A NEW TOOL FOR AUTOMATED QUALITY CONTROL OF ENVIRONMENTAL DATA IN OPEN WEB SERVICES

N. Kaffashzadeh, F. Kleinert, M. G. Schultz

July 22, 2019 | Najmeh Kaffashzadeh | Jülich Supercomputing Center (JSC)

Outline

1. Motivation
2. Background
   - Why do we need quality control (QC) for environmental data?
   - Current approach to environmental data QC
   - Requirements for environmental QC software
3. Objectives
   - A new methodology
   - The AutoQC4Env software framework
4. Results of a case study
5. Summary and conclusions

MOTIVATION

- Assembling (air quality) data from many different sources requires to ensure common quality for assessments
- Training deep neural networks requires good quality data

MOTIVATION

- Assembling (air quality) data from many different sources requires to ensure common quality for assessments
- Training deep neural networks requires good quality data

BACKGROUND

Why do we need quality control (QC) for environmental data?

- Obvious artifacts appear even in published data

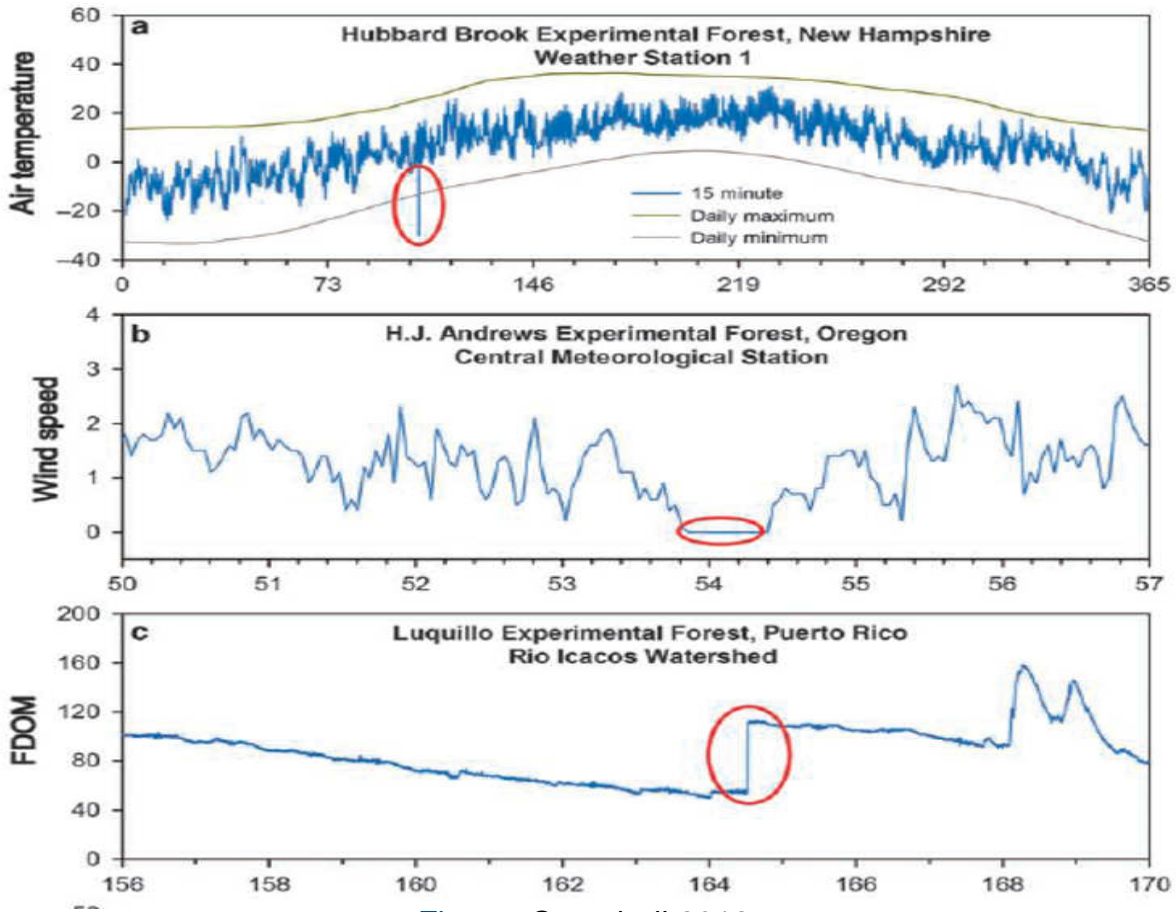


Figure: Campbell 2013

BACKGROUND

Current approach to environmental data QC

- Usually developed with a specific application focus
- Embedded in specific data processing workflows
- Inconsistent QC flags in different agencies
- Not fully transparent to data users

EBAS

Group 0: Valid data

Flag	Validity	Description
000	V	Valid measurement

Group 1: Exception flags for accepted, irregular data

Flag	Validity	Description
110	V	Episode data checked and accepted by data originator. Valid measurement
147	V	Below theoretical detection limit or formal QA limit, but a value has been measured and reported and is considered valid

Group 3: Flags for aggregated datasets

Flag	Validity	Description
390	V	Data completeness less than 50%
392	V	Data completeness less than 75%
394	V	Data completeness less than 90%

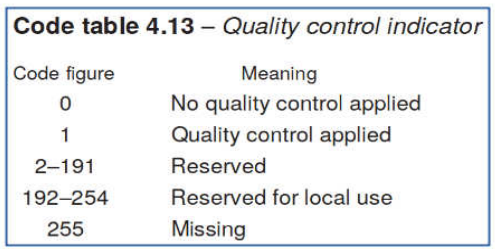
Group 5: Chemical problem

Flag	Validity	Description
559	V	Unspecified contamination or local influence, but considered valid
599	I	Unspecified contamination or local influence

