TOAR QUALITY CONTROL PROCEDURES AND PLANS

ENVRIPLUS, 06 NOV 2018

I MARTIN SCHULTZ AND NAJMEH KAFFASHZADEH

What is TOAR?

TOAR
tropospheric
ozone
assessment
report

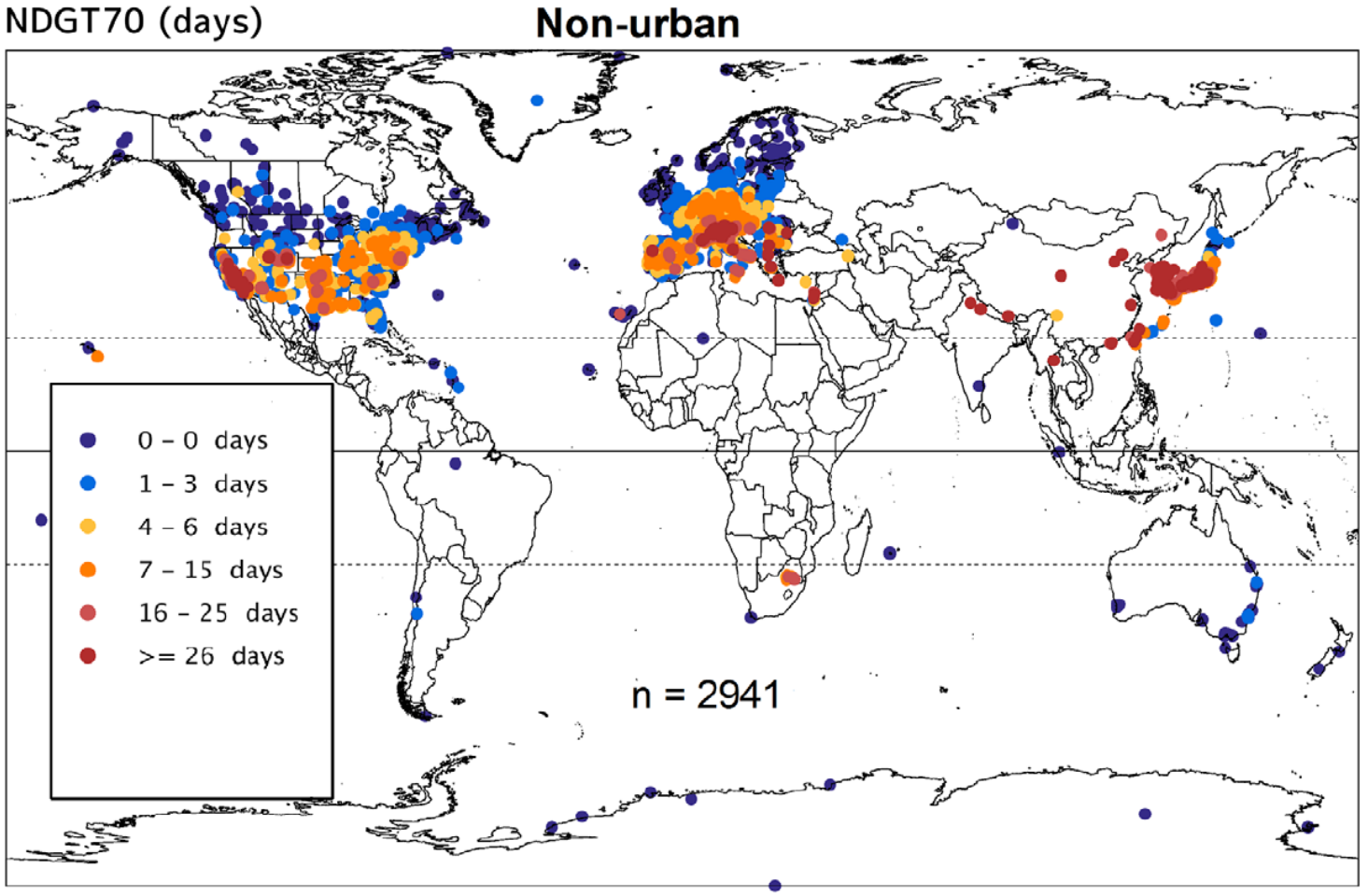
The „Tropospheric Ozone Assessment Report“ has been created by ~220 scientists from 36 countries to:

1. Produce the first tropospheric ozone assessment report based on the peer-reviewed literature and new analyses.

2. Generate easily accessible, documented data on ozone exposure and dose metrics at hundreds of measurement sites around the world (urban and non-urban), freely accessible for research on the global-scale impact of ozone on climate, human health and crop/ecosystem productivity.

