11th Seminar for Homogenization and Quality Control in
Climatological Databases, Budapest 2023

AQUAS – Austria Quality Service

A data quality tool at GeoSphere Austria

© M. Kopecky

Data Quality and Digitization
Niko Filipović*, Anita Paul, Alexandra Fritz, Martin Auer
niko.filipovic@geosphere.at

10. May 2023

Outline

- Overview: Data Quality Management and AQUAS
- Examples of quality control of
  - wind speed data (10 min) (real-time)
  - Global radiation and sunshine duration data (daily) (offline)
- Outlook

GeoSphere Austria – observational network

GeoSphere Austria

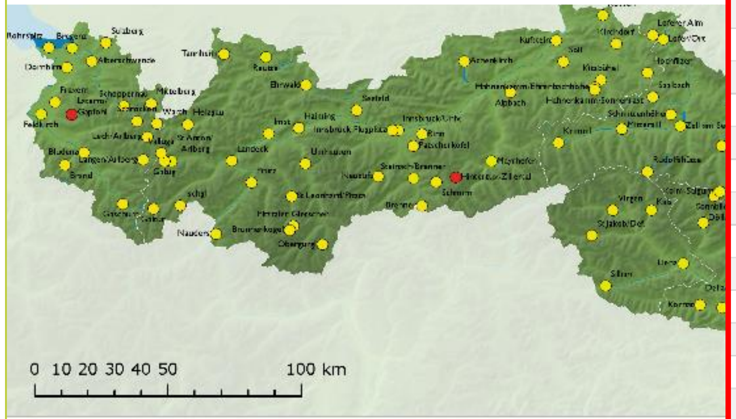
~ 280 stations

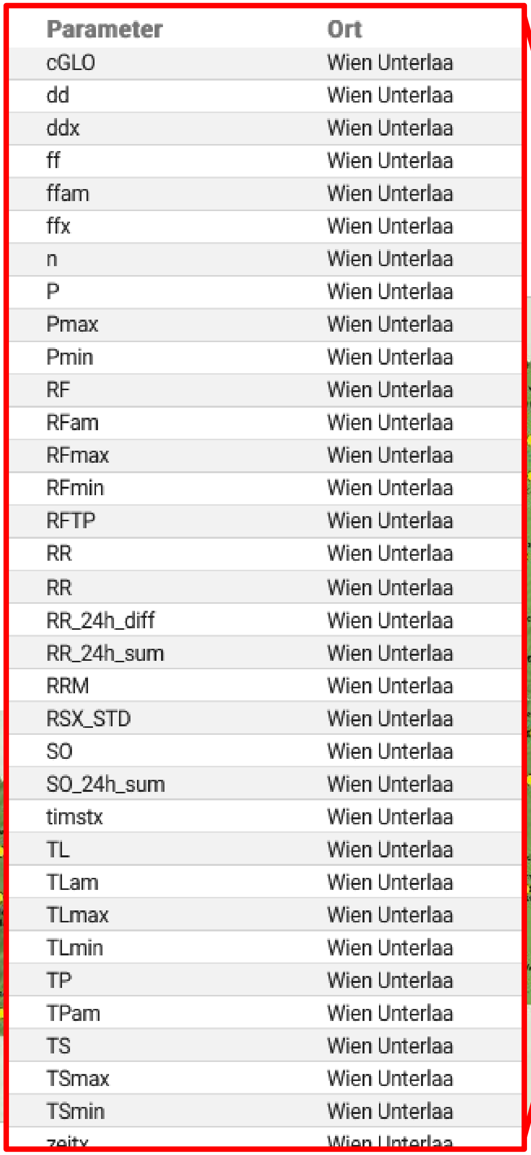
- ● 206 semi-automated weather stations – TAWES
- ● 60 full-automated weather stations – VAMES (including aviation-meteorologically important sensors for visibility, weather phenomena and cloud conditions)
- ●● 12 third-party network stations

- ○ global radiation
- ○ wind (speed, gust, direction)
- ○ air pressure
- ○ rel. Humidity
- ○ precipitation (amount + monitor)
- ○ sunshine duration
- ○ Temperature (2m; + 5cm; soil (3 levels))
- ○ dew point temperature

Challenging operational QC due to non-uniform instrumentation across the network:

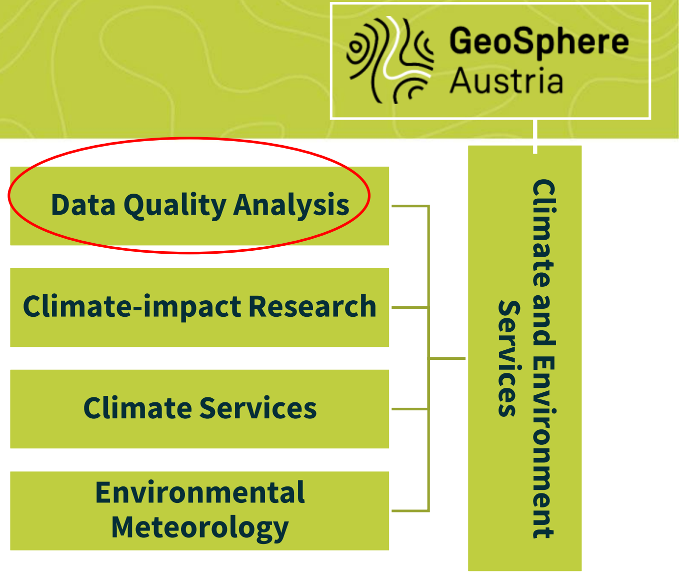
- tipping bucket and weighing rain gauges
- sonic and cup anemometer
- different types of humidity sensors
- high-end and low-cost sensors (third-party network)
- different time resolution (1-, 10-, 15-min)





Parameter	Ort
cGLO	Wien Unterlaa
dd	Wien Unterlaa
ddx	Wien Unterlaa
ff	Wien Unterlaa
ffam	Wien Unterlaa
ffx	Wien Unterlaa
n	Wien Unterlaa
P	Wien Unterlaa
Pmax	Wien Unterlaa
Pmin	Wien Unterlaa
RF	Wien Unterlaa
RFam	Wien Unterlaa
RFmax	Wien Unterlaa
RFmin	Wien Unterlaa
RFTP	Wien Unterlaa
RR	Wien Unterlaa
RR	Wien Unterlaa
RR_24h_diff	Wien Unterlaa
RR_24h_sum	Wien Unterlaa
RRM	Wien Unterlaa
RSX_STD	Wien Unterlaa
SO	Wien Unterlaa
SO_24h_sum	Wien Unterlaa
timstx	Wien Unterlaa
TL	Wien Unterlaa
TLam	Wien Unterlaa
TLmax	Wien Unterlaa
TLmin	Wien Unterlaa
TP	Wien Unterlaa
TPam	Wien Unterlaa
TS	Wien Unterlaa
TSmax	Wien Unterlaa
TSmin	Wien Unterlaa
zeitx	Wien Unterlaa

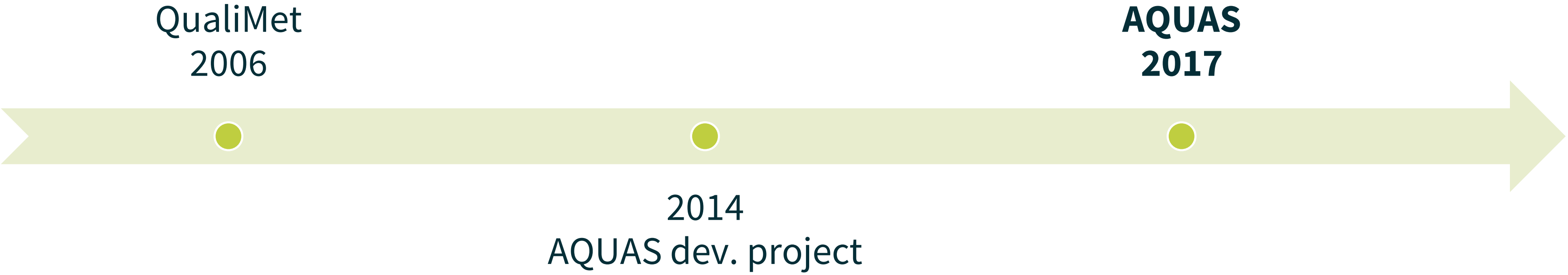
Department Data Quality Analysis



Our tasks within the scope of the Data Quality Management:

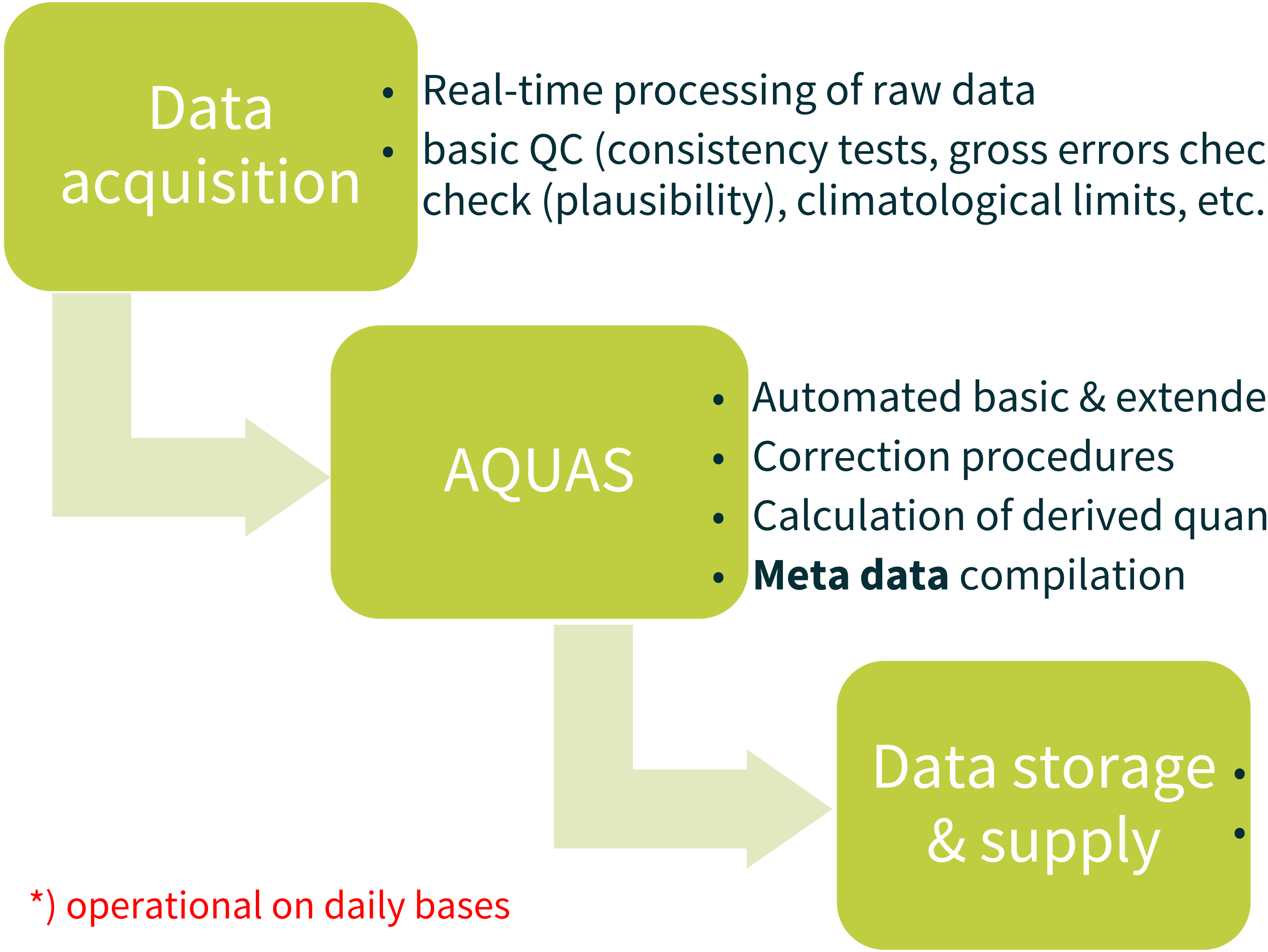
- Quality control
- Quality assurance
- Development and maintenance of test procedures
- Storage of raw and quality-checked data
- Documentation of all data modifications (**metadata**)

AQUAS



Data Quality Management at GeoSphere Austria

Comprehensive system for quality control and quality assurance



Data acquisition

- Real-time processing of raw data
- basic QC (consistency tests, gross errors check, range check (plausibility), climatological limits, etc.)