Group 9: Missing flags

Flag	Validity	Description
999	M	Missing measurement, unspecified reason

WMO



Code table 4.13 – *Quality control indicator*

Code figure	Meaning
0	No quality control applied
1	Quality control applied
2–191	Reserved
192–254	Reserved for local use
255	Missing

> 150 Qualifier Code in EPA

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 analyzor 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: QC flags from various environmental agencies

BACKGROUND

Requirements for environmental QC software

- Many heterogeneous datasets with different sampling times, statistical distributions, etc. $\rightarrow$ need flexibility
- Should be independent of subjective human decisions
- Needs to process large amounts of data in short time
- Should be applicable to archived as well as real-time data

OBJECTIVES

A new methodology

Probability concept

- Estimates the likelihood of a value's validity
- Attempt to provide a robust theoretical underpinning to quality control
- Foundation: every QC test is in some way a statistical test
- Assumption: we can use uncertainty information from statistical testing to estimate the likelihood of a value's validity

A NEW METHODOLOGY

How to estimate the probability?

Example 1: statistical p-value as a *proxy*.

$$\text{if test } t \text{ is passed:} \quad prob_t = 1 - min(p\text{-}value, 0.5) \tag{1}$$

$$\text{if test } t \text{ fails:} \quad prob_t = 0 + min(p\text{-}value, 0.5) \tag{2}$$

- low p-value indicates robust test result, i.e. strong confidence about the data validity
- large p-value indicates unclear test result, therefore little information on data validity

A NEW METHODOLOGY

How to estimate the probability?

Example 2: extreme value distribution as basis for value-range check
Using 1−CDF[1] from the statistical distribution as a *proxy*

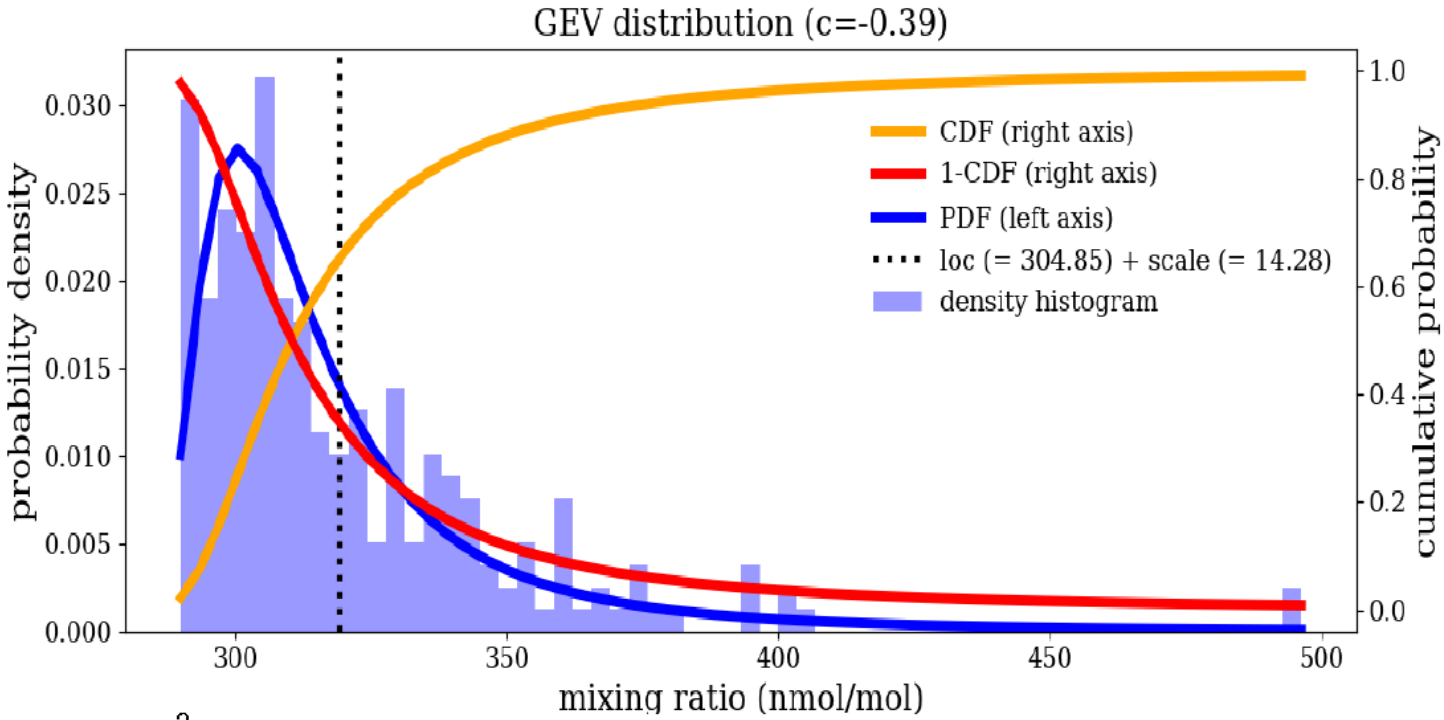


Figure: GEV[2] distribution derived from the 1000 largest ozone values measured after 1990 from the TOAR database.

[1] Cumulative Density Function
[2] Generalized Extreme Value

A NEW METHODOLOGY

How to estimate the overall probability (P)?

Assuming non-independent tests:

$$P = min(prob_t; t = 1, 2, ..., n) \tag{3}$$

n: the number of performed QC tests

OBJECTIVES

The AutoQC4Env software framework

Create a modern software package, which

- is easy to set up,
- allows user configuration,
- is well-documented,
- is applicable to various environmental time series,
- assists users in the right choice of statistical parameters,
- and is free and open-source.

Work in progress!