Web link: http://www.igacproject.org/activities/TOAR

The TOAR database



- A data portal for advanced analysis of harmonized global air quality data
- A second-level archive combining data from various sources
- Relies on proper data management within national and regional networks
- Focus on ozone and its precursors (+ PM2.5)

Currently ~10000 stations

Why QA?

TOAR
tropospheric ozone assessment report

Theory: Data should be fully QA‘d by primary data centers

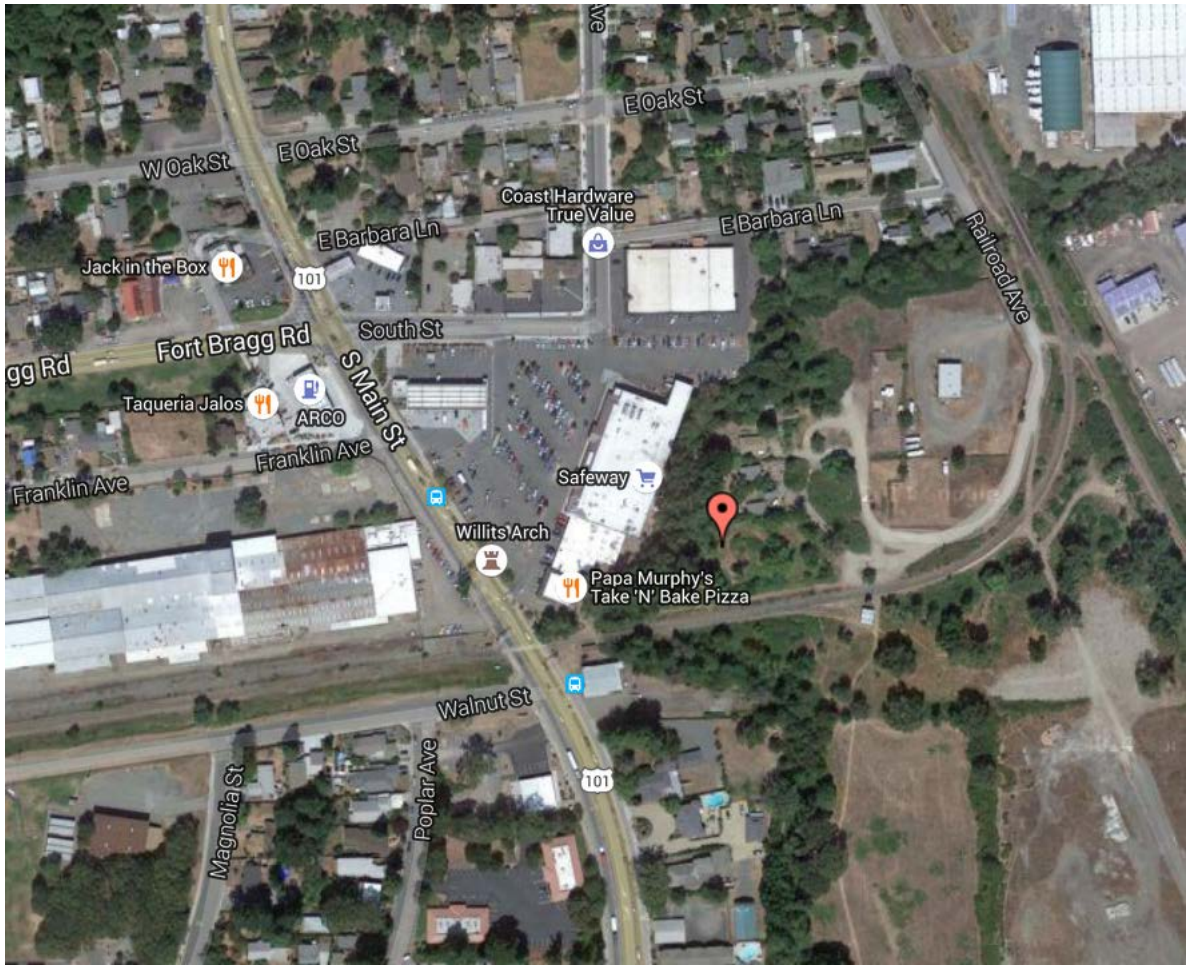
Practice: They normally are, but …

Analysis of global ozone records revealed:

