A STATISTICAL MODEL FOR AUTOMATED QUALITY ASSESSMENT OF TOAR-II

NAJMEH KAFFASHZADEH*[1], KAI-LAN CHANG[2], SABINE SCHRÖDER[1], AND MARTIN G. SCHULTZ[1]

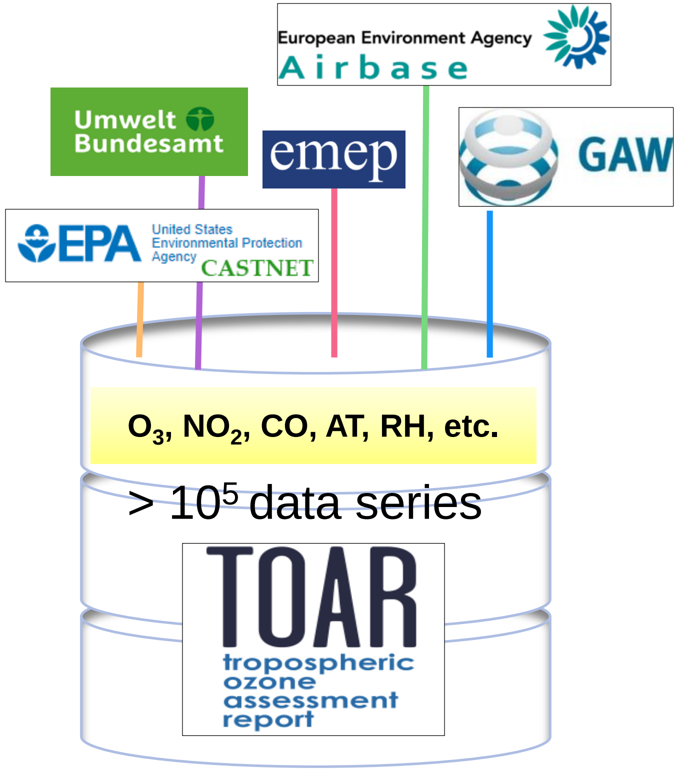
1 Jülich Supercomputing Centre, Forschungszentrum Jülich GmbH, Jülich, Germany
2 NOAA Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder/NOAA Chemical Sciences Laboratory, Boulder, CO, USA
* Corresponding author: Najmeh Kaffashzadeh

EGU General Assembly 2020

JÜLICH Forschungszentrum

SUMMARY

Motivation: assembling (air quality) data from many different sources requires a common data quality assessment (QA). [1]



Problem: big data (in terms of volume and variety) stresses the impossibility of manual QA

Solution: building an automated QA tool

AutoQA4Env

How: by developing data-driven statistical methods, where the tests assess the values' plausibility based on natural properties of the data

Demonstration: a schematic of implemented AutoQA4Env tool in the (TOAR) data ingestion workflow