THE AutoQC4Env SOFTWARE FRAMEWORK

QC test workflow in AutoQC4Env

QC tests are categorized in different groups:

Group0: pre-screening tests (range, constant value, step, etc.) with liberal thresholds to remove large gross error

Group1: single value tests (negative value, range)

Group2: neighboring values tests (step, z-test, q-test, spike)

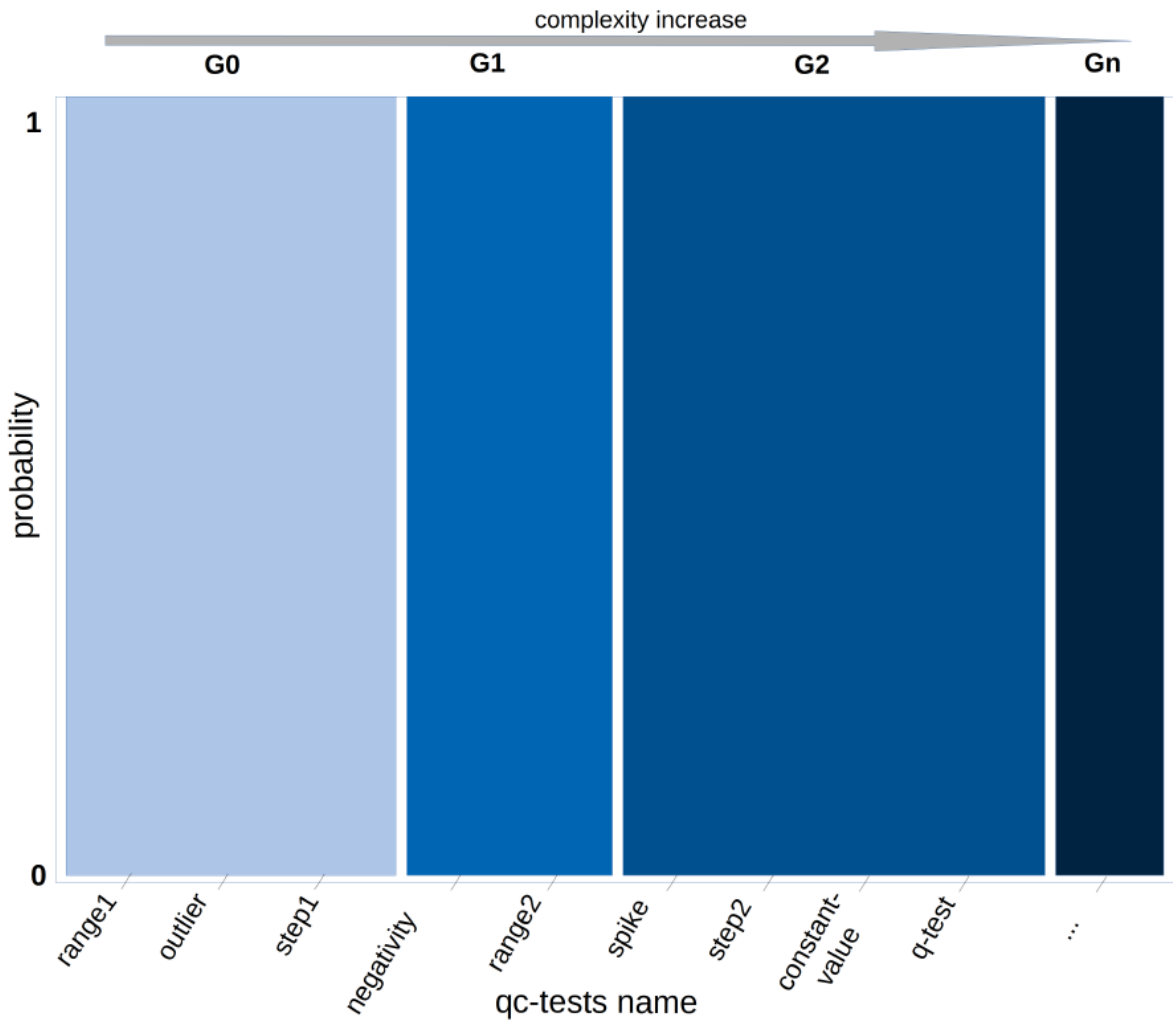
Group3: spatial consistency tests (statistical distributions)

Group4: internal consistency tests (correlation)