- Severe, obvious problems with ~1 % of the data
- Hundreds of unflagged, rather unlikely outliers, episodes with constant values, etc.
- Inconsistencies between different parts of the timeseries
- Inconsistencies among data in different archives
- Unexplicable small-scale regional differences
- Numerous metadata inconsistencies

Metadata Quality Control Example

AQS site 06-045-0009 (Willits, CA)
“LOCATED IN BACK OF SAFEWAY COMPLEX.”



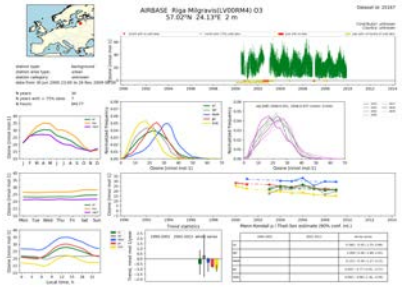
lat: 39.403056
lon: -123.349167
reported_alt: **1377**
google_alt: 417

lon and lat fit perfectly
→ status = 0 (verified)

station altitude
erroneously listed in ft.
→ alt_flag = 1 (google)

What has been done so far?

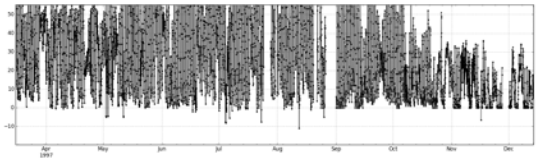
Visual inspection of data
(standardized data summary plots)

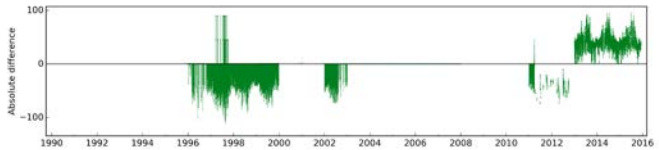


Ranking of data to identify
extreme issues

	A	B	C	D	E
1	Year	max. O3	Station ID	Station name	Station country
2	1980	490	06-037-0016	Glendora	United States of America
3	1981	360	06-071-0003	695 E. 3RD ST., SAN B	United States of America
4	1981	360	06-071-9003	172 W 3RD ST, SAN B	United States of America
5	1981	360	06-071-1004	Upland	United States of America
6	1982	390	06-037-1601		United States of America
7	1983	390	06-037-0002	Azusa	United States of America
8	1984	340	06-037-0016	Glendora	United States of America
9	1984	340	06-071-0005	Crestline	United States of America
10	1985	390	06-037-0016	Glendora	United States of America
11	1986	441	MX_PED	Pedregal	Mexico
12	1987	350	MX_MER	Merced	Mexico