AQUAS

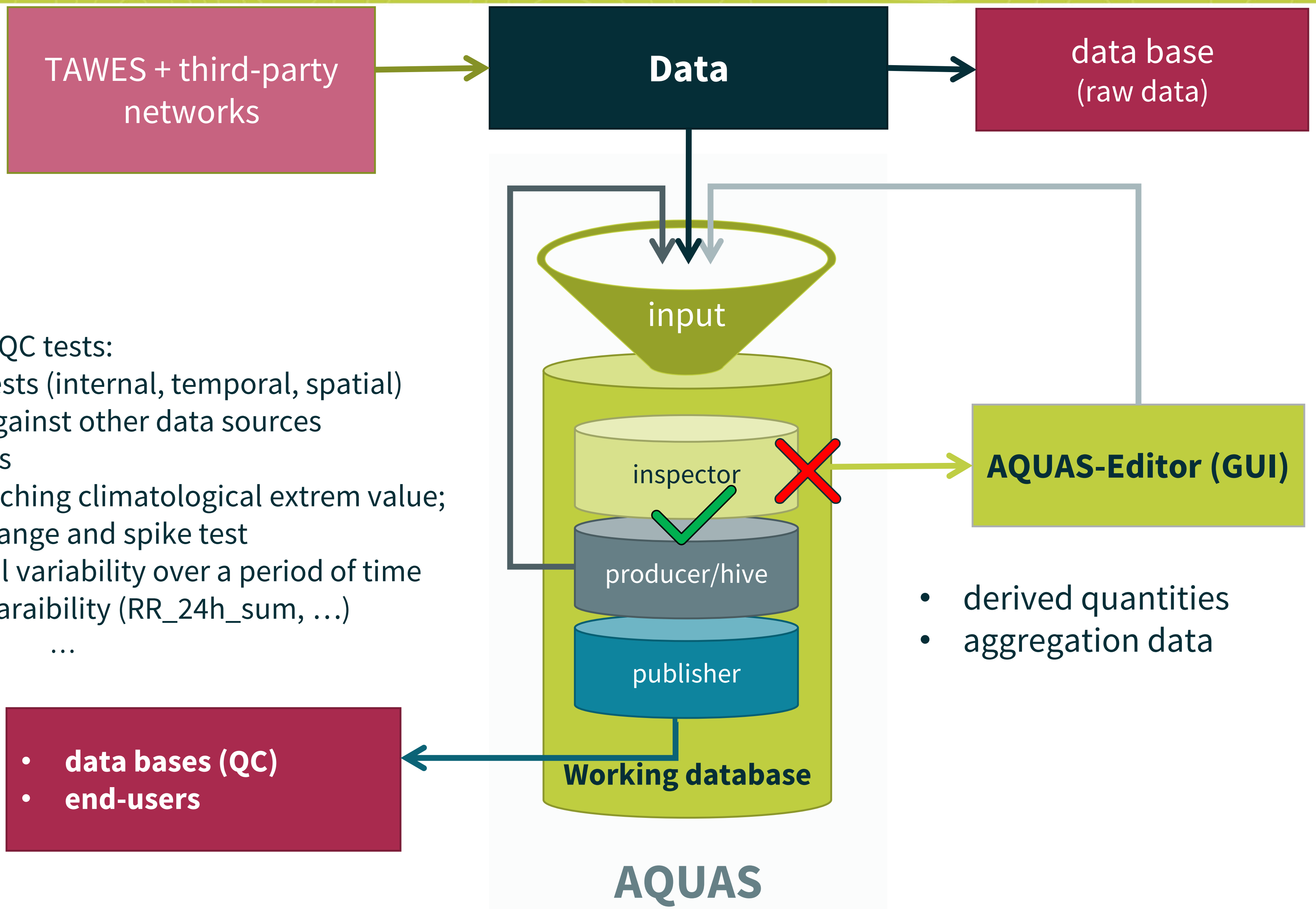
- Automated basic & extended QC of processed data **in near real-time***
- Correction procedures
- Calculation of derived quantities (hourly, daily data)
- **Meta data** compilation

Data storage & supply

- Storage of raw and processed data
- Data distribution to users

*) operational on daily bases

AQUAS – system structure



TAWES + third-party networks

Data

data base (raw data)

input

inspector

producer/hive

publisher

Working database

AQUAS

AQUAS-Editor (GUI)

- **data bases (QC)**
- **end-users**

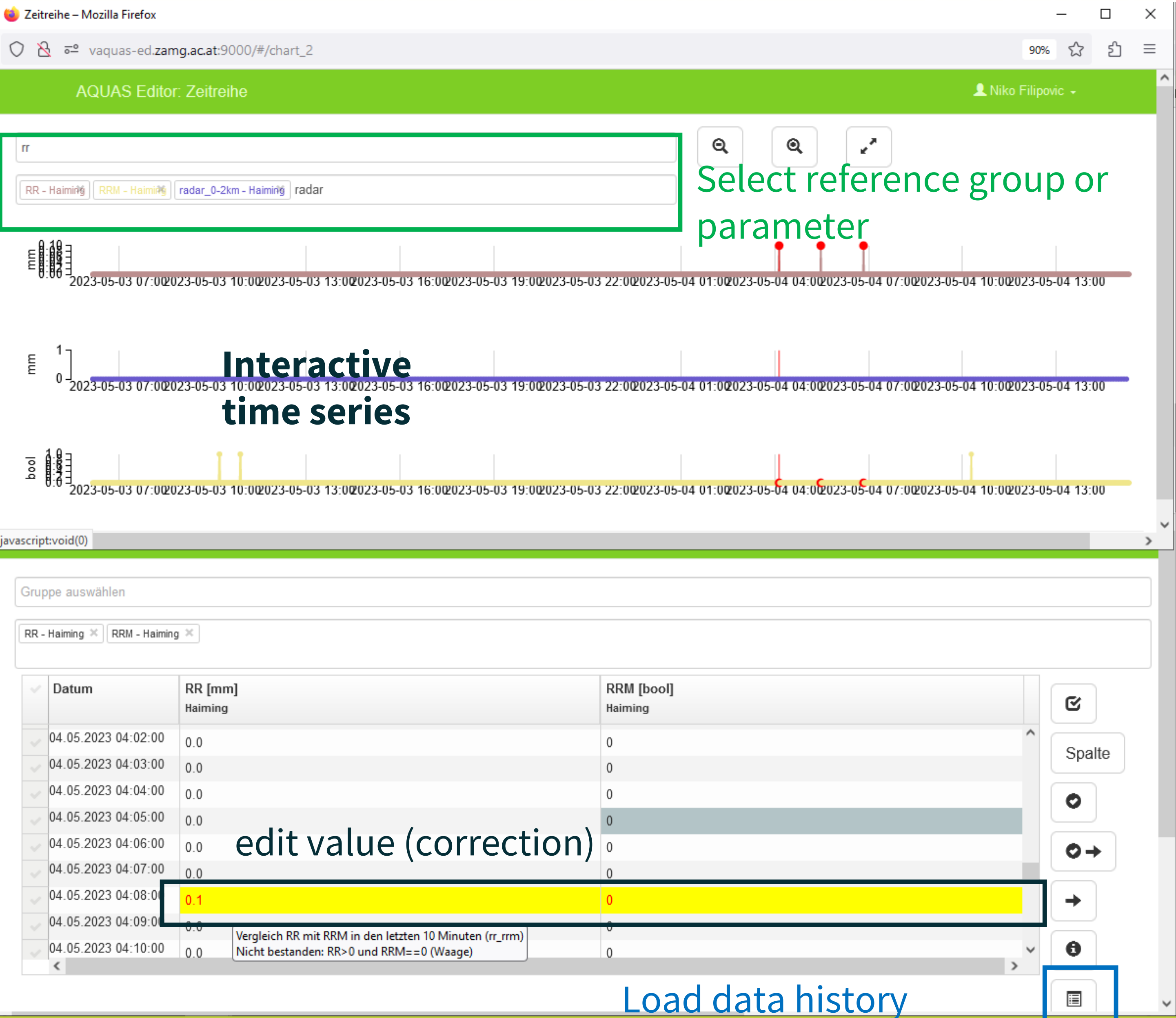
Basic & extended QC tests:

- Consistency tests (internal, temporal, spatial)
- Cross-check against other data sources
- Statistical tests
  - T approaching climatological extrem value;
  - Rapid change and spike test
  - Temporal variability over a period of time
  - Spatial varaibility (RR_24h_sum, …)

…

- derived quantities
- aggregation data

AQUAS – Web editor



Select reference group or parameter

Interactive time series

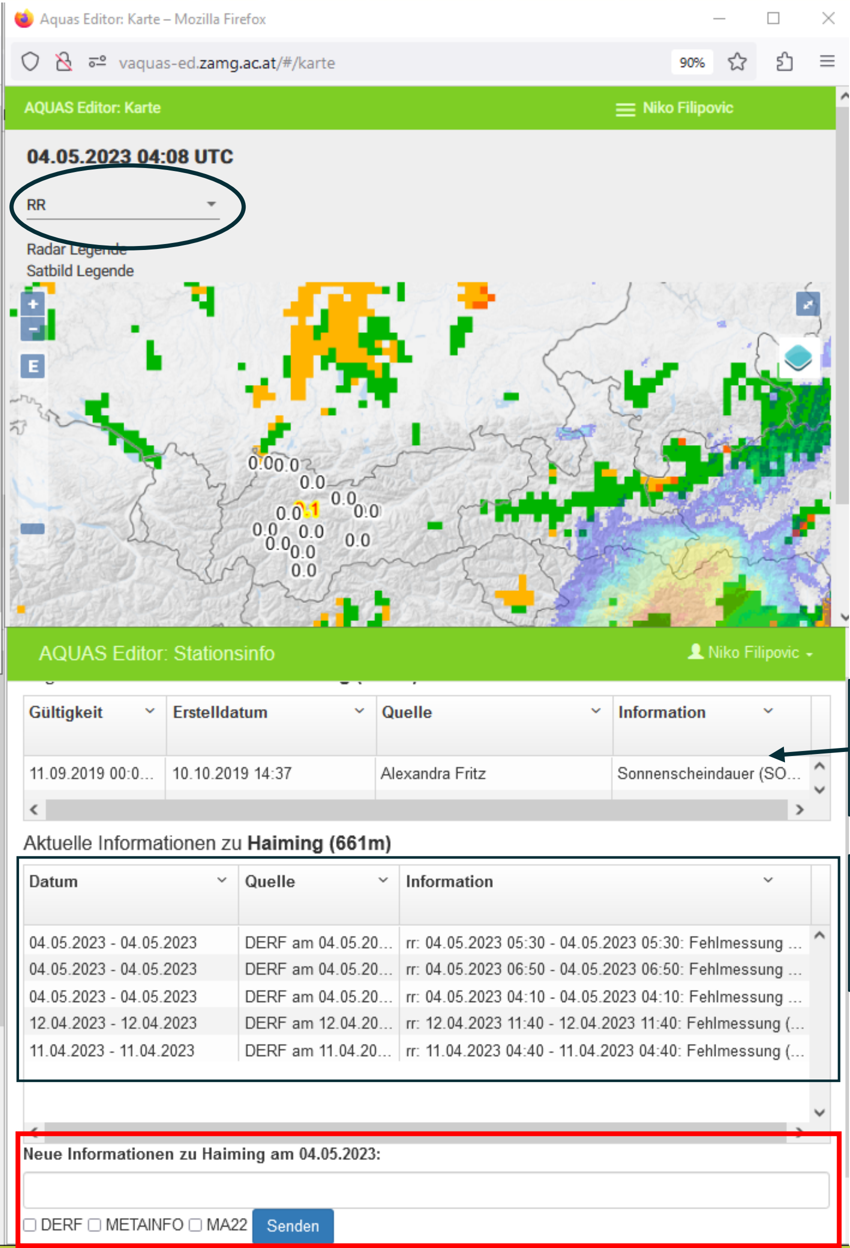
RR - Haiming ✕ RRM - Haiming ✕

Datum	RR [mm] Haiming	RRM [bool] Haiming
04.05.2023 04:02:00	0.0	0
04.05.2023 04:03:00	0.0	0
04.05.2023 04:04:00	0.0	0
04.05.2023 04:05:00	0.0	0
04.05.2023 04:06:00	0.0	0
04.05.2023 04:07:00	0.0	0
04.05.2023 04:08:00	0.1	0
04.05.2023 04:09:00	0.0	0
04.05.2023 04:10:00	0.0	0

Vergleich RR mit RRM in den letzten 10 Minuten (rr_rrm)
Nicht bestanden: RR>0 und RRM==0 (Waage)

edit value (correction)

Load data history


Aquas Editor: Karte – Mozilla Firefox
vaquas-ed.zamg.ac.at/#/karte
90%
AQUAS Editor: Karte
Niko Filipovic
RR
Radar Legende
Satbild Legende
E
AQUAS Editor: Stationsinfo
Niko Filipovic

04.05.2023 04:08 UTC

Gültigkeit	Erstelldatum	Quelle	Information
11.09.2019 00:0...	10.10.2019 14:37	Alexandra Fritz	Sonnenscheindauer (SO...