Group5: deep learning

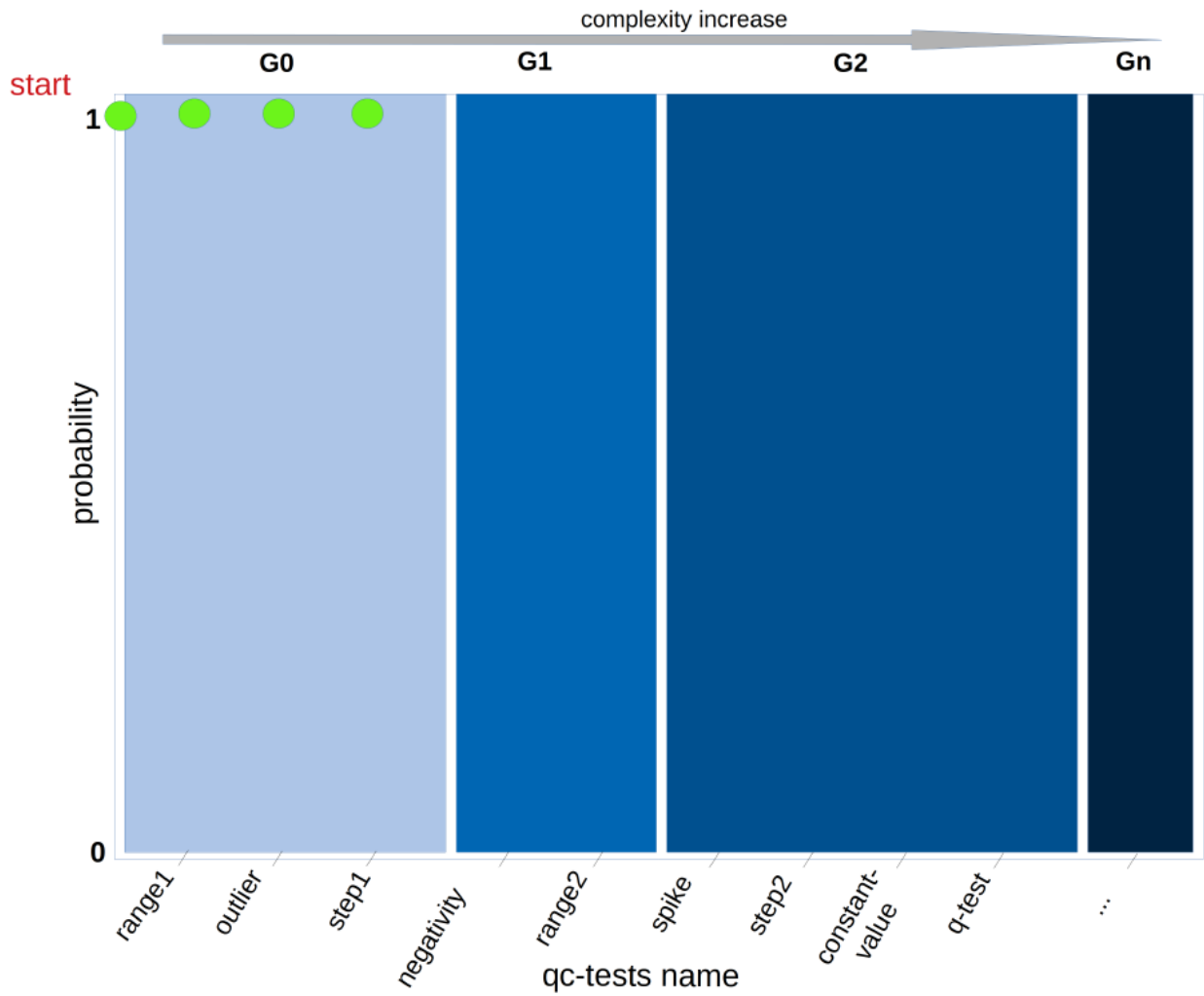
THE AutoQC4Env SOFTWARE FRAMEWORK

Implementation of the concept



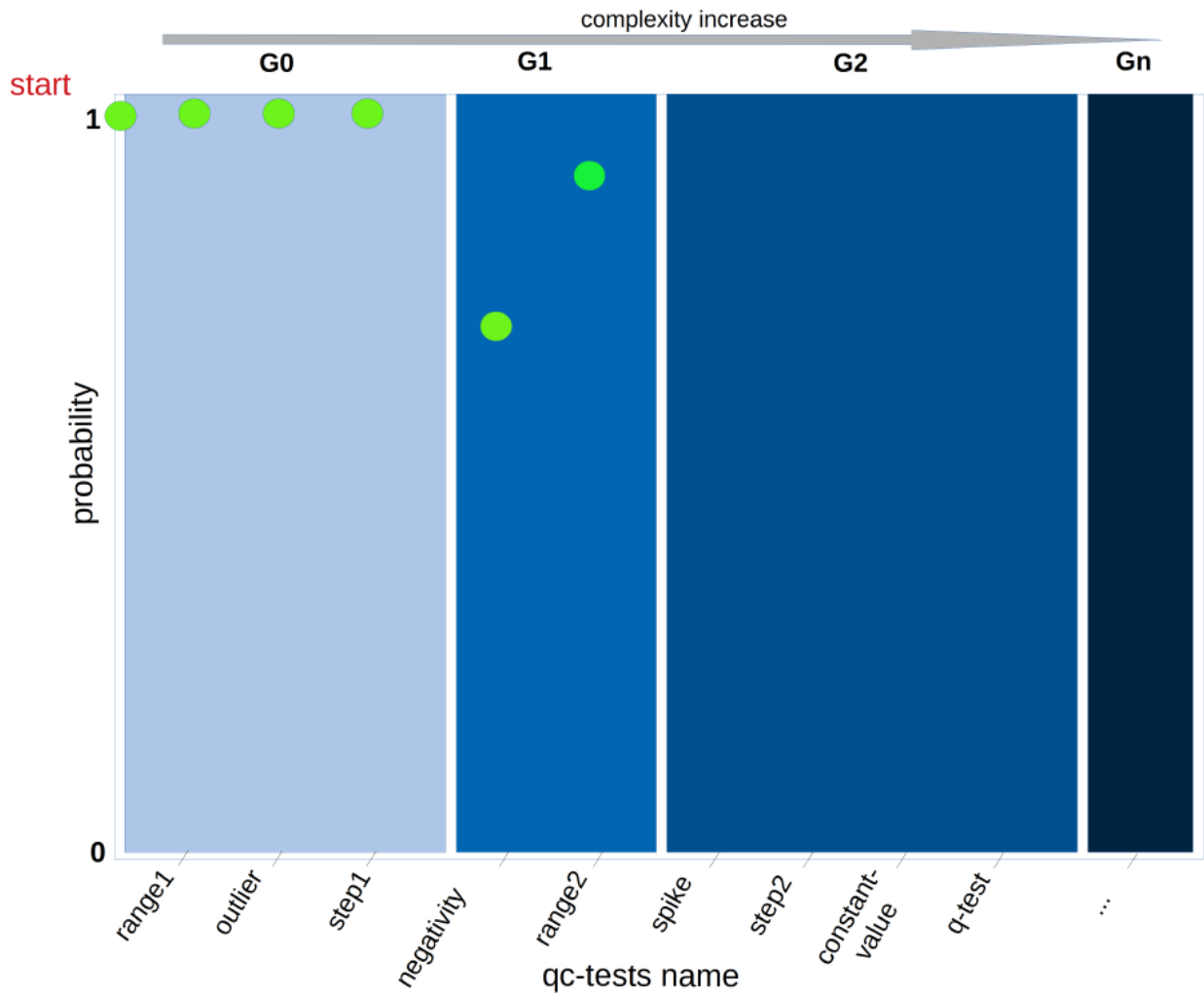
IMPLEMENTATION OF THE CONCEPT

Performing tests in G0



