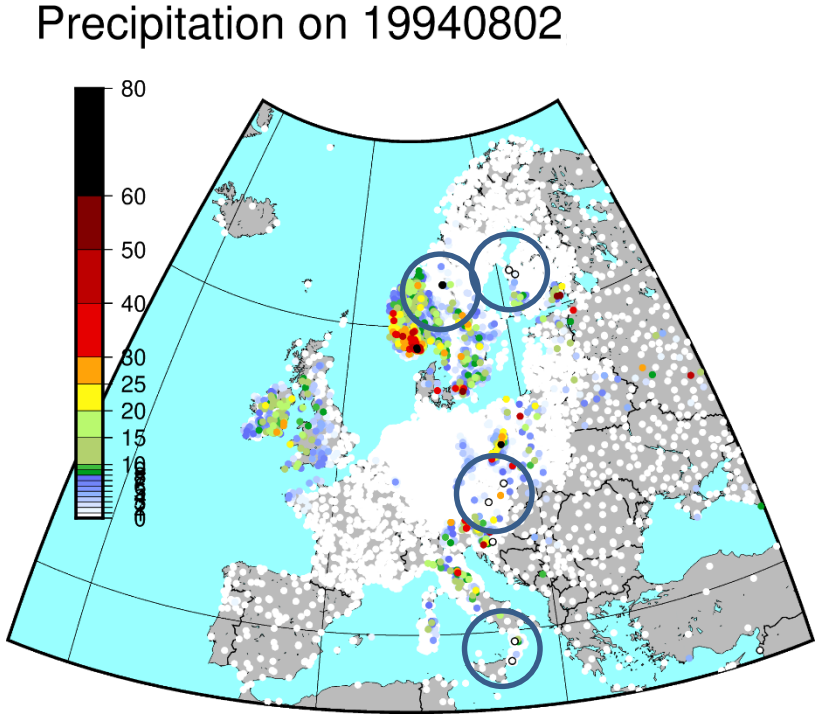
3 'error' values for 1976/07/07

- Norwegian series: not sure what is happening here – perhaps wrong metadata?

Climate Change

Implementation in ECA&D: Precipitation





'error' values in summer likely coincide with convective events -> *no flagging done*

ECMWF Copernicus Europe's eyes on Earth European Commission

Climate Change

General recommedations and comments

- Quality control must not be a “black box”, the user has to have full control and should be informed in detail about the process
- In averages (even daily ones) errors are masked. It is recommended to test unprocessed, directly measured data (e.g. observed hourly data)
- Crucial is selection of reference (neighbour) stations
- For automated method to give acceptable results, it should combine several statistical approaches
- Automated methods of QC are necessary for large datasets, but the user still needs to have a full control about the process
- Graphical outputs are beneficial
- More complicated meteorological elements (e.g. precipitation) should be validated on sufficiently dense station network. *Caution is required*