Detailed investigation of specific
data issues and contact to data
providers





Visual inspection of station
locations and internet searches

Data flagging

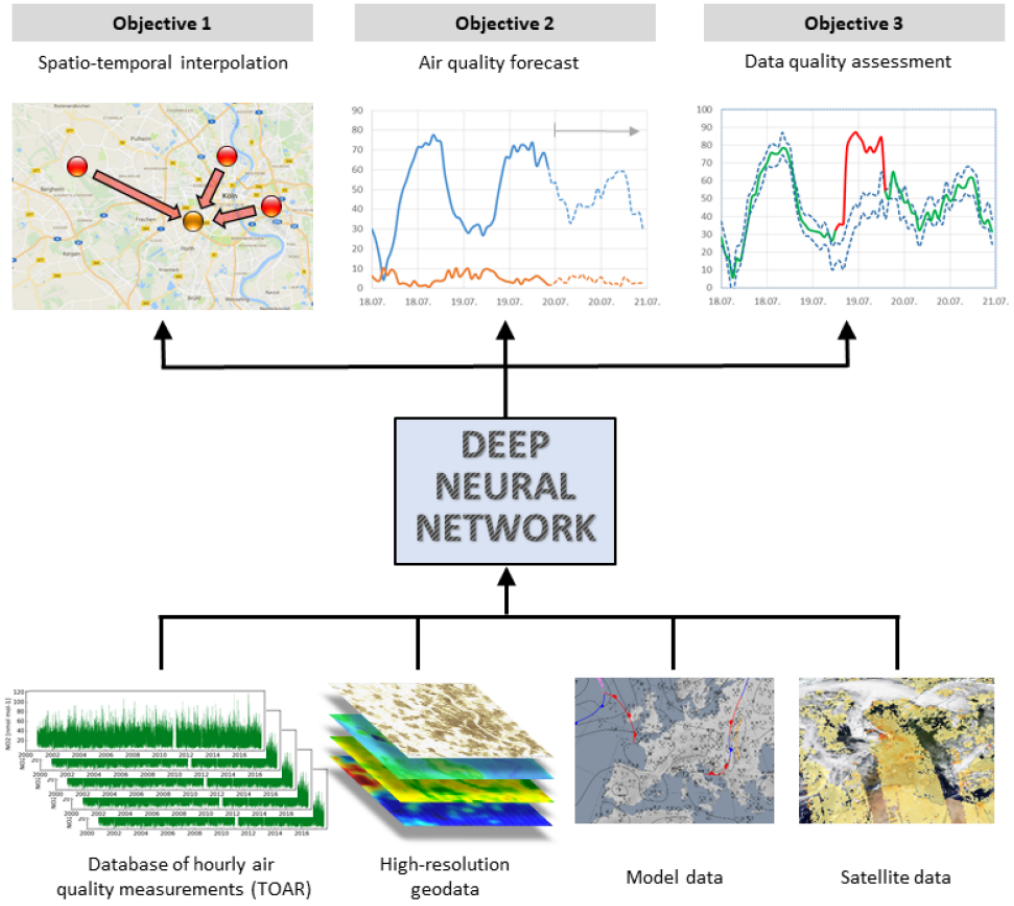
Adopted WMO code table 033 020:

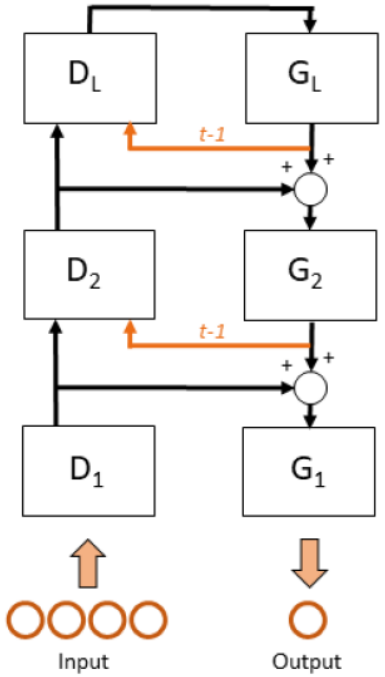
Flag value	Meaning
0	OK
1	inconsistent
2	doubtful
3	erroneous
4	not checked
5	changed
6	estimated
7	missing value

- Mapped „local“ flags to common scheme
- Only use 3 flag values in practice
- Subjective decision

Just beginning: use of KI to improve data coverage and quality control

TOAR tropospheric ozone assessment report





European Research Council
Established by the European Commission

Advanced Grant
ERC-2017-ADG
#787576

Figure 1: Schematic diagram of the IntelliAQ objectives and general concept. The project will build an unprecedented database of information related to air quality and will use these data in cutting-edge deep neural networks to enable the interpolation and prediction of air quality and to improve the quality assessment of global air pollution measurements.

TOAR database roadmap

TOAR
tropospheric
ozone
assessment
report

European Research Council
Established by the European Commission

Advanced Grant
ERC-2017-ADG
#787576

- Interface with openaq to obtain near realtime data globally
- Enhance database model to capture more metadata
- Implement automatic quality control procedures
- Introduce data quality scores
- Document all data modifications

Automatic data quality procedures

European Research Council
Established by the European Commission

Advanced Grant
ERC-2017-ADG
#787576

Python framework for test workflows

Group 1: single value tests (range, outlier)

Group 2: neigbourhood tests (constant, step change, variance)

Group 3: consistency tests (freq. distributions, diurnal/seasonal cycles, …)

Group 4: spatial similarity tests (nearby stations, station clusters)

Group 5: multi-species consistency tests (e.g. ozone titration)

Group 6: deep learning

CURRENT STATUS

QC-Tests

Group1

Sensor-gross range test: finds data points exceeds the prescribed minimum and maximum value.