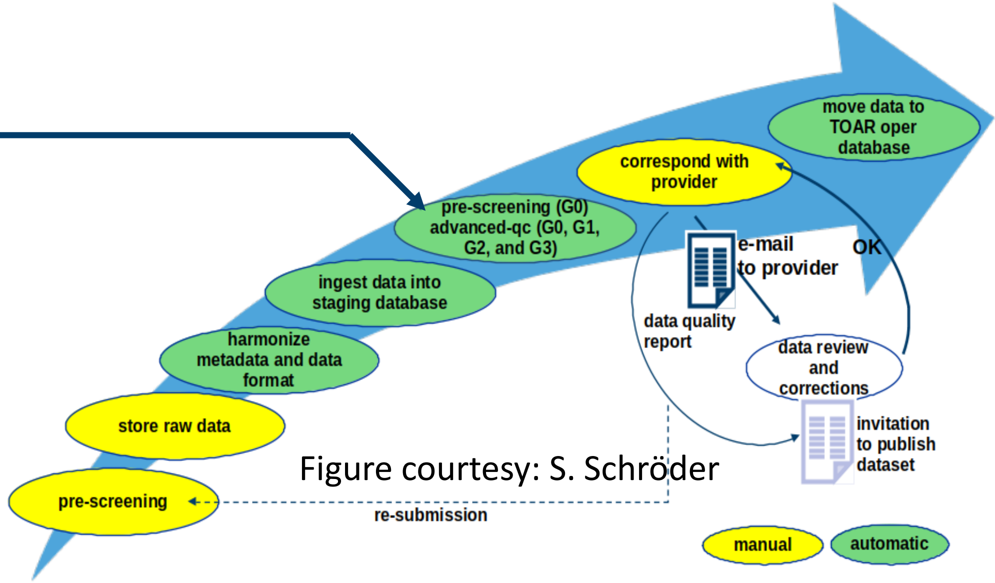


Figure courtesy: S. Schröder

n.kaffashzadeh@fz-juelich.de

PROBLEM STATEMENT

Erroneous values in a data time series

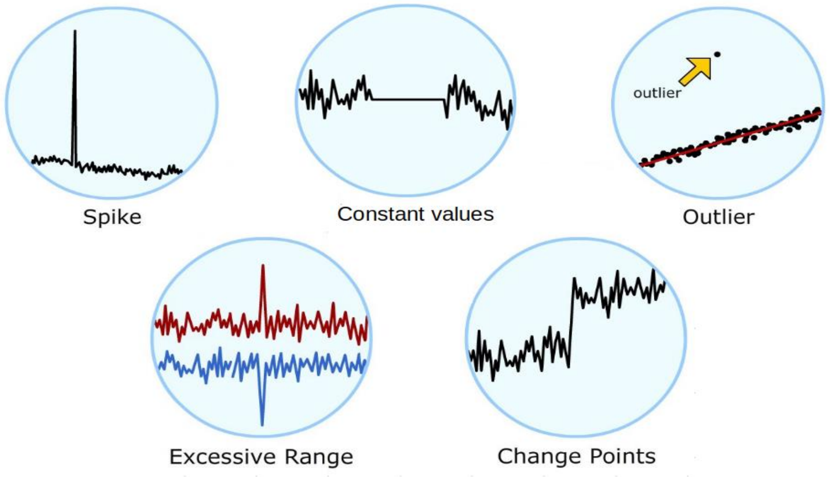


Figure 1. A schematic of regular data errors in a data time series. [2]

Figure 2. The measurements sensors can be damaged or destroyed by natural phenomena such as floods, fire, lightning strikes, and animal activity. [3]

Big data (volume and variety)

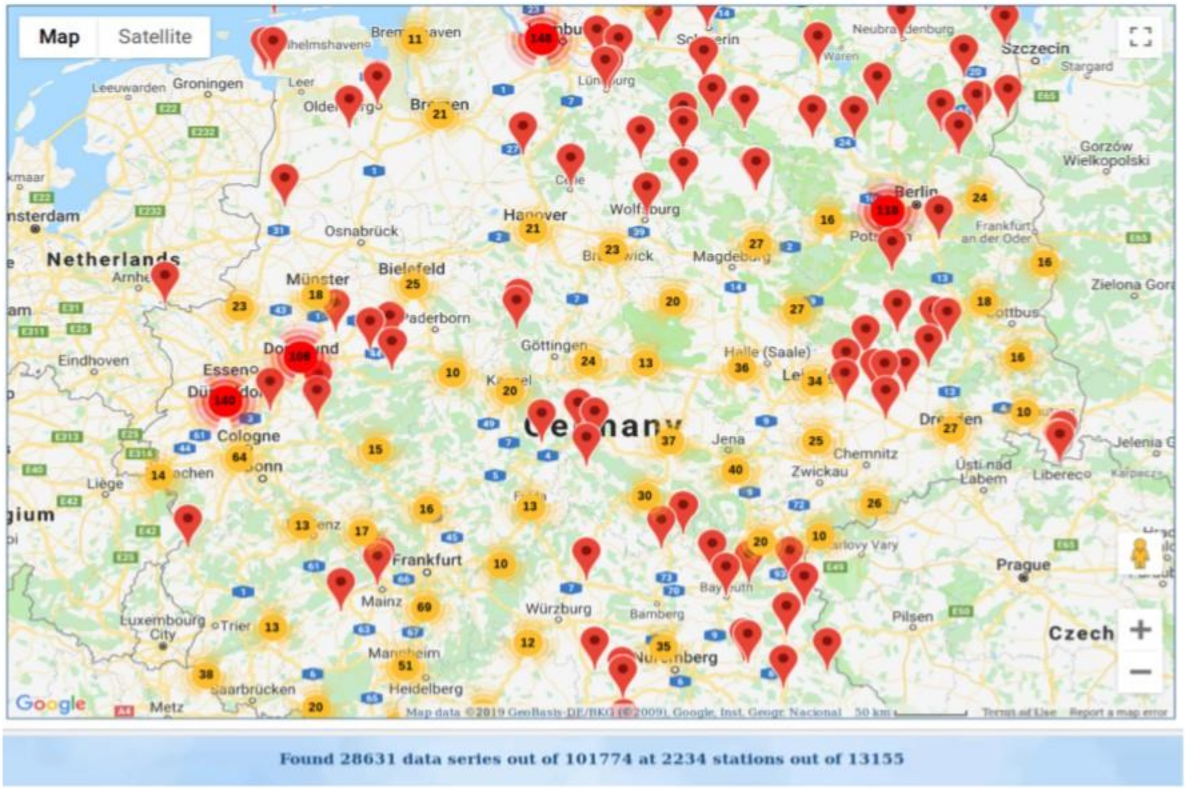


Figure 3. An snapshot of the stations over Germany in the TOAR database.

n.kaffashzadeh@fz-juelich.de

FIRST RESEARCH QUESTION

Can we have a tool to automatically, i.e. not manually, check the quality of big data?

Aims:

- to reduce delays in releasing data products

- to enhance data integrity and reliability

- to ensure consistency and reduce human bias

n.kaffashzadeh@fz-juelich.de

AUTOQC4ENV FRAMEWORK

Several statistical tests were classified into a few sub-groups as:

G0: mixed tests (range, constant value, step, etc.) with liberal thresholds to exclude large gross errors before further analysis

G1: single value tests (negative value, range)

G2: neighboring tests (constant value, step, spike)

G3: spatial consistency tests (statistical distributions)

G4: internal consistency tests (correlation)

G5: deep learning

Figure 4. A schematic of the AutoQA4Env framework

Implemented (only flagging system)!