Aktuelle Informationen zu **Haiming (661m)**

Datum	Quelle	Information
04.05.2023 - 04.05.2023	DERF am 04.05.20...	rr: 04.05.2023 05:30 - 04.05.2023 05:30: Fehlmessung ...
04.05.2023 - 04.05.2023	DERF am 04.05.20...	rr: 04.05.2023 06:50 - 04.05.2023 06:50: Fehlmessung ...
04.05.2023 - 04.05.2023	DERF am 04.05.20...	rr: 04.05.2023 04:10 - 04.05.2023 04:10: Fehlmessung ...
12.04.2023 - 12.04.2023	DERF am 12.04.20...	rr: 12.04.2023 11:40 - 12.04.2023 11:40: Fehlmessung (...
11.04.2023 - 11.04.2023	DERF am 11.04.20...	rr: 11.04.2023 04:40 - 11.04.2023 04:40: Fehlmessung (...

Neue Informationen zu Haiming am 04.05.2023:

☐ DERF ☐ METAINFO ☐ MA22 Senden

show other data sources (radar, satellite, hydro-data, …)

Permanent station info

Current info

Message to technical staff or metadata

AQUAS in action

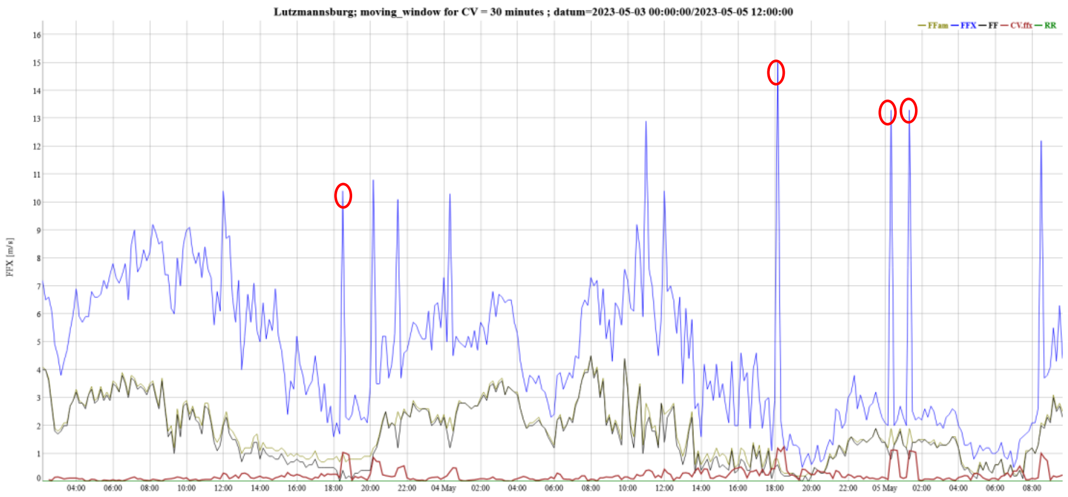
Examples

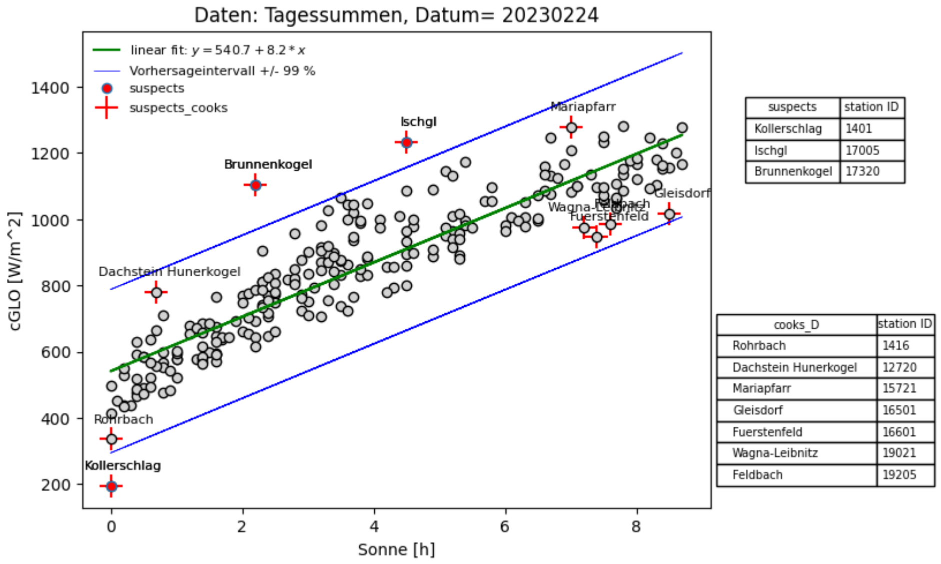
1. Real-time QC of 10-min wind speed data (spike-test)

2. Non-real-time QC of sunshine and global radiation daily data (linear regression)

Test-phase

Both check-routines run in a pre-operational mode

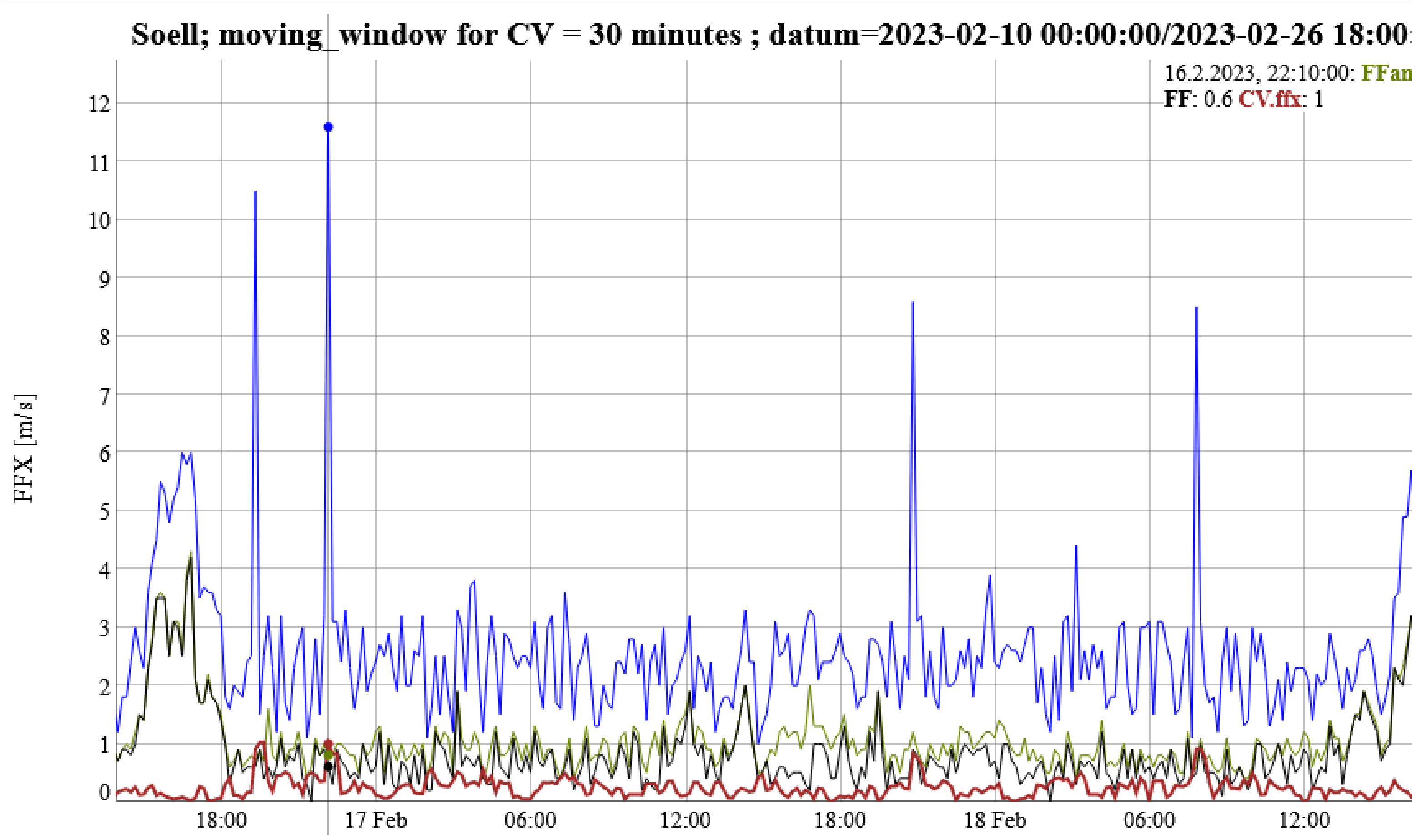

Lutzmannsburg; moving_window for CV = 30 minutes ; datum=2023-05-03 00:00:00/2023-05-05 12:00:00


Daten: Tagessummen, Datum= 20230224

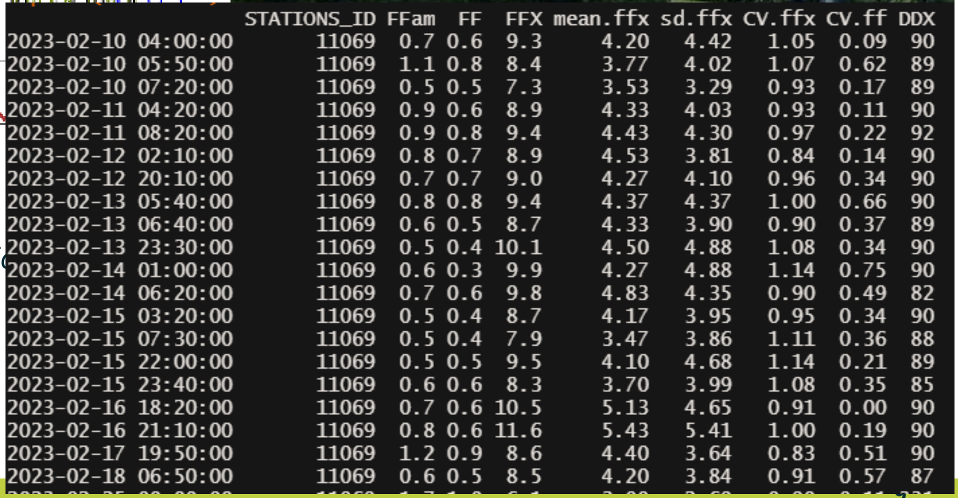
suspects	station ID
Kollerschlag	1401
Ischgl	17005
Brunnenkogel	17320

cooks_D	station ID
Rohrbach	1416
Dachstein Hunerkogel	12720
Mariapfarr	15721
Gleisdorf	16501
Fuerstenfeld	16601
Wagna-Leibnitz	19021
Feldbach	19205