User-gross range test: finds data points exceeds the user prescribed minimum and maximum value.

Range test: identifies the data points falls within the expectation range. The range is determined by the mean and standard deviation of the prior data.

Group2

Rate of change test: checks the rate of change in the variance structure in a given period.

Flat line test: identifies where several data stuck at the a singular value.

QC-Test Settings for Ozone (experimental)

Sensor-gross range test: $-5 \leqslant validvalue \leq 500$

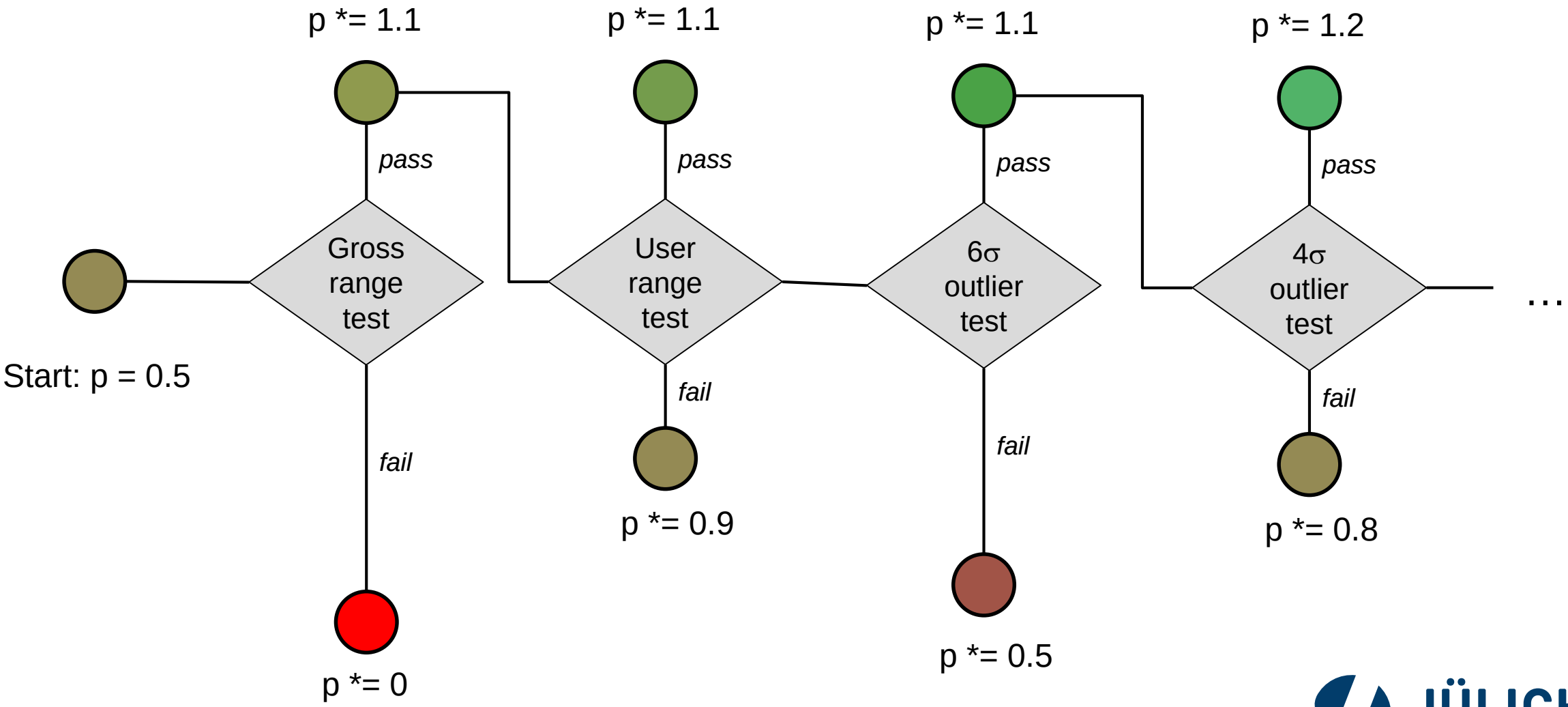
User-gross range test: $0 \leqslant validvalue \leq 120$