Not implemented yet!

n.kaffashzadeh@fz-juelich.de

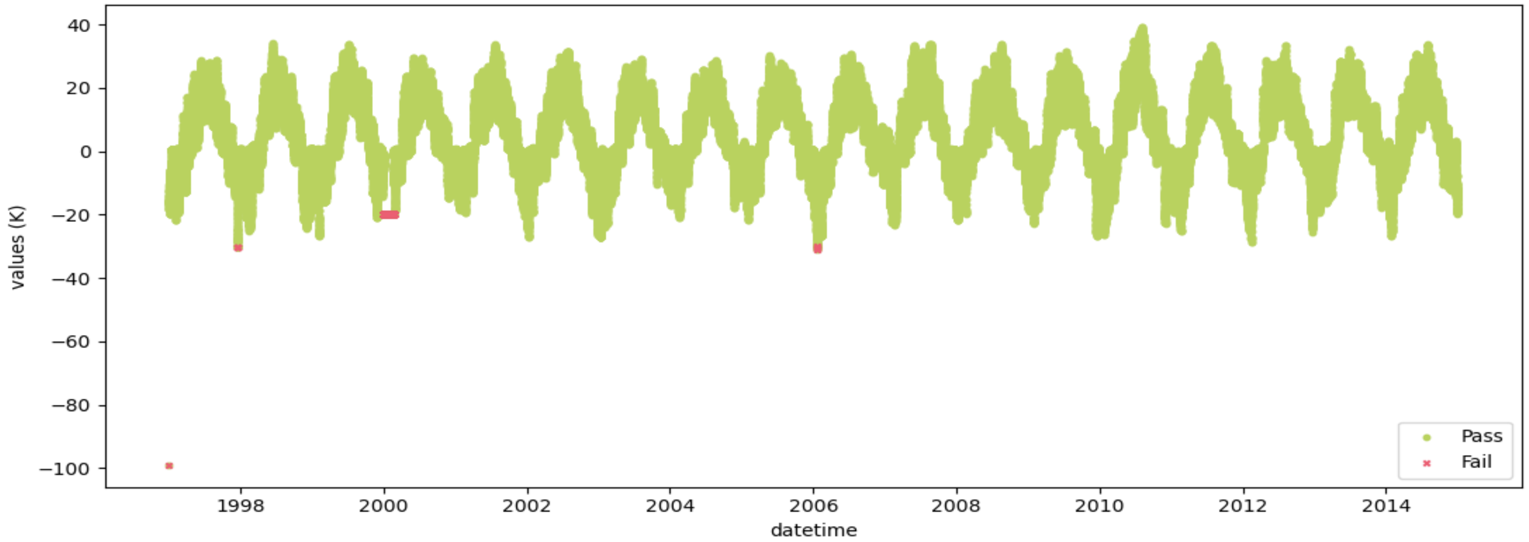
AN EXAMPLE OUTPUT FROM AUTOQA4ENV

TestsGroupG0.json

```
{
  "RangeTest": {
                "range_min": -60,
                "range_max": 80,
                "qc_group_name": "g0"
               }
}
```

TestsGroupG1.json

```
{
  "NegativeValueTest":
                     {
                      "negative_value_first": -40,
                      "negative_value_second": -30,
                      "qc_group_name": "g1"
                     },
  "ConstantValueTest":
                      {
                       "constant_value_stuck_points": 48,
                       "qc_group_name": "g1"
                      }
}
```



	the number of pass (%)	the number of fail (%)
g1_constant_value_test	[156567, ' (99.24)']	[1201, ' (0.76)']
g1_negative_value_test	[157756, ' (99.99)']	[12, ' (0.01)']
g0_range_test	[157767, ' (100.0)']	[1, ' (0.0)']

Figure 5. The QA results from a hourly time series of temperature at an unknown station. The data was retrieved from TOAR database and a stretch of constant values and a value out of range were added to the time series for the demonstration purposes. The constant value test were customized based on the suggested approach in [4].

These results can be regenerated by using the code and sample data in:
https://b2share.fz-juelich.de/records/f79417f0a7eb4db7818e6e4e3c0163e7
Last access: 29.04.2020

SCAN ME

n.kaffashzadeh@fz-juelich.de

AN EXAMPLE OUTPUT FROM AUTOQA4ENV

Although here an advanced flagging system was used, still fixed quality classifications are used!

TestsGroupG1.json

```
{
  "GrossRangeTest": {
                     "sensor_min": -10,
                     "sensor_max": 500,
                     "user_min": -2,
                     "user_max": 200,
                     "qc_group_name": "g1"
                     }
}
```