Example 1: Maximum wind speed - spike test



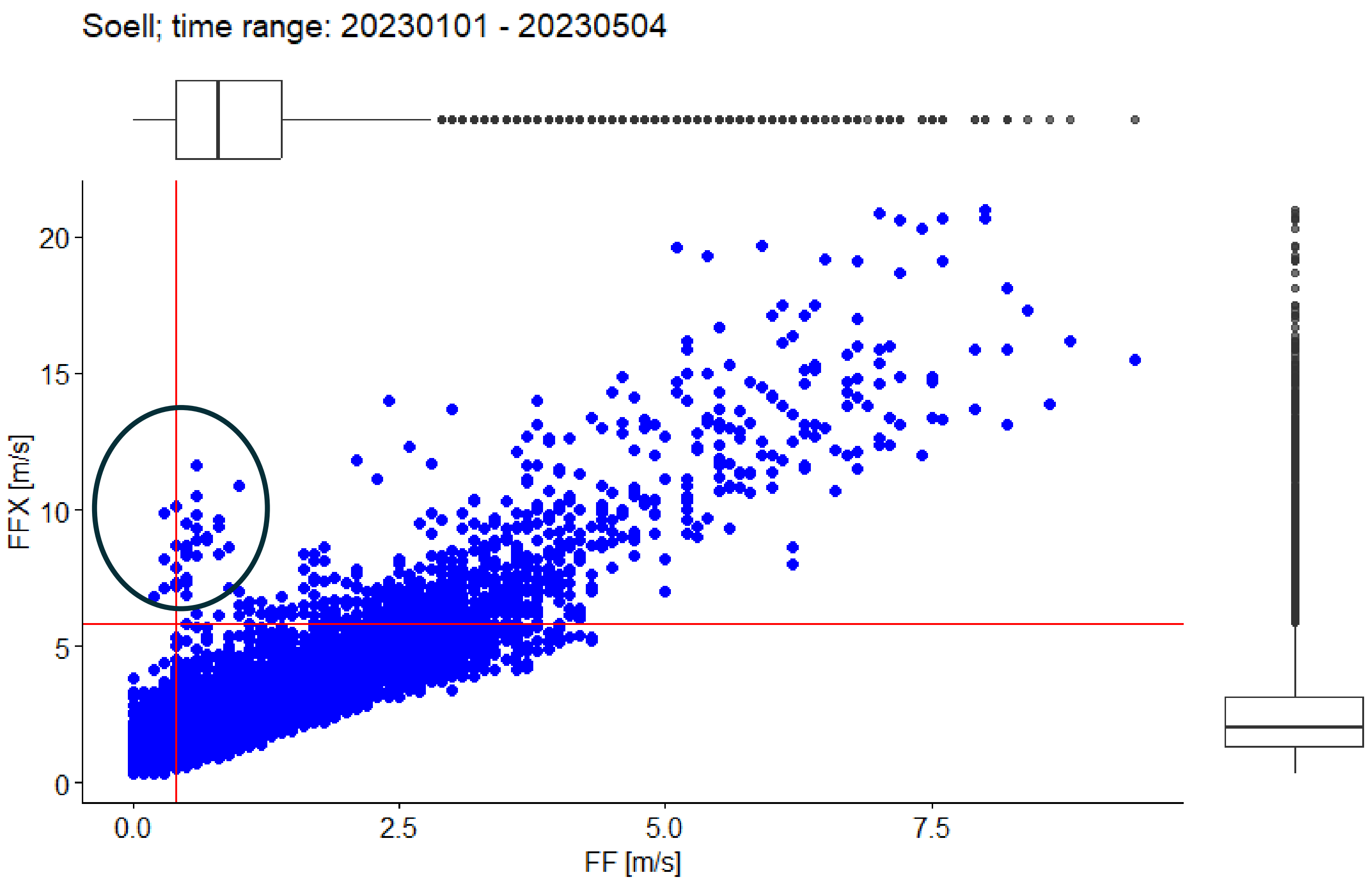
TAWES-weather stati

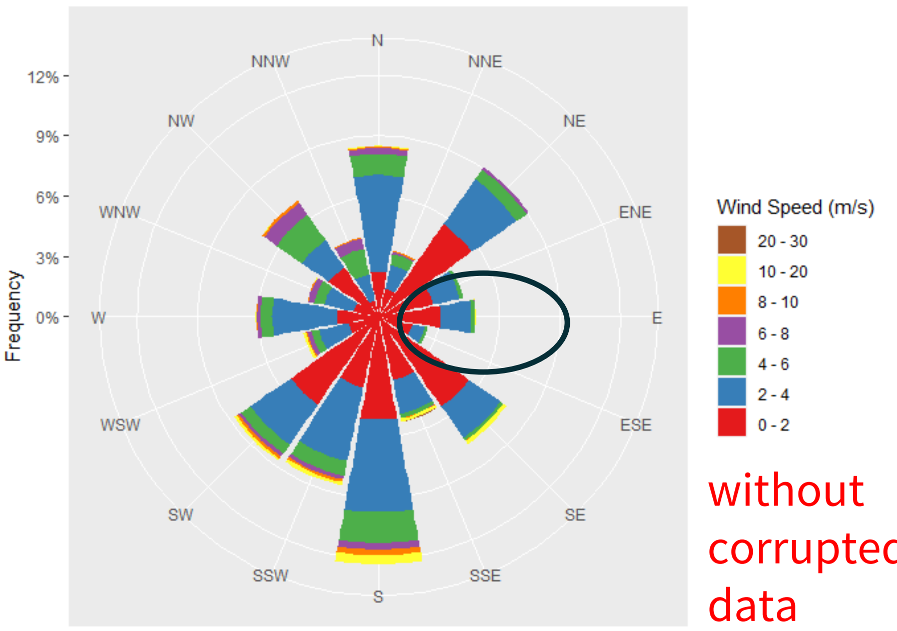


```
                    STATIONS_ID FFam  FF  FFX mean.ffx sd.ffx CV.ffx CV.ff DDX
2023-02-10 04:00:00       11069  0.7 0.6  9.3     4.20   4.42   1.05  0.09  90
2023-02-10 05:50:00       11069  1.1 0.8  8.4     3.77   4.02   1.07  0.62  89
2023-02-10 07:20:00       11069  0.5 0.5  7.3     3.53   3.29   0.93  0.17  89
2023-02-11 04:20:00       11069  0.9 0.6  8.9     4.33   4.03   0.93  0.11  90
2023-02-11 08:20:00       11069  0.9 0.8  9.4     4.43   4.30   0.97  0.22  92
2023-02-12 02:10:00       11069  0.8 0.7  8.9     4.53   3.81   0.84  0.14  90
2023-02-12 20:10:00       11069  0.7 0.7  9.0     4.27   4.10   0.96  0.34  90
2023-02-13 05:40:00       11069  0.8 0.8  9.4     4.37   4.37   1.00  0.66  90
2023-02-13 06:40:00       11069  0.6 0.5  8.7     4.33   3.90   0.90  0.37  89
2023-02-13 23:30:00       11069  0.5 0.4 10.1     4.50   4.88   1.08  0.34  90
2023-02-14 01:00:00       11069  0.6 0.3  9.9     4.27   4.88   1.14  0.75  90
2023-02-14 06:20:00       11069  0.7 0.6  9.8     4.83   4.35   0.90  0.49  82
2023-02-15 03:20:00       11069  0.5 0.4  8.7     4.17   3.95   0.95  0.34  90
2023-02-15 07:30:00       11069  0.5 0.4  7.9     3.47   3.86   1.11  0.36  88
2023-02-15 22:00:00       11069  0.5 0.5  9.5     4.10   4.68   1.14  0.21  89
2023-02-15 23:40:00       11069  0.6 0.6  8.3     3.70   3.99   1.08  0.35  85
2023-02-16 18:20:00       11069  0.7 0.6 10.5     5.13   4.65   0.91  0.00  90
2023-02-16 21:10:00       11069  0.8 0.6 11.6     5.43   5.41   1.00  0.19  90
2023-02-17 19:50:00       11069  1.2 0.9  8.6     4.40   3.64   0.83  0.51  90
2023-02-18 06:50:00       11069  0.6 0.5  8.5     4.20   3.84   0.91  0.57  87
```

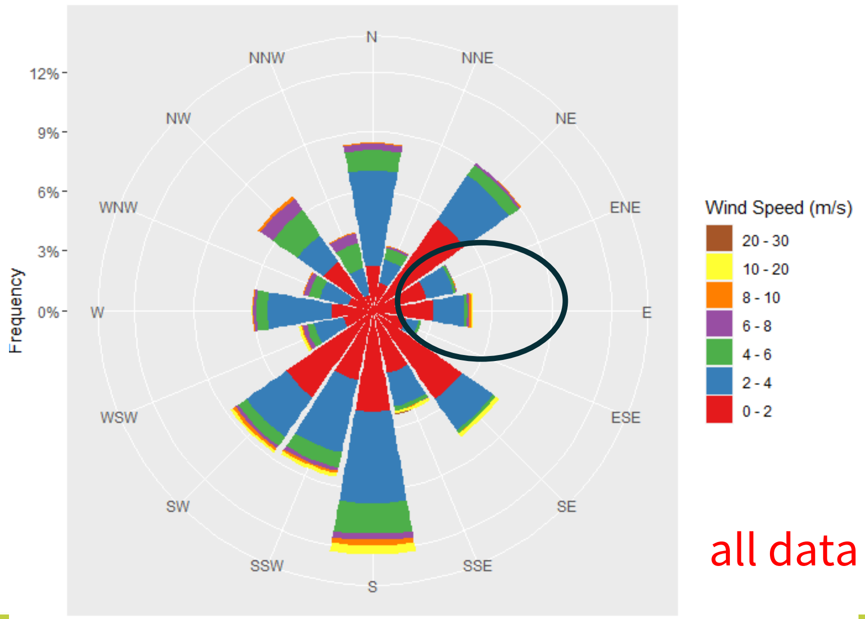
Dubious FFX-spikes occuring with $ddx \cong 90°$
(FFX … maximum wind speed during the last 10 minutes)

FFX – spike test





without corrupted data



all data

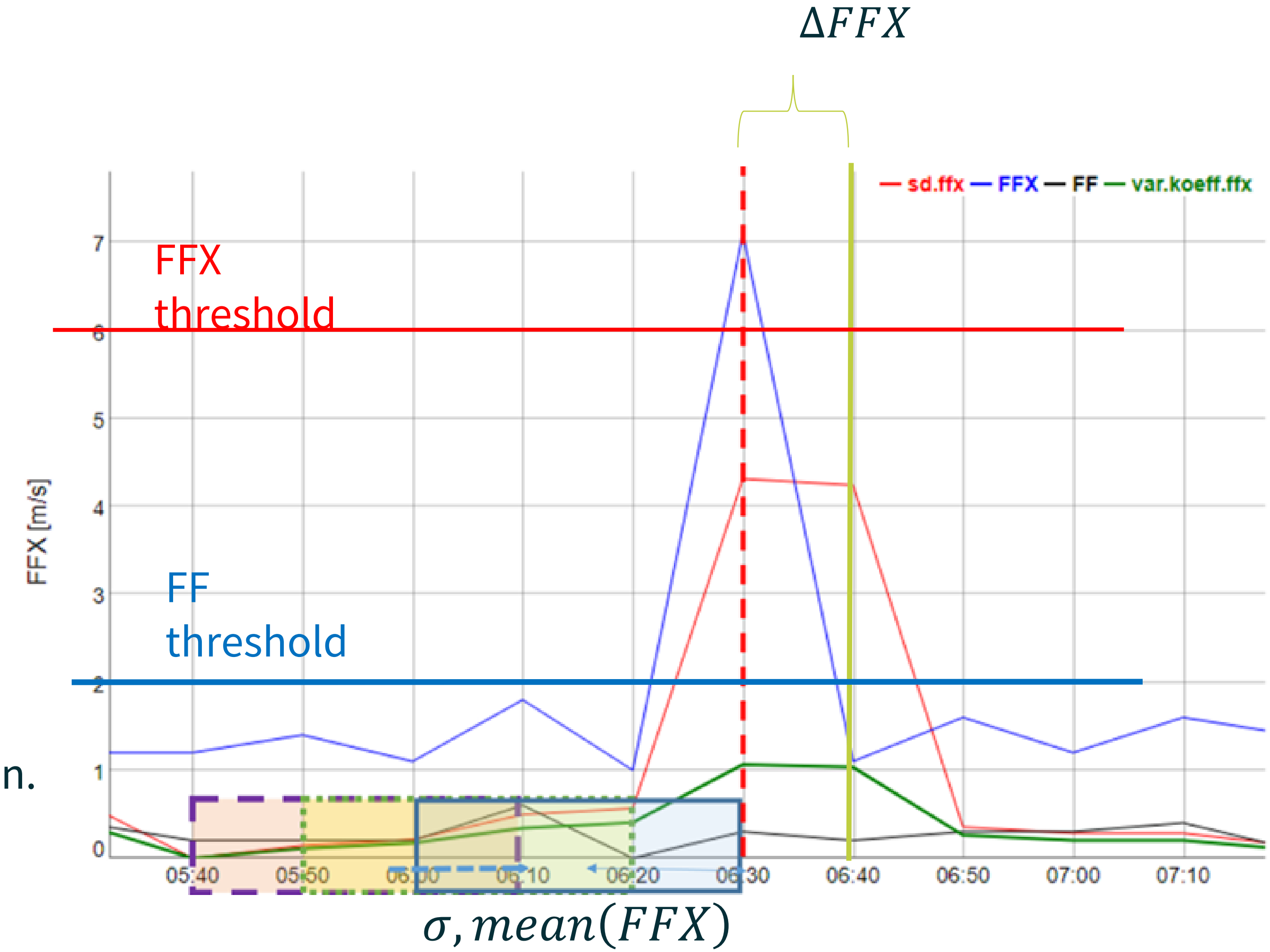
Example of a wind speed spike-check (test phase)