Range test: $\mu_{5day} - n\sigma_{5days} \leqslant validvalue \leqslant \mu_{5day} + n\sigma_{5days}$

Rate of change test: $vat_t - val_{t-1} \leqslant n\sigma_{5days}$

Flat line test: $val_t - val_{t-n} \geqslant 0.01$

Assign a probability for „valid data“

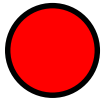


Flagging: map final probability to flag values

(also allow user probability thresholds for data selection)

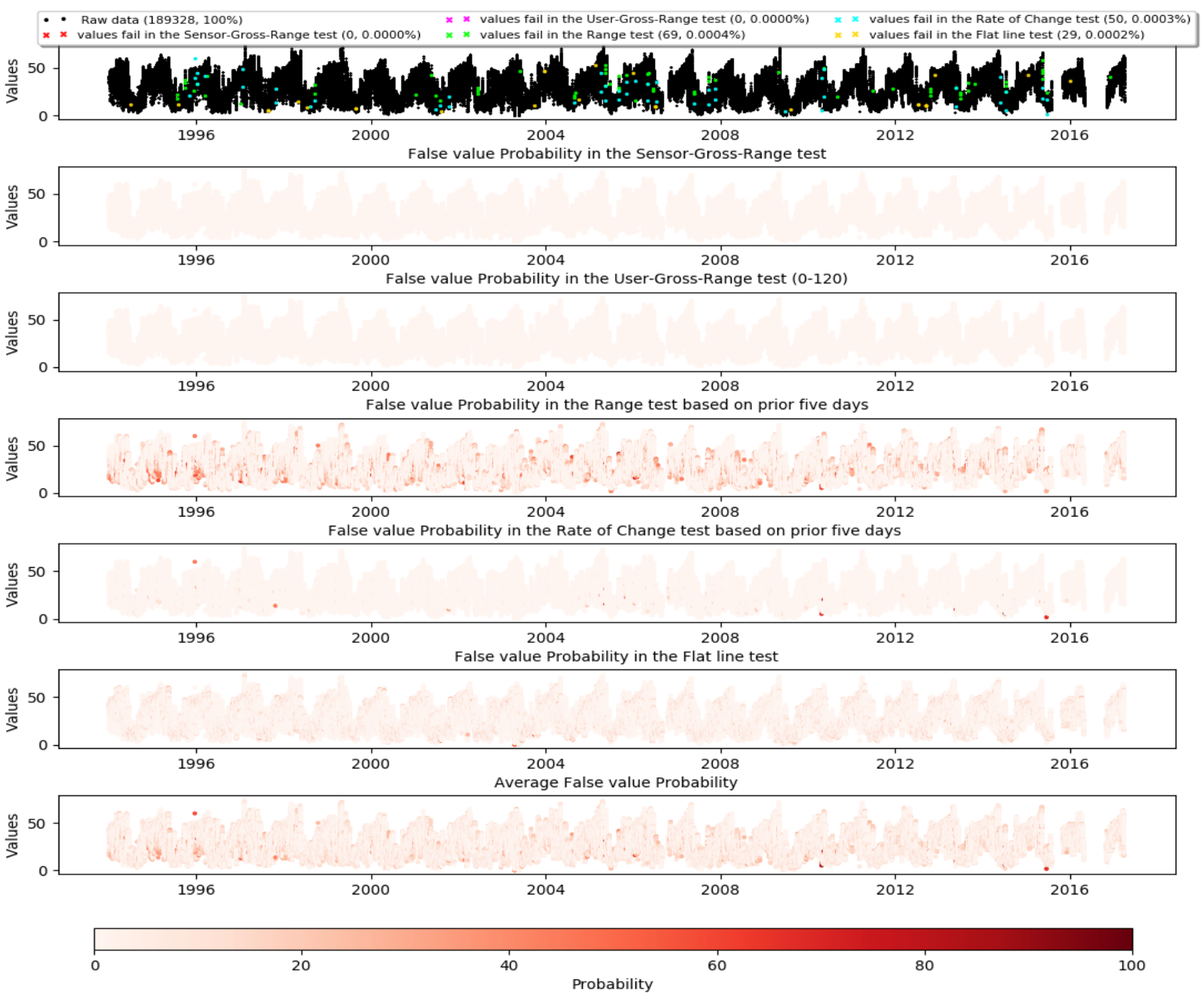
$p \geq 0.8$: valid value (flag = 0)