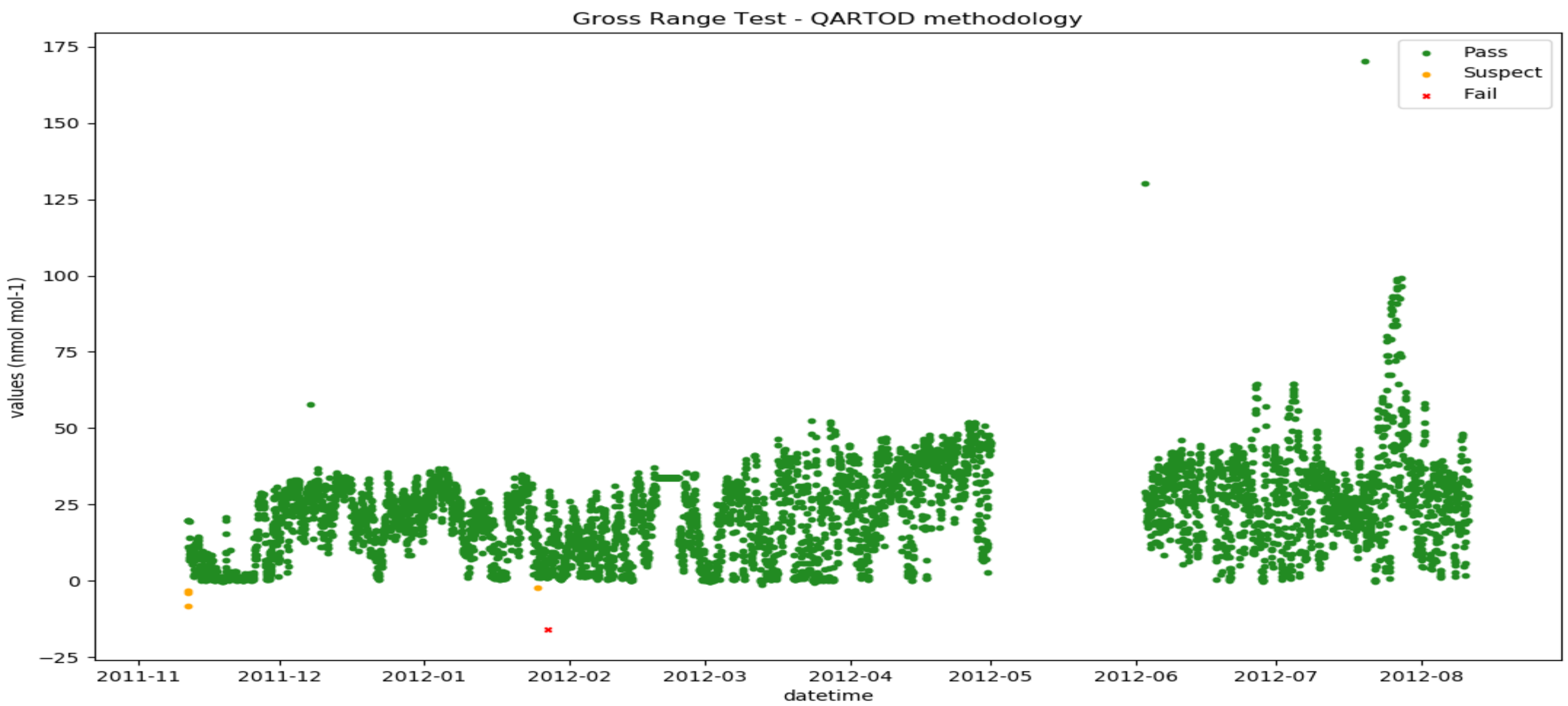


Figure 6. The QA results from a hourly time series of temperature at an unknown station. The data was retrieved from TOAR database and a few negative values (out of range) were added to the time series for the demonstration purposes. The gross range test were implemented based on the suggested approach in [5].

These results can be regenerated by using the code and sample data in:
https://b2share.fz-juelich.de/records/9afba748f2f943f5a73e6b6b919ce3c2
Last access: 30.04.2020

SCAN ME

n.kaffashzadeh@fz-juelich.de

SECOND RESEARCH QUESTION

Can we quantify the quality of a data, e.g. in a range of (0, 1), instead of using fixed quality classifications?

Aims:

- to provide a practical measure of the data quality

- to take into account the (tests and data) uncertainties

> 100 qualifiers code

Qualifier Code	Qualifier Description	Qualifier Type	Qaulifier Type Code	Still Active	Legacy Code
1	Deviation from a CFR/Critical Criteria Requirement.	Quality Assurance Qualifier	QA	YES	
1C	A 1-Point QC check exceeds acceptance criteria but there is compelling evidence that the analyzer data is valid.	Null Data Qualifier	NULL	YES	
1V	Data reviewed and validated.	Quality Assurance Qualifier	QA	YES	
2	Operational Deviation.	Quality Assurance Qualifier	QA	YES	
3	Field Issue.	Quality Assurance Qualifier	QA	YES	
4	Lab Issue.	Quality Assurance Qualifier	QA	YES	
5	Outlier.	Quality Assurance Qualifier	QA	YES	
6	QAPP Issue.	Quality Assurance Qualifier	QA	YES	
7	Below Lowest Calibration Level.	Quality Assurance Qualifier	QA	YES	
8	QA/QC Unknown.	Quality Assurance Qualifier	QA	NO	
9	Negative value detected - zero reported.	Quality Assurance Qualifier	QA	YES	
A	High Winds.	Informational Only	INFORM	NO	
AA	Sample Pressure out of Limits.	Null Data Qualifier	NULL	YES	9967
AB	Technician Unavaliable.	Null Data Qualifier	NULL	YES	9968
AC	Construction/Repairs in Area.	Null Data Qualifier	NULL	YES	9969

Figure 7. A snapshot of qualifiers code taken from EPA [6]
Last access: 27.04.2020

n.kaffashzadeh@fz-juelich.de

METHODOLOGY

Probability concept

- it estimates the likelihood of a value’s validity or plausibility

- it provides a robust theoretical underpinning to the data quality

n.kaffashzadeh@fz-juelich.de

SPECIFIC PROBLEM STATEMENT

Data persistence

The occurrence of successive constant values episode (CVE) can be an indicative of sensor (system) failures or other measurement errors.