FFX – spike test

- *Calculate FFX-difference $\Delta\ FFX$ to previous time step*
- *If $\Delta\ FFX < 0$:*
  - *If $(FFX > threshold\ \ \&\ FF < threshold)^{*}$ then calculate the Coefficient of Variation (CV) from the last 30 min*

$$CV = \frac{\sigma_{FFX}}{mean(FFX)}$$

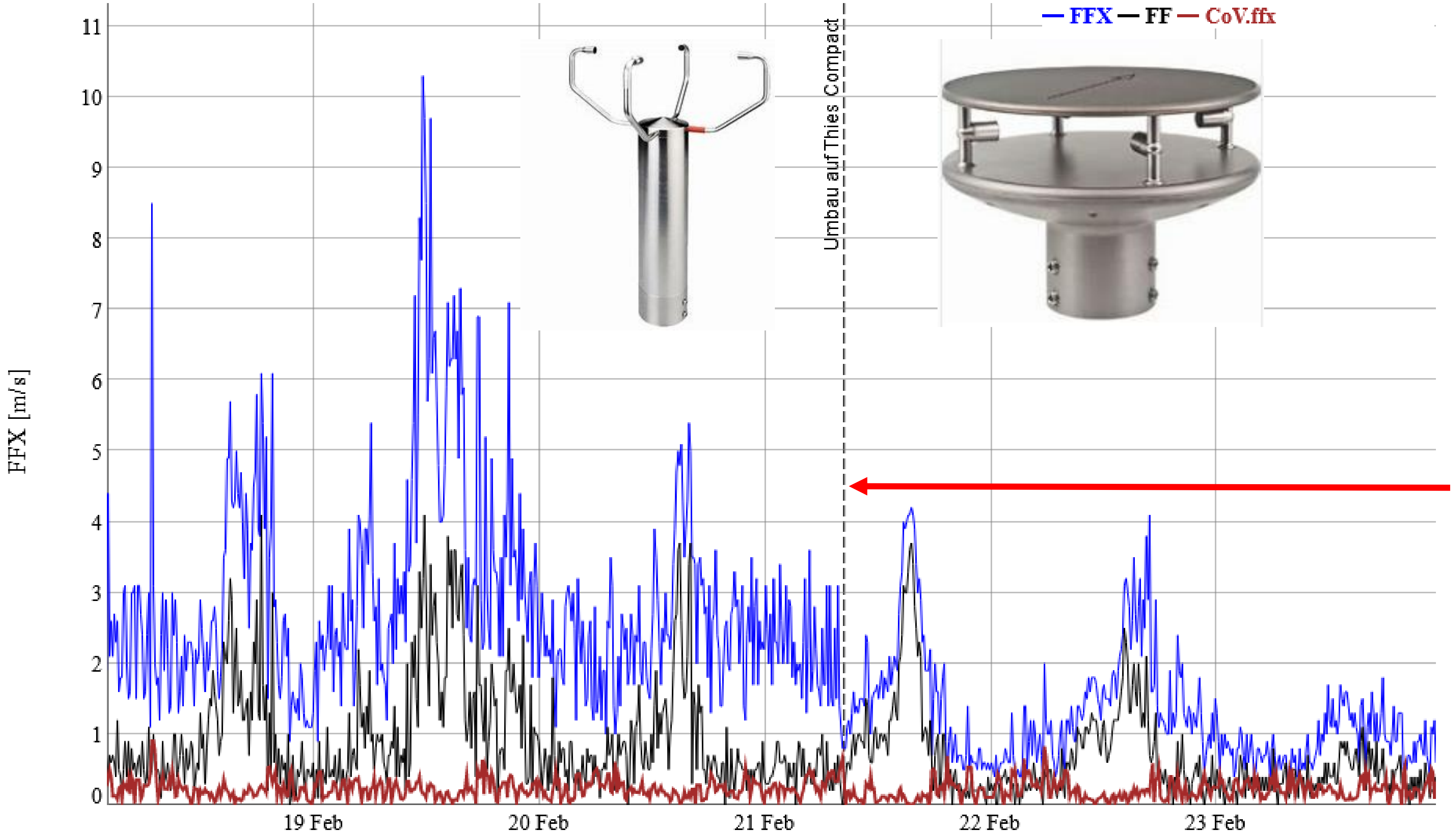
  - If CV exceeds a threshold (currently set by 0.7) the observation is flagged as suspect and a message appears in AQUAS web editor for manual inspection.



**) empirical thresholds from our data*

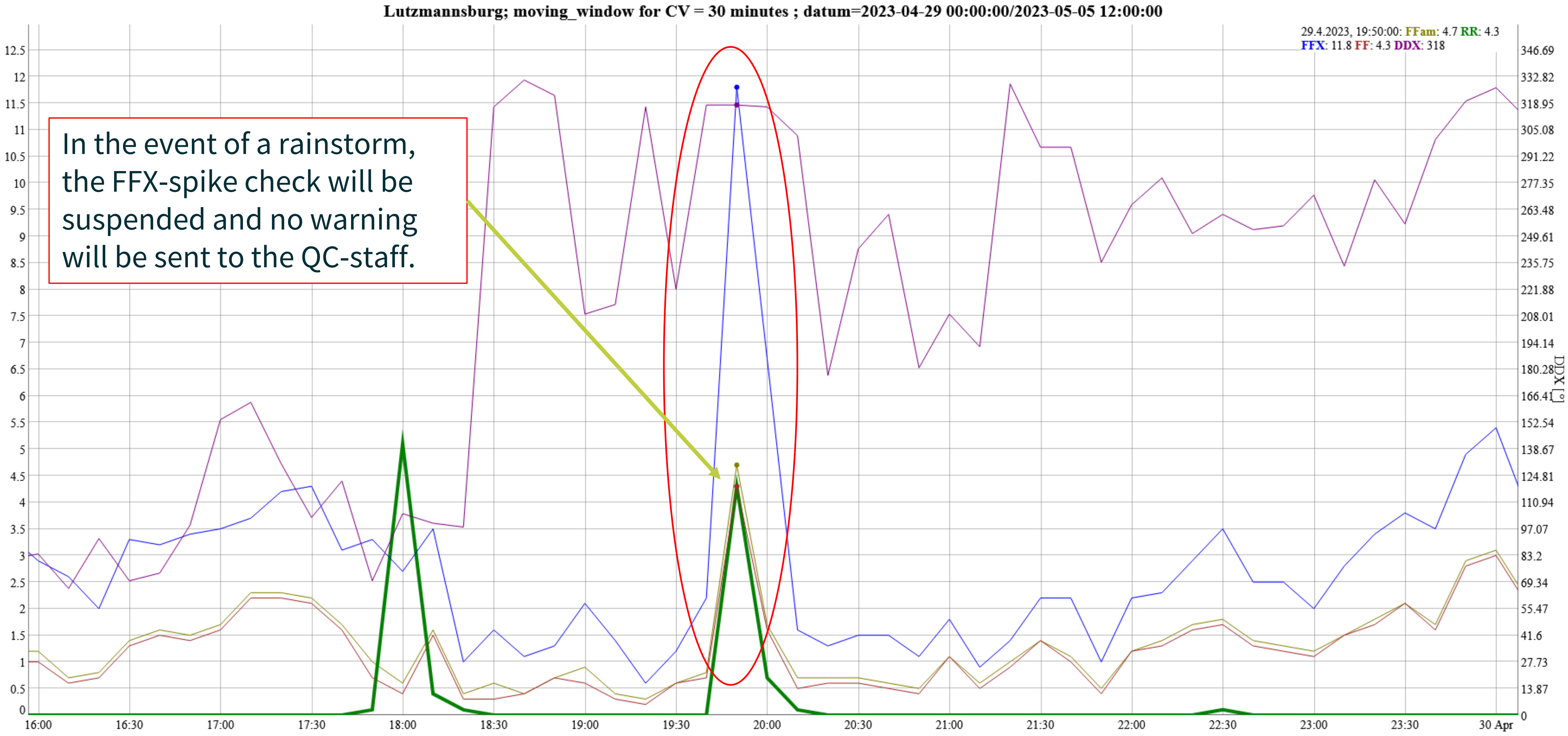
Wind sensor change

Soell; moving_window for CoV = 30 minutes ; datum=2023-01-01 06:00:00/2023-03-09 06:00:00



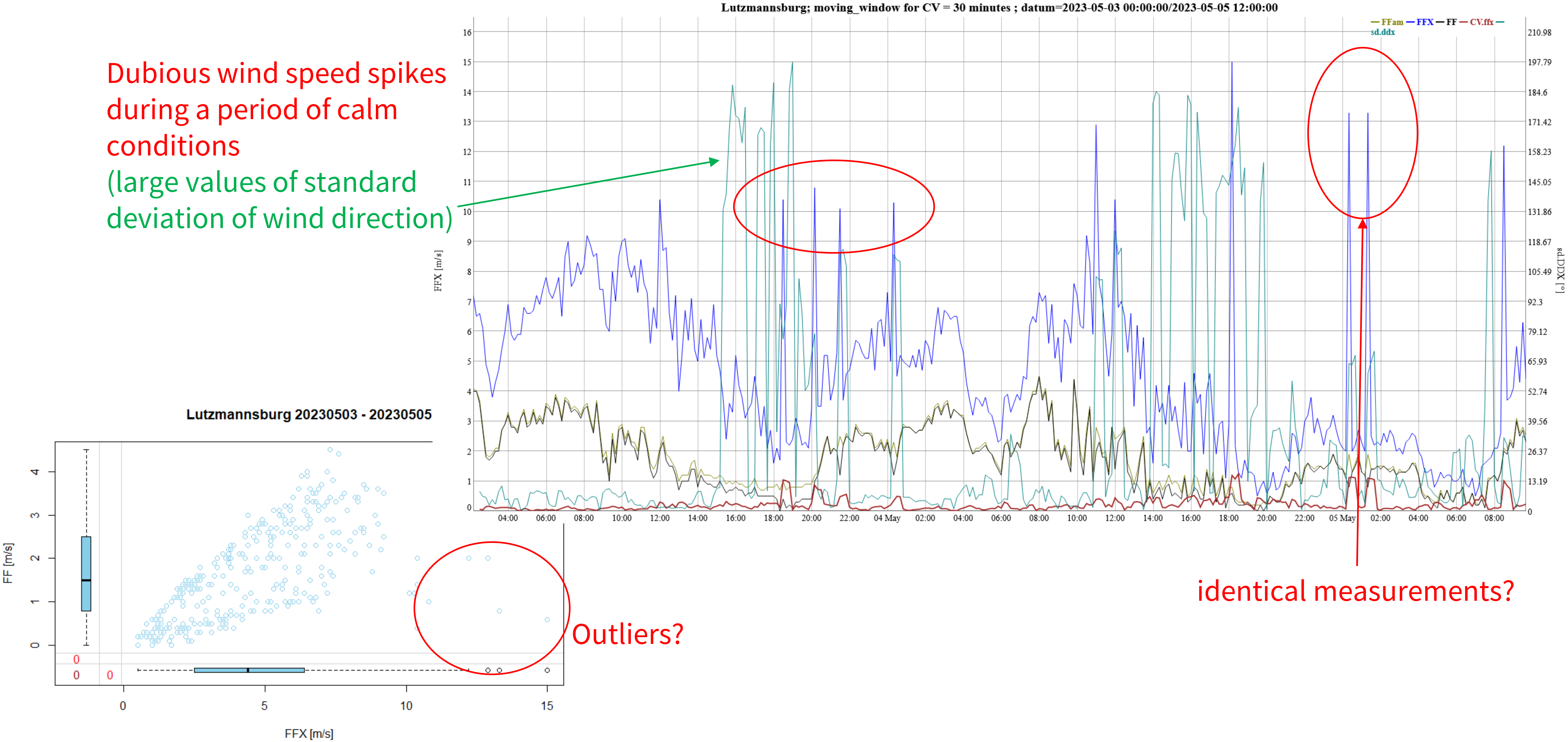
Time of the sensor change

FFX-spike with rain shower



In the event of a rainstorm, the FFX-spike check will be suspended and no warning will be sent to the QC-staff.

FFX-spike



Dubious wind speed spikes during a period of calm conditions

(large values of standard deviation of wind direction)

identical measurements?

Outliers?