$0.8 > p \geq 0.3$: suspect or high interest value (flag = 3)

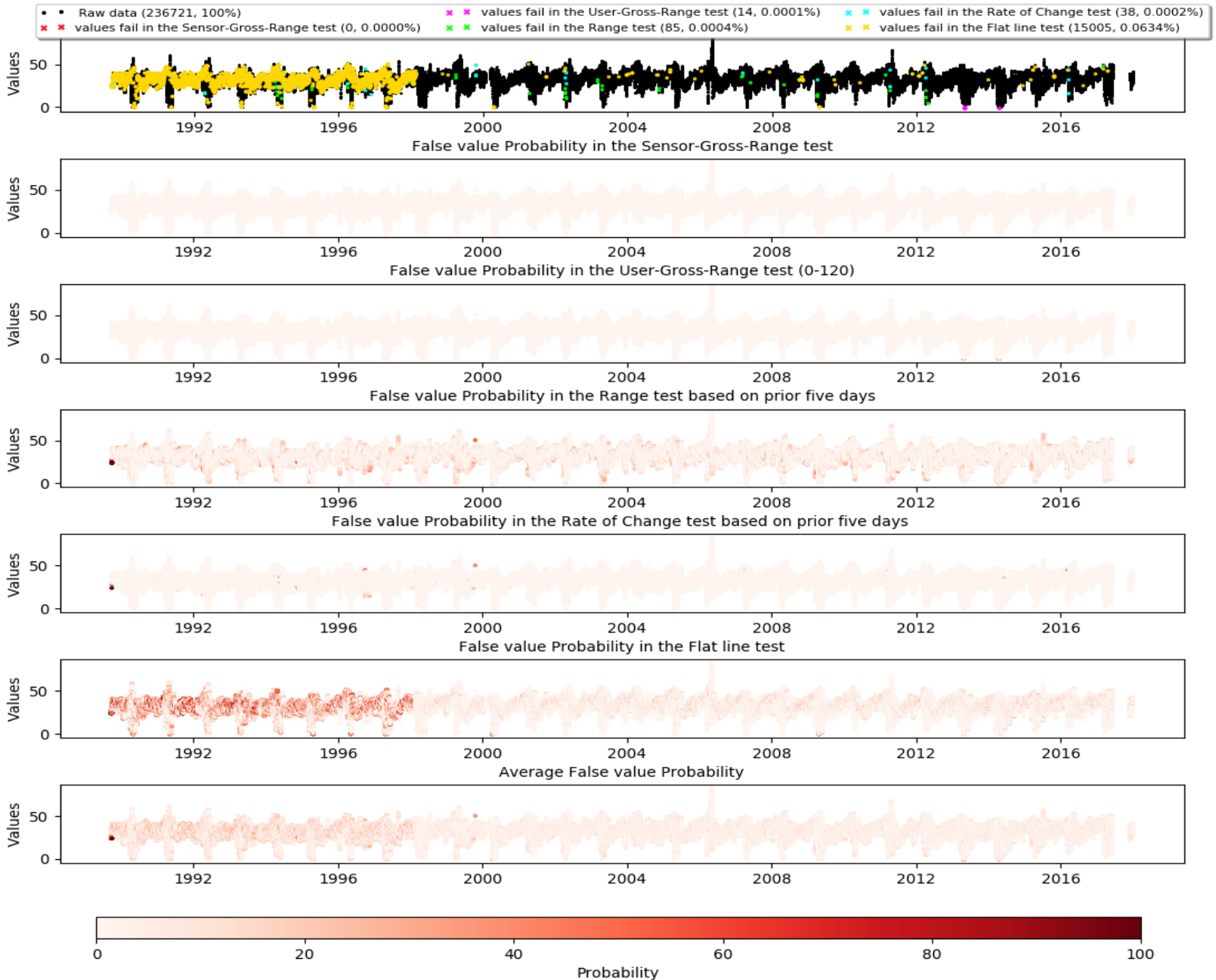


$0.3 > p$: erroneous value (flag = 4)