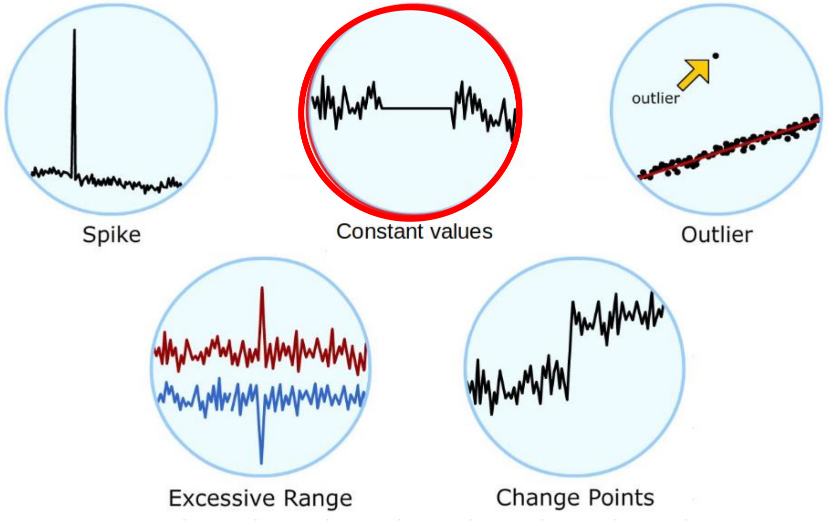


Figure 8. A schematic of regular data errors in a data time series. [2]
A red circle shows the focus of the next slides, i.e. constant values.

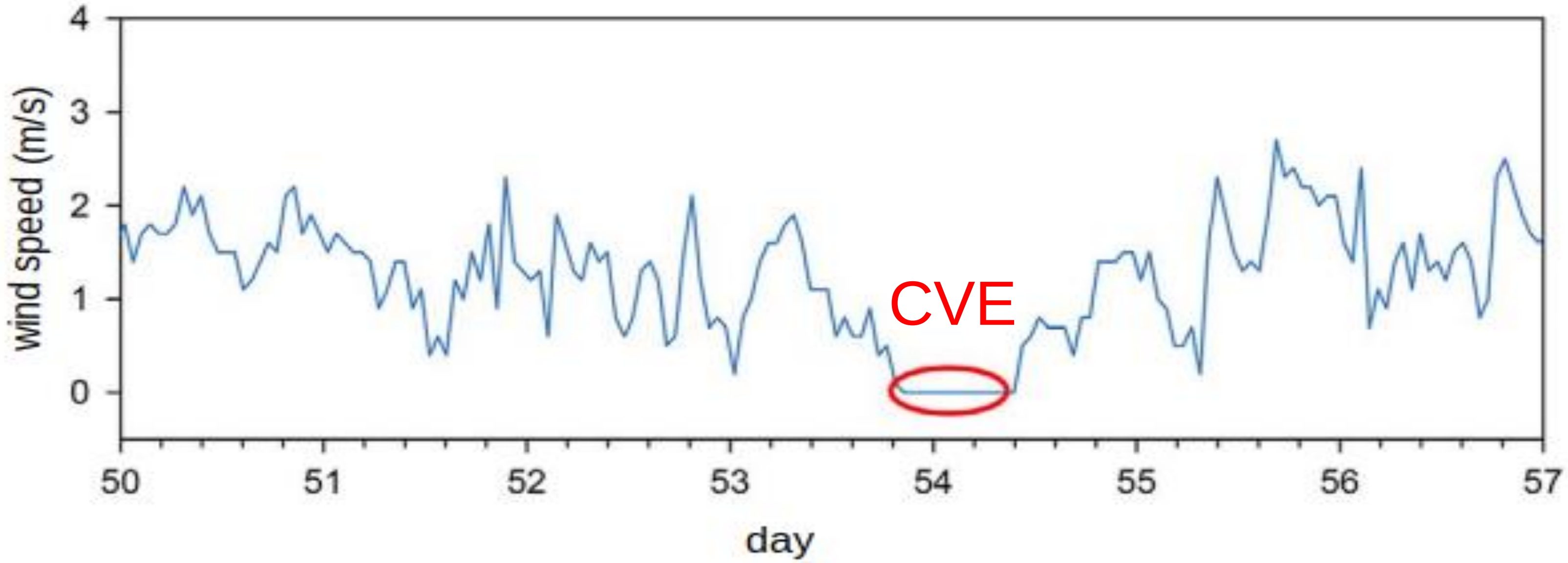


Figure 9. Snow and ice coated the anemometer propeller causing a constant zero wind speed reading. [3]

n.kaffashzadeh@fz-juelich.de

BOTTLENECK OF THE CVE

Too many CVEs in this time series! In the classical QA procedure, all the CVEs are excluded from the data. Why? Are they all erroneous data? Obviously Not.