Examples of wind checks in AQUAS

GeoSphere Austria

test		
Range of values	check if values are within a range manually check gusts > 30 m/s	$ff, ffam, ffx \in [0, 100]\ \frac{m}{s}$ $dd, ddx \in [0{,}360]°$
🛠 vectorial vs. scalar wind speed	comparison between vectorial and scalar mean of wind speed	FAILED if: ffam-ff <0 or ffam-ff > threshold
🛠 maximum (ffx) vs. mean (ff) wind speed	Check for gustiness intensity	FAILED: ff > ffx or FAILED: ffx > 19m/s and ffx/ff >= threshold
Sample rate wind speed	If N < 270 -> suspicious measurements	Time frame 30 minutes
Temporal variability	check if values are changing more or less than expected within a timeframe	Checks every 30 min
🛠 Wind speed spikes	step check for dubious wind speed spikes	check if coefficient of variation (moving window) exceeds a threshold
…	…	…

🛠 Work in progress

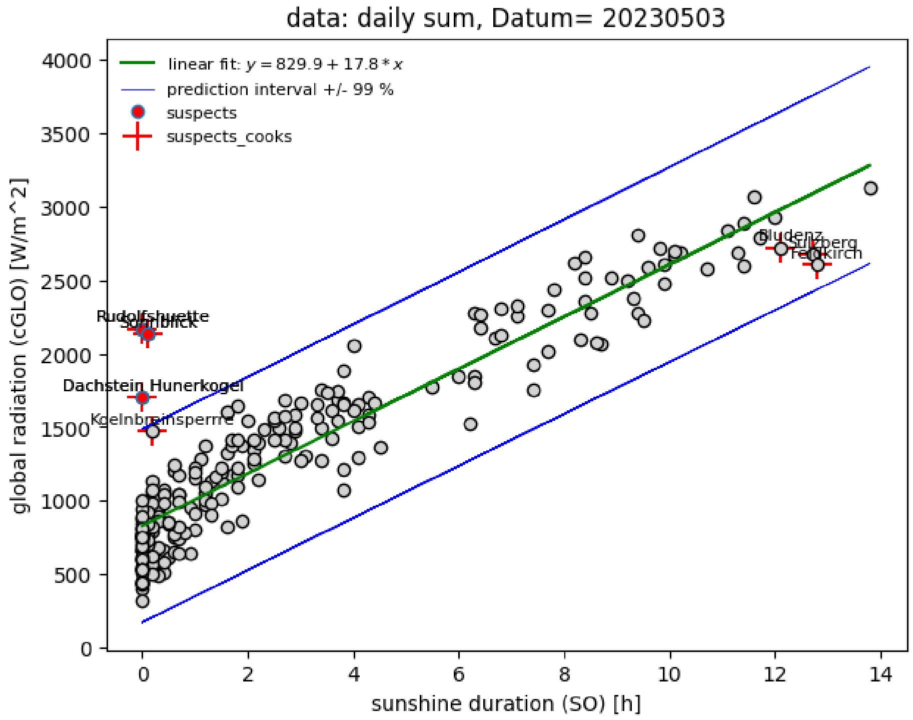
Example 2: sunshine and global radiation data

Method

At the end of each day a test runs loading daily sums of sunshine duration and global radiation data of the previous day.

Linear regression is calculated with lower and upper prediction intervalls using all available stations.

- ➢ Stations outside the 99%-predictions bounds are identified as suspect.
- ➢ Another subset of stations which are influential for linear regression regarding their Cook‘s distance (cooks_D) are flagged as „potentially suspect“



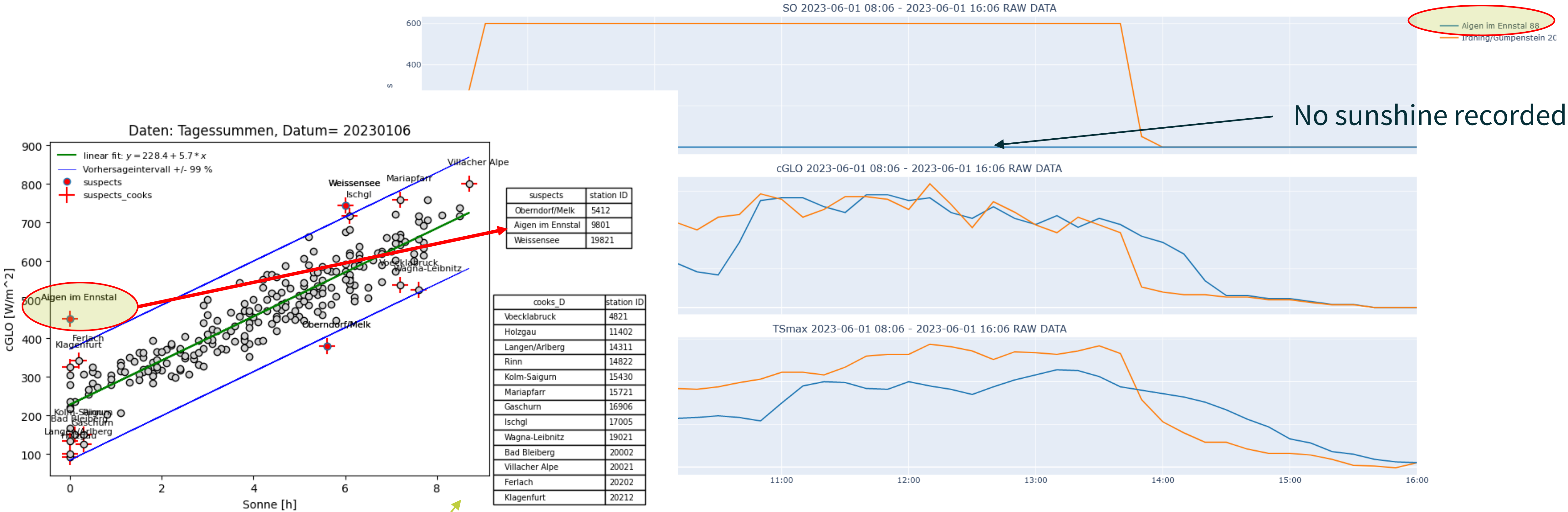
suspects	station ID
Dachstein Hunerkogel	12720
Rudolfshuette	15321
Sonnblick	15411

cooks_D	station ID
Sulzberg	8301
Feldkirch	11115
Bludenz	14103
Koelnbreinsperrre	15640

Cooks distance gives a comprehensive information about the change of a regression model after removing a particular observation.

Example: Non-real time check on the basis of daily data

Daily sum of
global radiation ~ sunshine duration



No sunshine recorded

suspects	station ID
Oberndorf/Melk	5412
Aigen im Ennstal	9801
Weissensee	19821

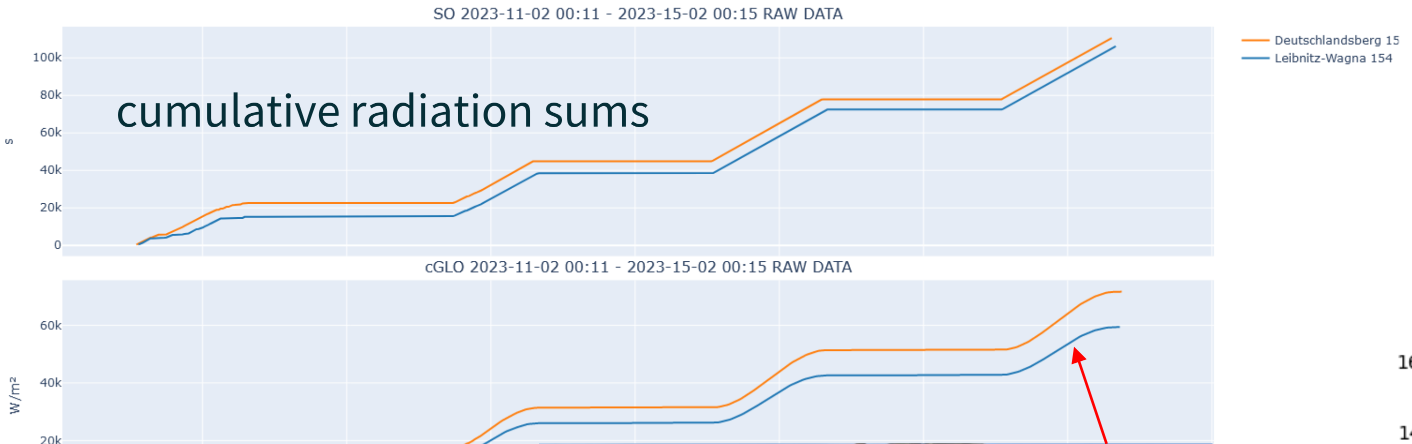
cooks_D	station ID
Voecklabruck	4821
Holzgau	11402
Langen/Arlberg	14311
Rinn	14822
Kolm-Saigurn	15430
Mariapfarr	15721
Gaschurn	16906
Ischgl	17005
Wagna-Leibnitz	19021
Bad Bleiberg	20002
Villacher Alpe	20021
Ferlach	20202
Klagenfurt	20212

Problem: Sunshine vs. Global radiation warning was erroneously confirmed as valid by the staff.

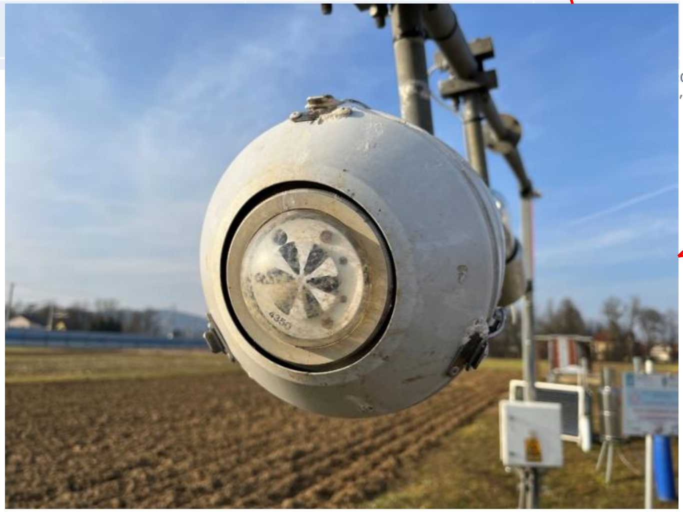
Potentially influential stations:
Cook‘s distance criterium: $> 4 * mean(all\ Cook's\ distances)$

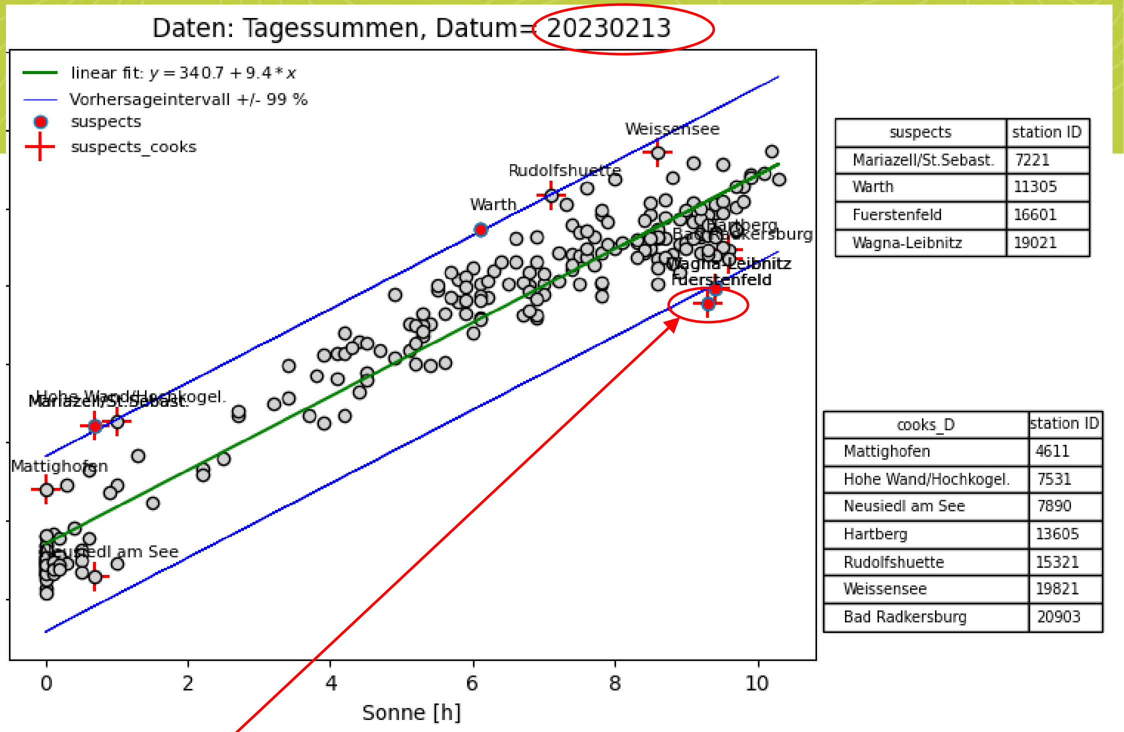
Global radiation vs. sunshine duration

Impact of a dirty sensor on the daily global radiation – gradually increasing error