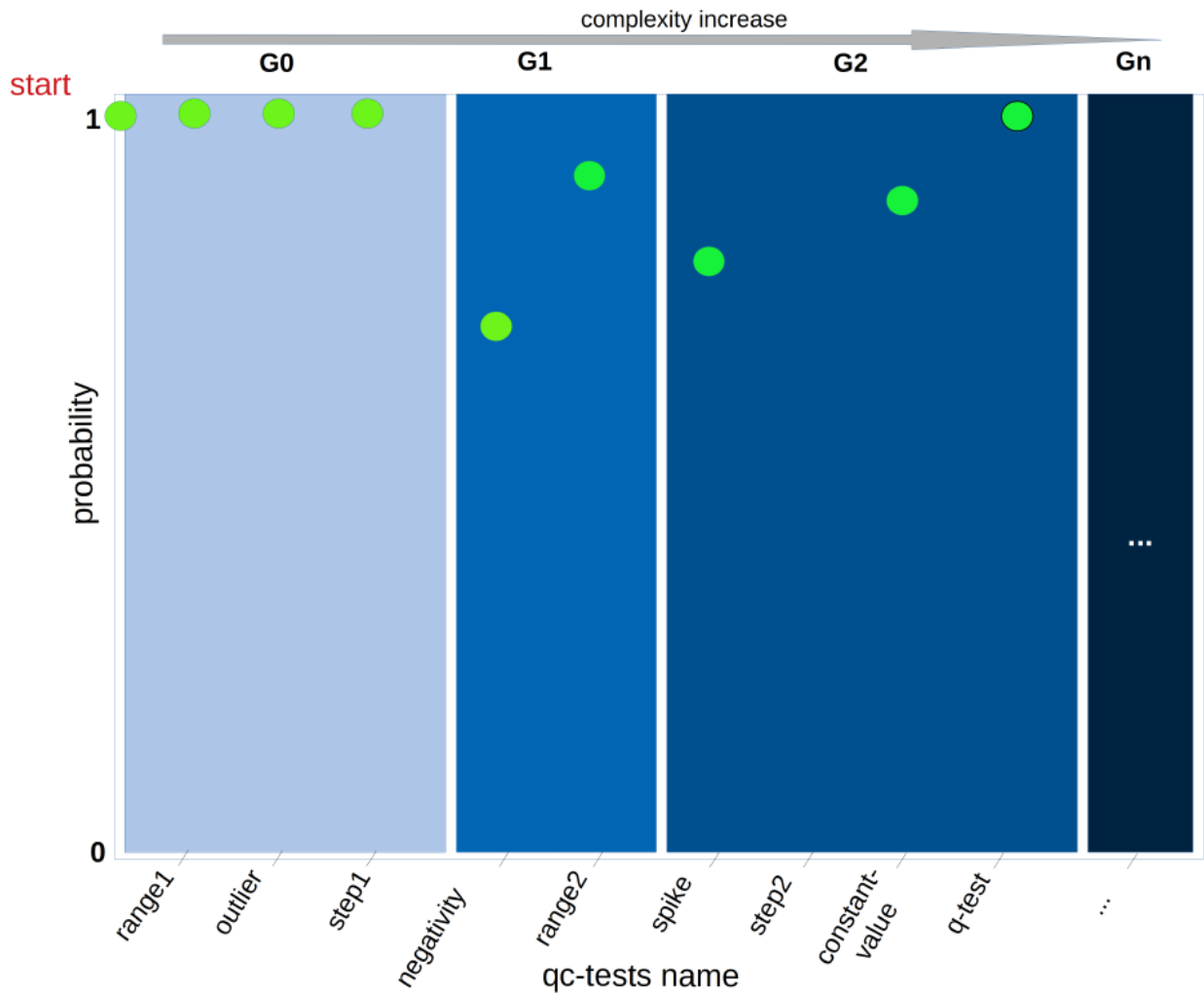
IMPLEMENTATION OF THE CONCEPT

Performing tests in G1



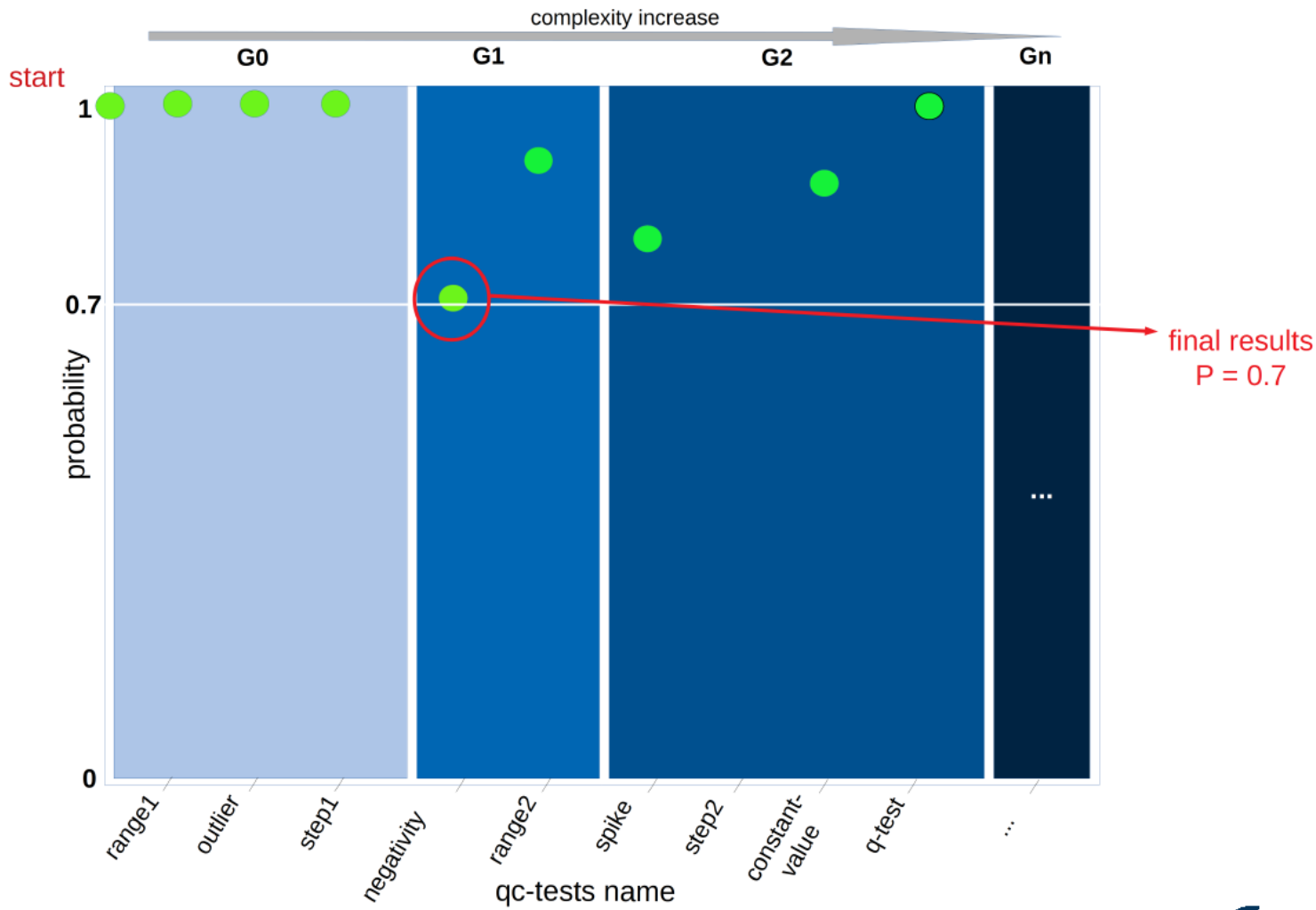
IMPLEMENTATION OF THE CONCEPT

Performing tests in G2



IMPLEMENTATION OF THE CONCEPT

Estimating the final probability (P)