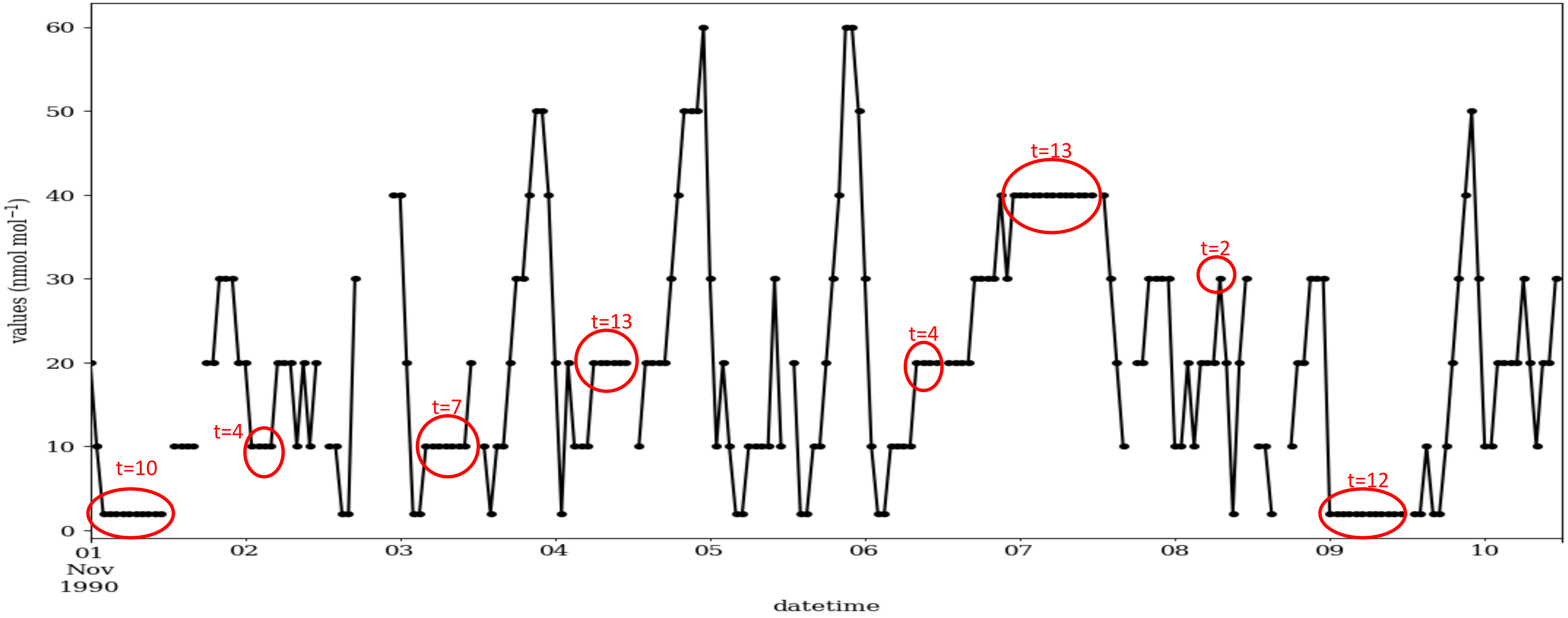


Figure 10. An hourly time series of ozone at the Azusa station where the data have been recorded at a low resolution, i.e. 10 ppb, in early period [7]. The data was retrieved from TOAR database. The red circles show several CVEs with a different length of t.

n.kaffashzadeh@fz-juelich.de

RESULT

A data-driven statistical test, constant value test (CVT), was developed to estimate the probability of CVEs plausibility.

The CVT:

- takes into account the uncertainty of the decision, data, tests, etc.

- is based on the statistical properties and a few assumptions of the data time series, e.g. stationarity.

- prevents excluding the valid CVEs in the QA procedures, which could lead to an additional bias in the analysis.