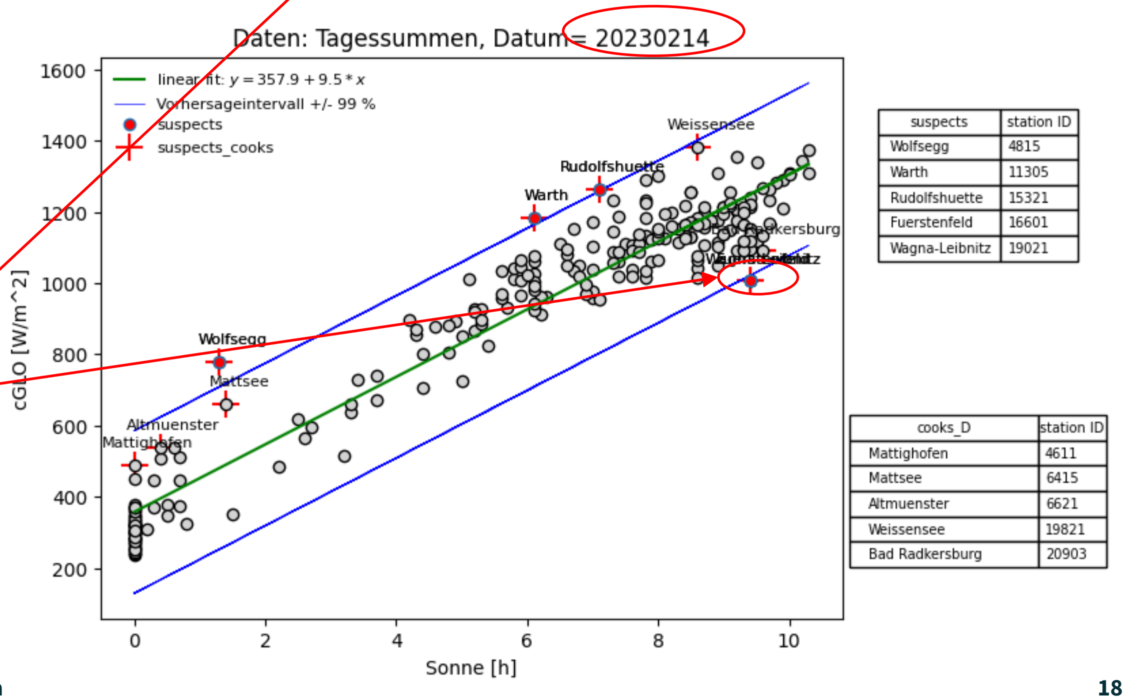
cumulative radiation sums





suspects	station ID
Mariazell/St.Sebast.	7221
Warth	11305
Fuerstenfeld	16601
Wagna-Leibnitz	19021

cooks_D	station ID
Mattighofen	4611
Hohe Wand/Hochkogel.	7531
Neusiedl am See	7890
Hartberg	13605
Rudolfshuette	15321
Weissensee	19821
Bad Radkersburg	20903

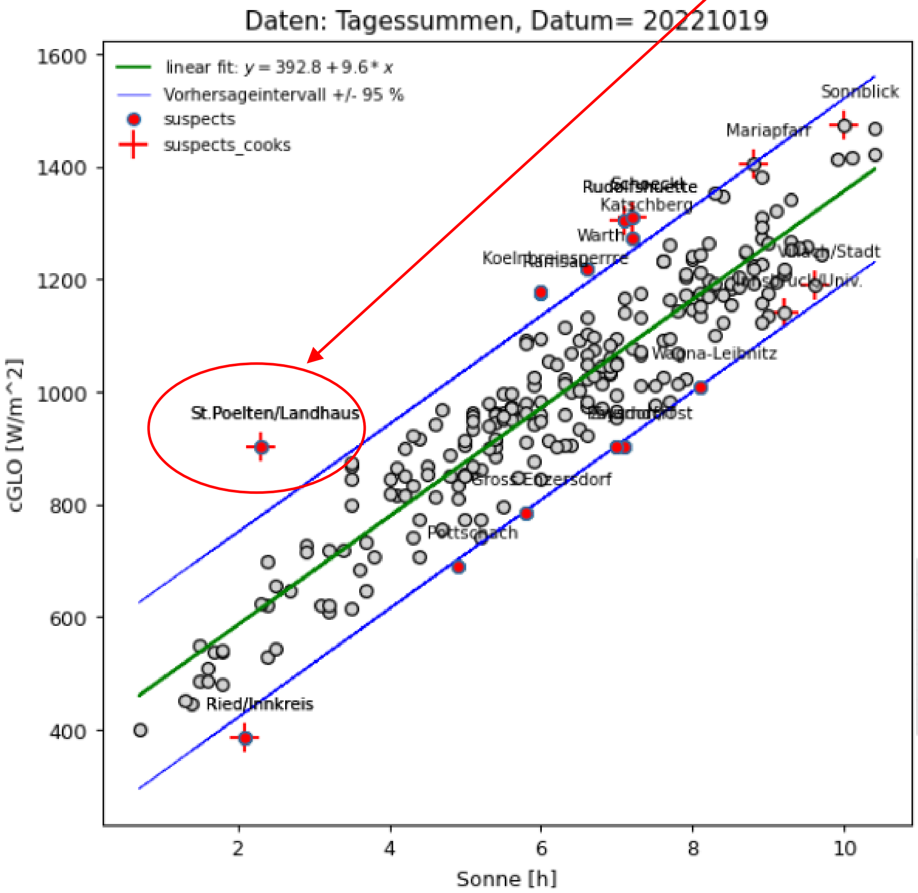


suspects	station ID
Wolfsegg	4815
Warth	11305
Rudolfshuette	15321
Fuerstenfeld	16601
Wagna-Leibnitz	19021

cooks_D	station ID
Mattighofen	4611
Mattsee	6415
Altmuenster	6621
Weissensee	19821
Bad Radkersburg	20903

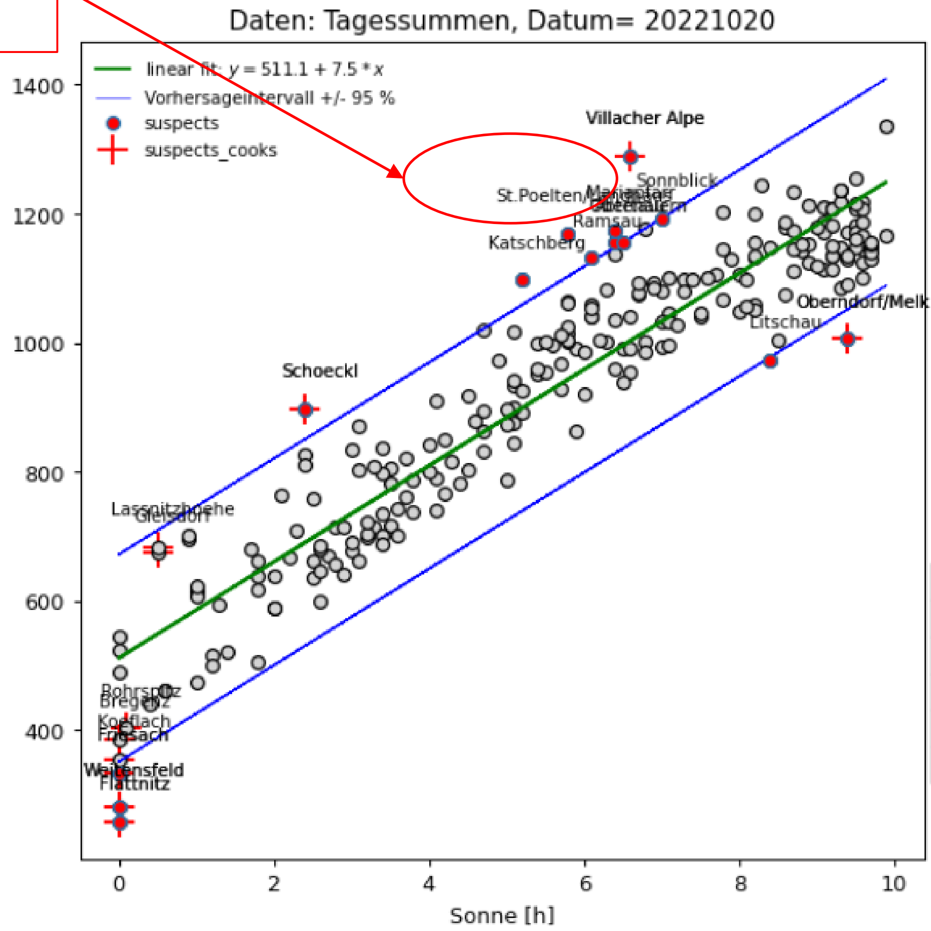
Example: impact of a levelling error on the sunshine duration

St. Pölten appears as suspect on two consecutive days.



suspects	station ID
Poysdorf/Ost	2503
Zwerndorf	4305
Ried/Innkreis	4705
St.Poelten/Landhaus	5609
Gross Enzersdorf	5972
Pottschach	10531
Warth	11305
Ramsau	12711
Rudolfshuette	15321
Koelnbreinsperrre	15640
Katschberg	15715
Schoeckl	16421
Wagna-Leibnitz	19021

cooks_D	station ID
Innsbruck/Univ.	11803
Sonnblick	15411
Mariapfarr	15721
Villach/Stadt	20123



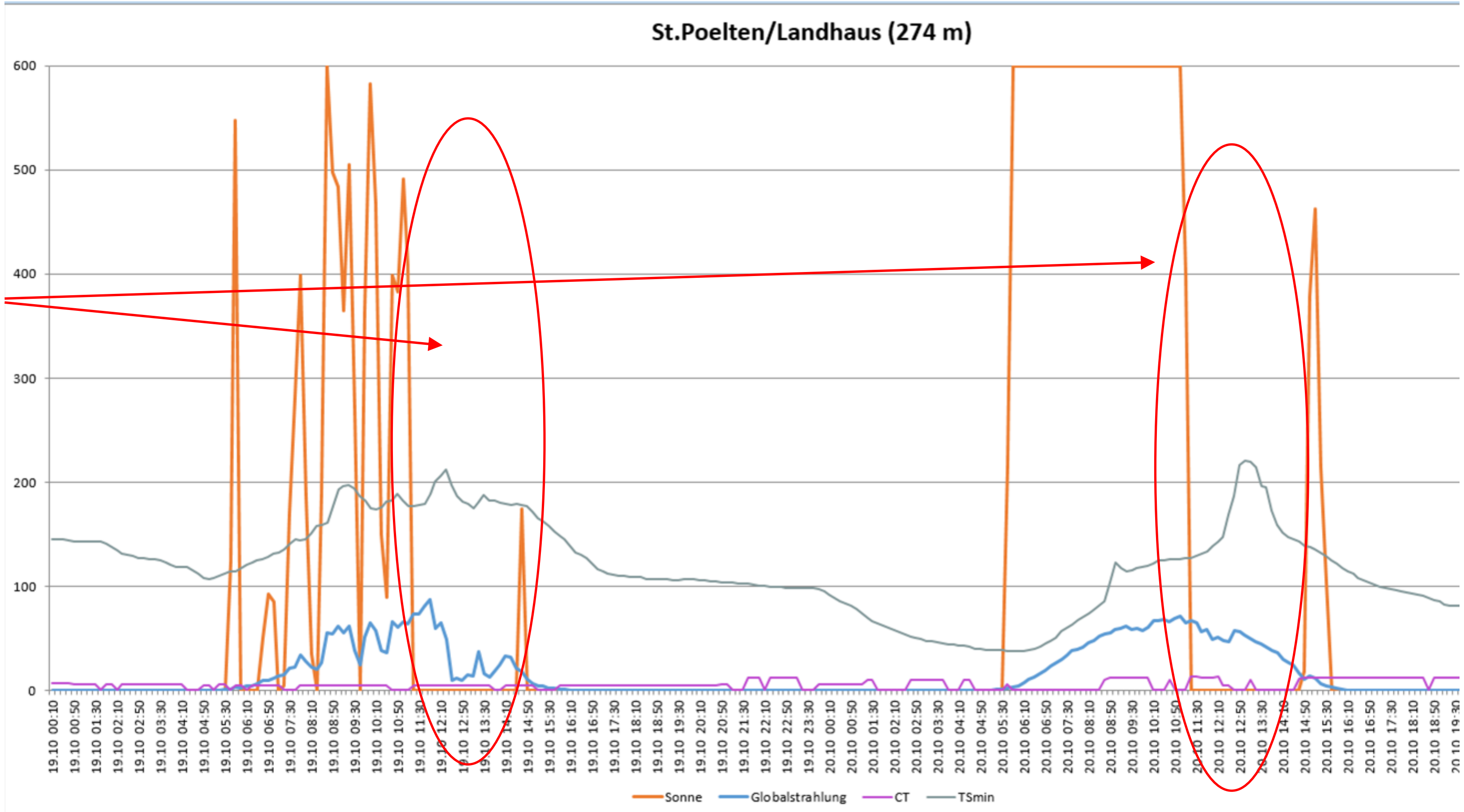
suspects	station ID
Litschau	500
Oberndorf/Melk	5412
St.Poelten/Landhaus	5609
Abtenau	9501
Ramsau	12711
Sonnblick	15411
Obertauern	15610
Katschberg	15715
Mariapfarr	15721
Schoeckl	16421
Flattnitz	18402
Weitensfeld	18502
Friesach	18601
Villacher Alpe	20021

cooks_D	station ID
Bregenz	11104
Rohrspitz	11146
Koeflach	16308
Gleisdorf	16501
Lassnitzhoehe	16511