RESULTS OF A CASE STUDY

Input data

Demonstration of typical environmental time series errors

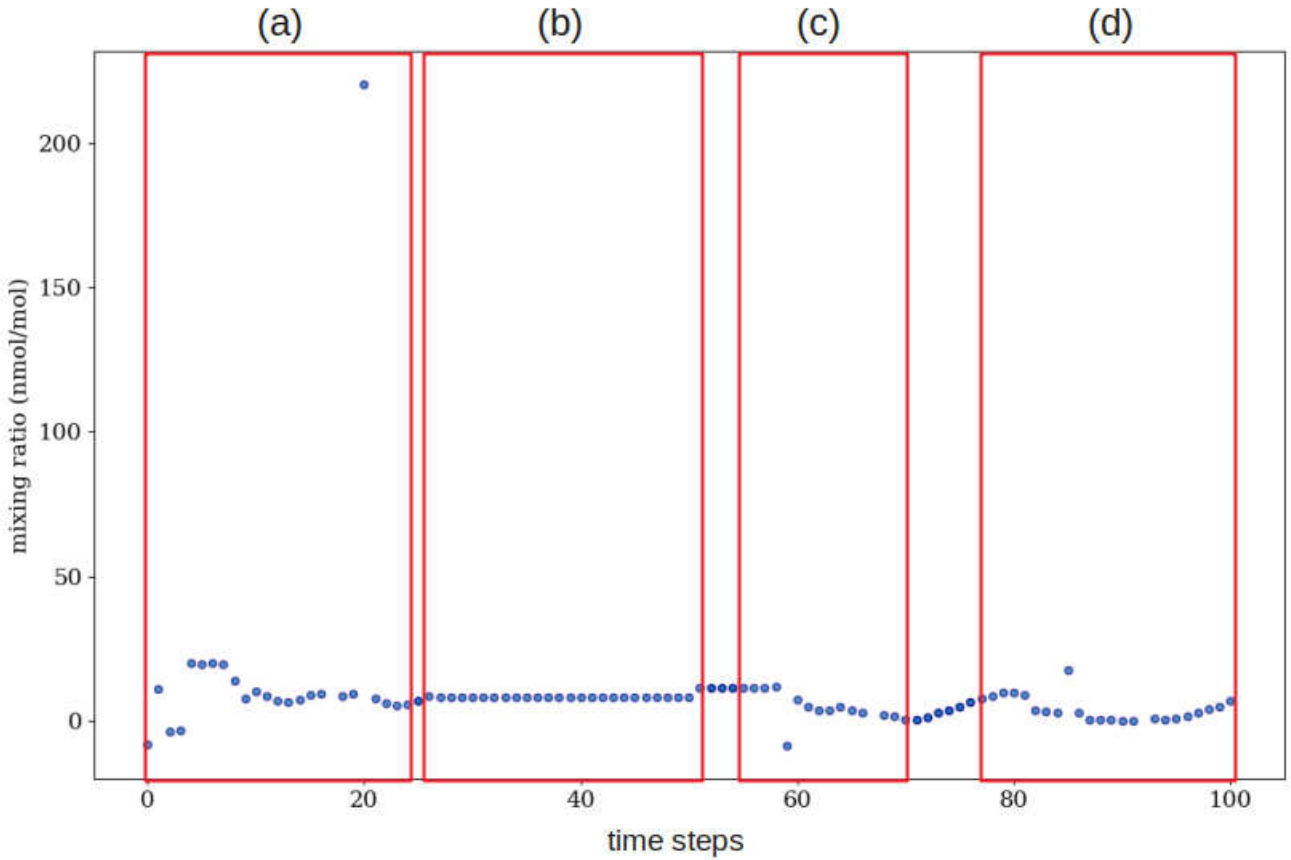


Figure: An arbitrarily selected ground-level ozone measurement series with typical error features