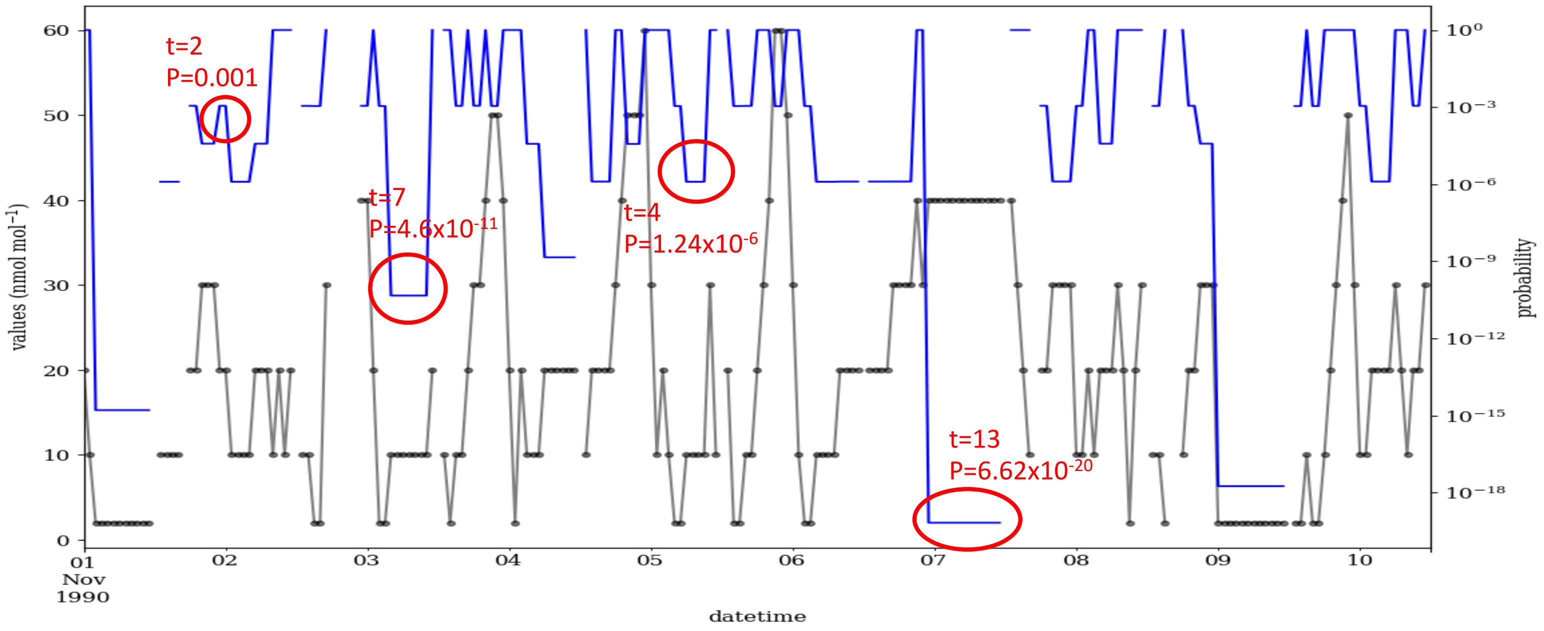


Figure 11. The results of performing the CVT on the ozone time series shown in Fig. 8. The black and blue lines show the time series and its associated probability. The red circles highlight several CVEs with a different length (t) and probability (P).

Paper in preparation!

n.kaffashzadeh@fz-juelich.de

RESULT

Here is another example of regular occurrence of CVEs in the temperature time series at the Cape Grim station.

None of these CVEs are an indicative of erroneous data. By estimating the probability via the CVT, there is more chance to not exclude them from the data series.