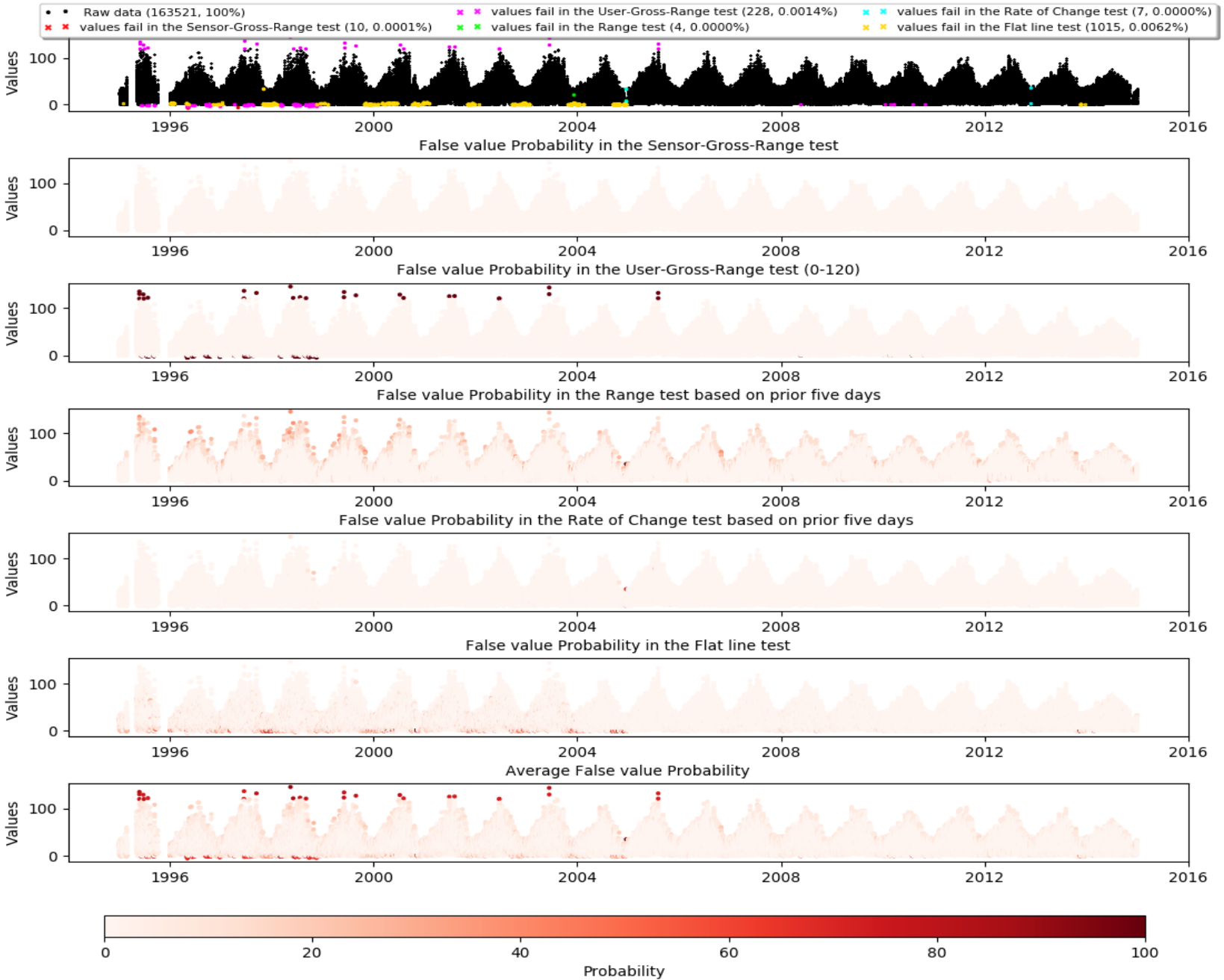
Preliminary results



O3, Zeppelin mountain (Ny-Ålesund)



O3, Montelibretti



QC scores (5-star concept)

Completeness of metadata

Length of timeseries

Data completeness

Data consistency

Absence of obvious artefacts (outliers, etc.)

Thanks for listening!