Example: impact of a levelling error on the sunshine duration

No sunshine registration during a certain period of the day.

Inspection by the technical staff reveald a levelling error (20° instead of 50°)



Requirements imposed on AQUAS

before AQUAS

- Processing data on daily basis
- Fixed time intervals: 1 or 10 minutes
- Station-wise processing
- fixed parameters

AQUAS

- arbitrary time intervals
- parameter-wise processing
- arbitrary parameters and check-routines
- automated check in real-time
- flexible implementation of new stations or stations from third-party networks
- Documentation of all manipulations of data for complete tracking of data changes

✓

Benefits and costs by implementing new system at GeoSphere

benefits	costs
Flexibility • parameter- and site-specific check, • easy extension of the scope through adding new stations or networks, …	enhanced configuration effort
near real-time operation	limited availability of reference values at the time of the analysis
• one single comprehended system • consistence between input parameters and derived products (from 1min up to monthly data)	Sometimes compromise solutions are needed

Outlook

Future attempts

- Wind:
  - Operational spike-test implementation
  - Imputation of wind speed/direction data
  - Dealing with dynamic thresholds
- Precipitation
  - Detection of weighing rain gauge malfunctions
    - spurious measurements
    - missing rainfall observations when the weight increases and other sources detect rainfall (RRM, PWS))
- Rel. humidity
  - Spatial check in regions of high station density (e.g. MA22-network)
- In general: determination of „natural neighbours“ based on statistical approaches

GeoSphere Austria

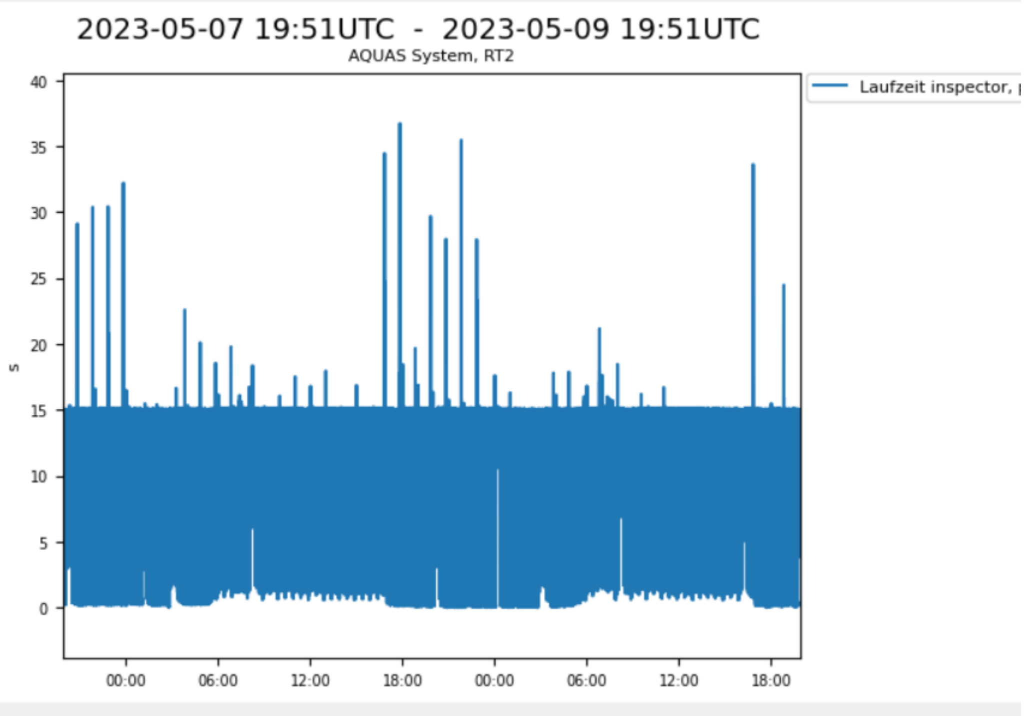
THANK YOU

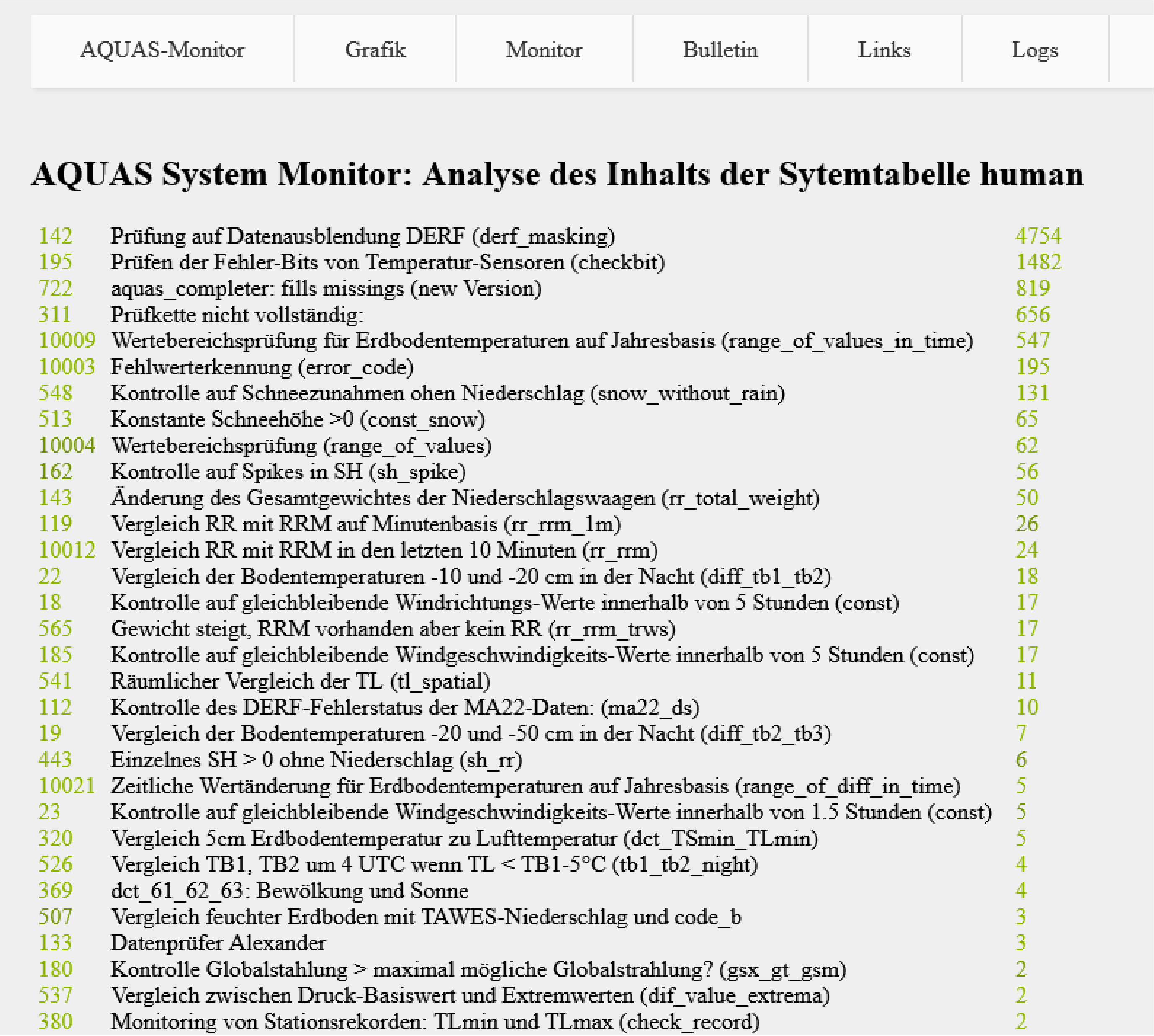
Data Quality and Digitization
Niko Filipović
niko.filipovic@geosphere.at

AQUAS Monitor

System monitoring tool for Data Quality staff to see which checks are active.

Many different options implemented for work management and monitoring of the system performance





AQUAS System Monitor: Analyse des Inhalts der Sytemtabelle human

142	Prüfung auf Datenausblendung DERF (derf_masking)	4754
195	Prüfen der Fehler-Bits von Temperatur-Sensoren (checkbit)	1482
722	aquas_completer: fills missings (new Version)	819
311	Prüfkette nicht vollständig:	656
10009	Wertebereichsprüfung für Erdbodentemperaturen auf Jahresbasis (range_of_values_in_time)	547
10003	Fehlwerterkennung (error_code)	195
548	Kontrolle auf Schneezunahmen ohen Niederschlag (snow_without_rain)	131
513	Konstante Schneehöhe >0 (const_snow)	65
10004	Wertebereichsprüfung (range_of_values)	62
162	Kontrolle auf Spikes in SH (sh_spike)	56
143	Änderung des Gesamtgewichtes der Niederschlagswaagen (rr_total_weight)	50
119	Vergleich RR mit RRM auf Minutenbasis (rr_rrm_1m)	26
10012	Vergleich RR mit RRM in den letzten 10 Minuten (rr_rrm)	24
22	Vergleich der Bodentemperaturen -10 und -20 cm in der Nacht (diff_tb1_tb2)	18
18	Kontrolle auf gleichbleibende Windrichtungs-Werte innerhalb von 5 Stunden (const)	17
565	Gewicht steigt, RRM vorhanden aber kein RR (rr_rrm_trws)	17
185	Kontrolle auf gleichbleibende Windgeschwindigkeits-Werte innerhalb von 5 Stunden (const)	17
541	Räumlicher Vergleich der TL (tl_spatial)	11
112	Kontrolle des DERF-Fehlerstatus der MA22-Daten: (ma22_ds)	10
19	Vergleich der Bodentemperaturen -20 und -50 cm in der Nacht (diff_tb2_tb3)	7
443	Einzelnes SH > 0 ohne Niederschlag (sh_rr)	6
10021	Zeitliche Wertänderung für Erdbodentemperaturen auf Jahresbasis (range_of_diff_in_time)	5
23	Kontrolle auf gleichbleibende Windgeschwindigkeits-Werte innerhalb von 1.5 Stunden (const)	5
320	Vergleich 5cm Erdbodentemperatur zu Lufttemperatur (dct_TSmin_TLmin)	5
526	Vergleich TB1, TB2 um 4 UTC wenn TL < TB1-5°C (tb1_tb2_night)	4
369	dct_61_62_63: Bewölkung und Sonne	4
507	Vergleich feuchter Erdboden mit TAWES-Niederschlag und code_b	3
133	Datenprüfer Alexander	3
180	Kontrolle Globalstahlung > maximal mögliche Globalstrahlung? (gsx_gt_gsm)	2
537	Vergleich zwischen Druck-Basiswert und Extremwerten (dif_value_extrema)	2
380	Monitoring von Stationsrekorden: TLmin und TLmax (check_record)	2