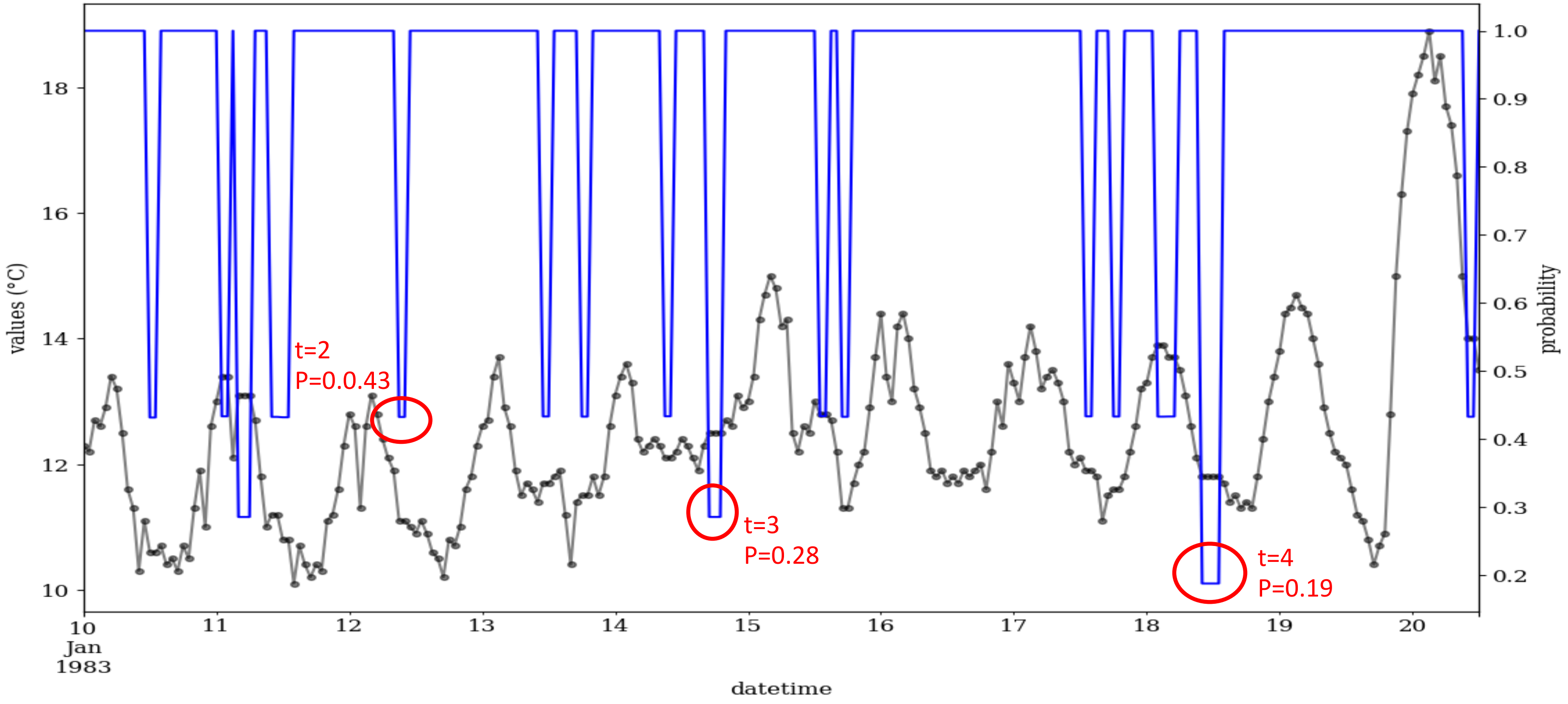


Figure 12. The results of performing the CVT on temperature time series at the Cape Grim station. The black and blue lines show the time series and its associated probability. The red circles highlight several CVEs with a different length (t) and probability (P).

Paper in preparation!

n.kaffashzadeh@fz-juelich.de

CONCLUSIONS

- So far, most of the QA procedures are manual and include subjective decisions.

- We need to build an automated tool for QA, e.g. AutoQA4Env, and likely for other data analysis in the era of big data.

- Developing data-driven statistical methods is one possible approach for building an automated tool.

- The CVT estimates the probability of CVEs based on the natural properties of the data time series.

- Using probability concept can prevent the exclusion of (many) valid values form the data series.

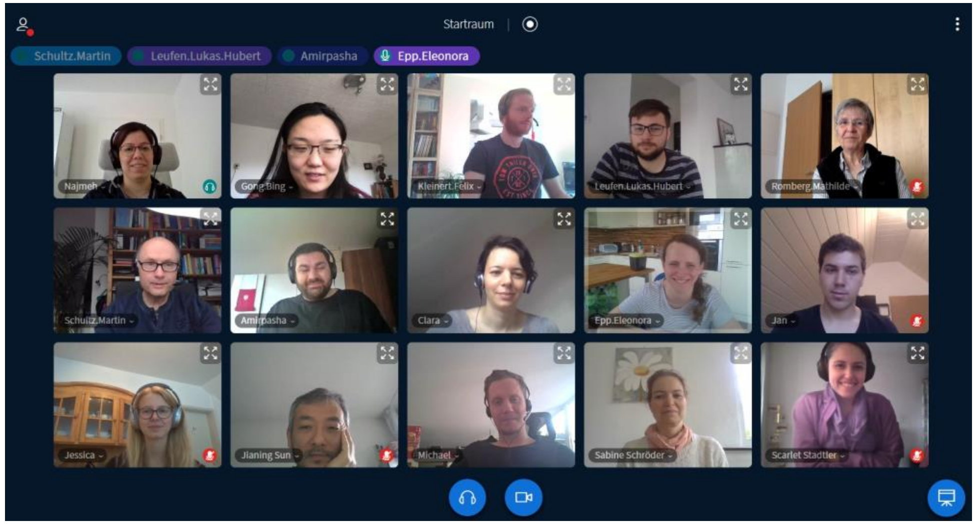
n.kaffashzadeh@fz-juelich.de

REFERENCES

1. Schultz, M.G. et al. (2017): Tropospheric Ozone Assessment Report: Database and Metrics Data of Global Surface Ozone Observations, Elementa Sci. Anthrop., https://doi.org/10.1525/elementa.244
2. https://www.climate.gov/maps-data/primer/processing-climate-data
3. J. Campbell: quantity is nothing without quality, BioScience,vol.63, no.7, pp.574–585,2013.
4. Gudmundsson et al., "The Global Streamflow Indices and Metadata Archive (GSIM) – Part 2: Quality Control, Time-Series Indices and Homogeneity Assessment."
5. Shkvorets, Igor & Bushnell, Mark & Worthington, Helen & Fredericks, Janet & Jaeger, Stephanie & Lankhorst, Matthias & Thomas, Julie. (2016). Manual for Real-Time Quality Control of In-situ Temperature and Salinity Data. A Guide to Quality Control and Quality Assurance for In-situ Temperature and Salinity Observations. 10.13140/RG.2.2.14814.95040.
6. https://aqs.epa.gov/aqsweb/documents/codetables/qualifiers.html
7. Tarasick D. et al. (2019): Tropospheric Ozone Assessment Report: Tropospheric ozone from 1877 to 2016, observed levels, trends and uncertainties. Elem Sci Anth, 7(1), p.39. DOI : 10.1525/elementa.376.

ACKNOWLEDGEMENTS

- ERC grant, number ERC-2017-ADG-787576, for financial support
- My colleagues at the JSC, in particular Earth System Data Exploration (ESDE) research group, for their technical support and advices in software development and data science
- Digital Earth community for encouraging us to step ahead in this field of research
- EGU conveners for helding the conference and creating an opportunity for sharing knowledge
- The Australian Bureau of Meteorology for providing the temperature time series data at the Cape Grim station
- The U.S. EPA for providing the ozone time series data at the Azusa station



A snapshot of a ESDE group meeting on 29.04.2020.

European Research Council
Established by the European Commission

IntelliAQ is funded by the EU's ERC programme, Grant Agreement ERC-2017-ADG-787576