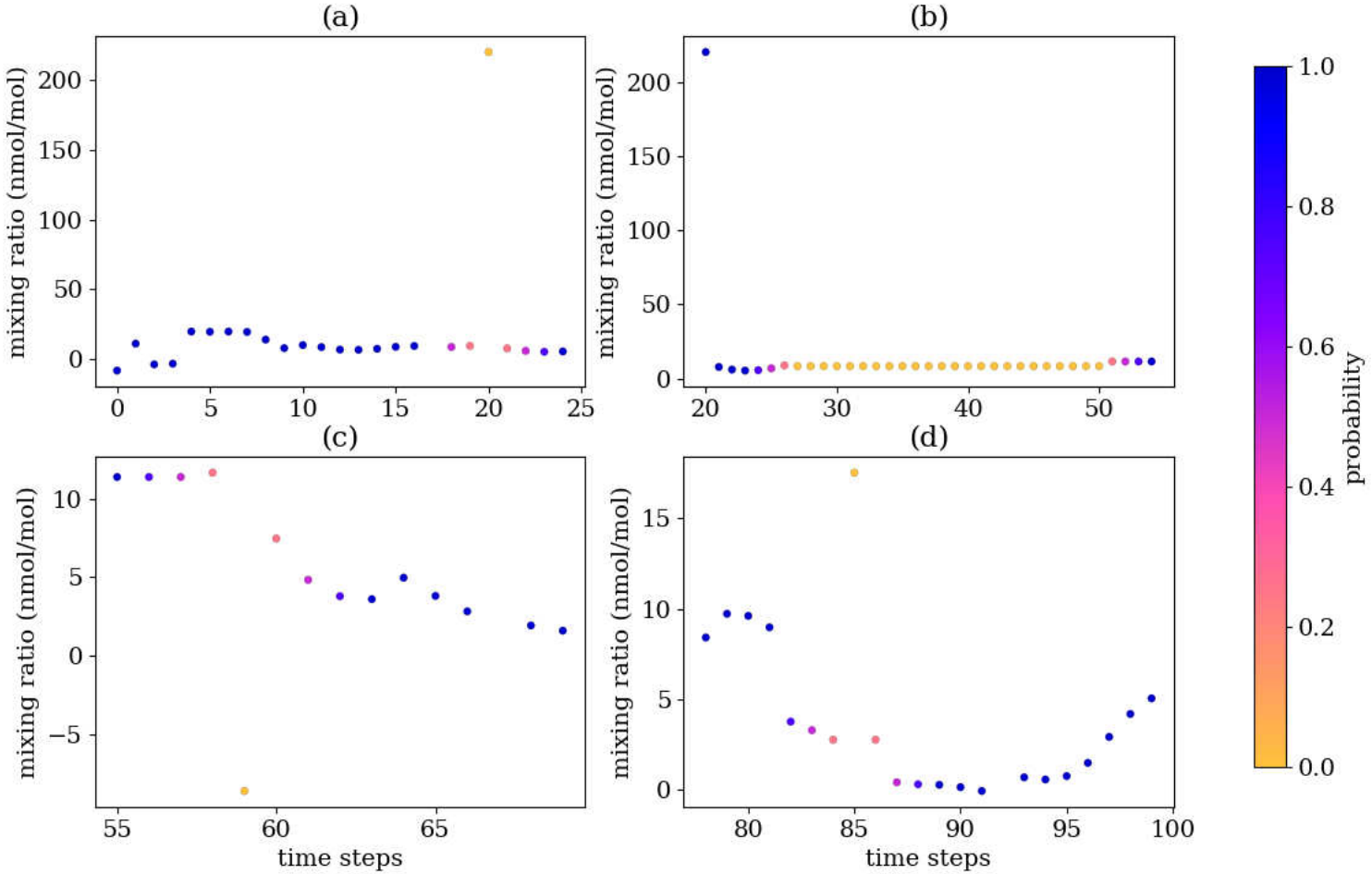
RESULTS OF A CASE STUDY

Sample user configuration settings

```
"RangeTest":{
            "range_min": -10,
            "range_max": 150,
            "range_neighboring_size": 3,
            "range_neighboring_side": "both",
            "qc_group_name": "g0"
            }
```

RESULTS OF A CASE STUDY

QC tests output

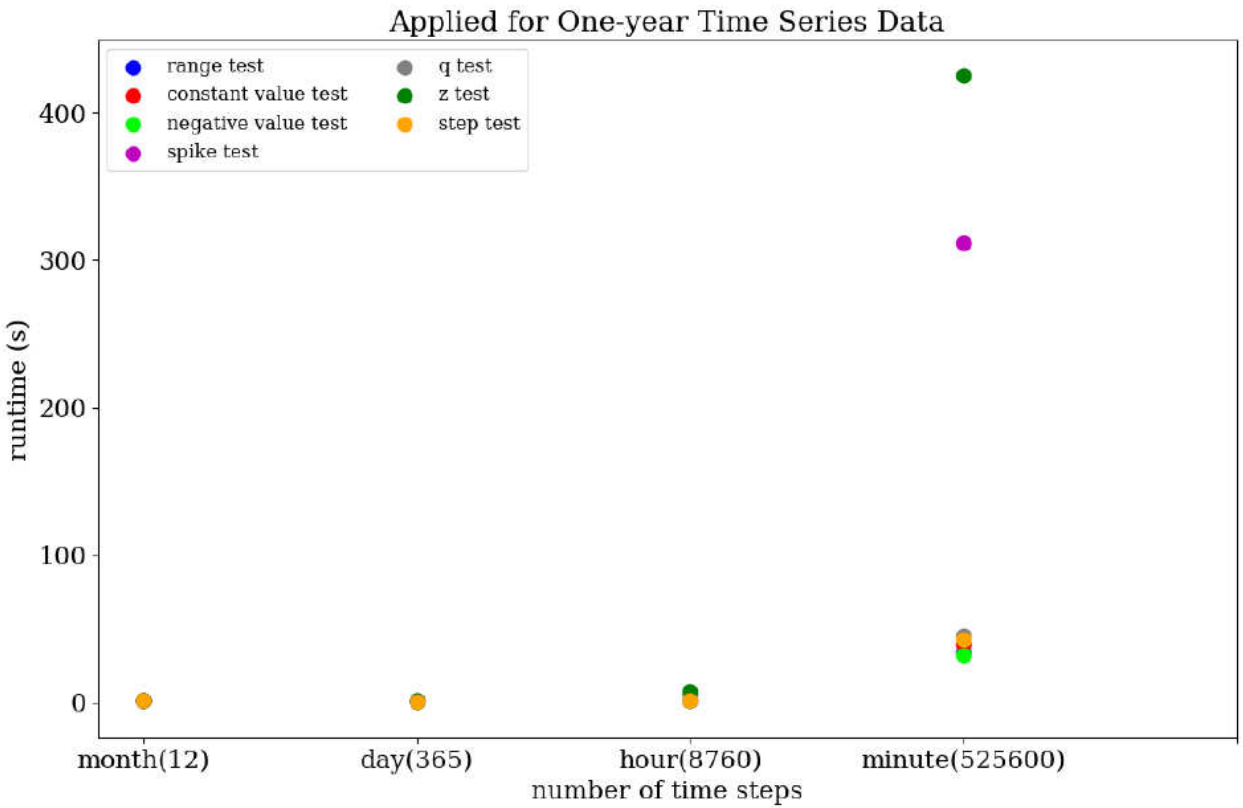


Users may then decide below which probability data should not be used in their analysis.

JÜLICH Forschungszentrum

RESULTS OF A CASE STUDY

Code performance



For large amounts of data, code-parallelization will be required.

Summary and Conclusions

- We introduced a novel probability concept
- We began to construct a software framework to implement this concept
- AutoQC4Env allows easy configuration for specific use cases
- AutoQC4Env is intended for a wide range of environmental applications

Acknowledgements

- Felix Kleinert
- Martin G. Schultz
- To my colleagues at the JSC

Figure: The ESDE[1] group members on 12 June 2019

European Research Council
Established by the European Commission

BUSINESS INFORMATION SYSTEMS 2019

[1] Earth System Data Exploration

How to access to the AutoQC4Env?

It is available in a jugit repository as:

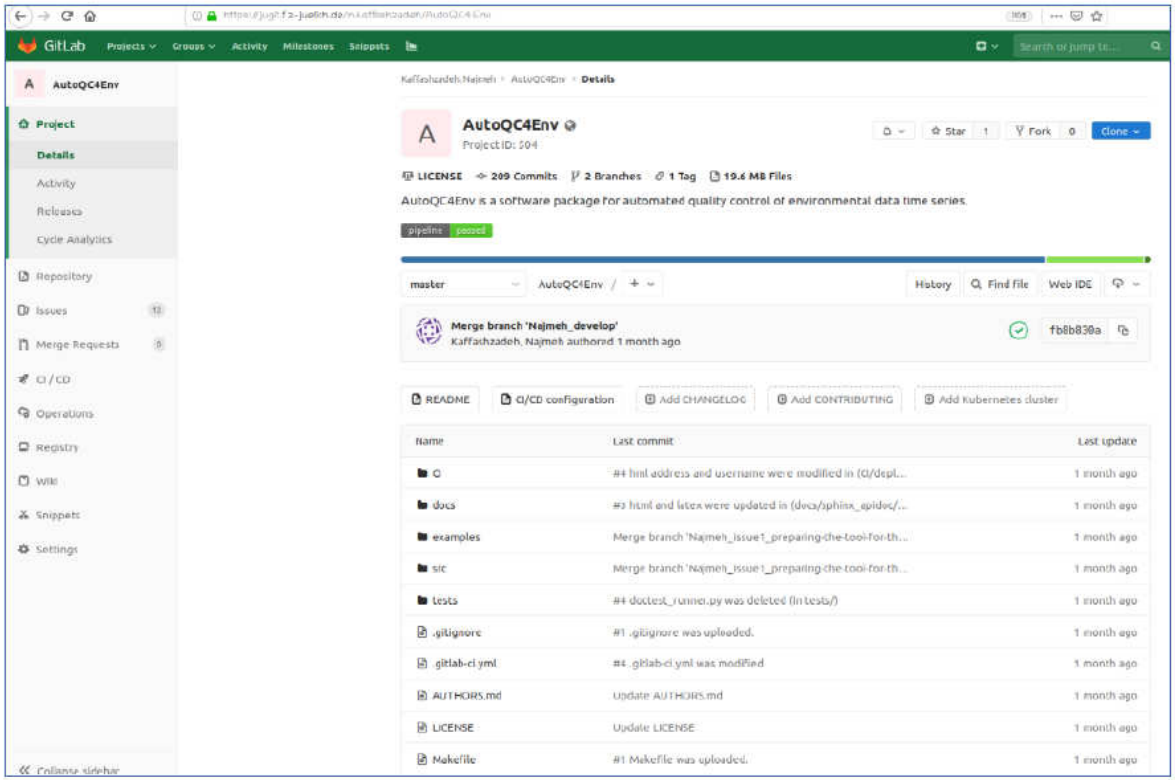


Figure: A screenshot of the repo at the
`https://jugit.fz-juelich.de/n.kaffashzadeh/AutoQC4Env`

Scan me

How to access to the AutoQC4Env?

It is available in a jugit repository as:

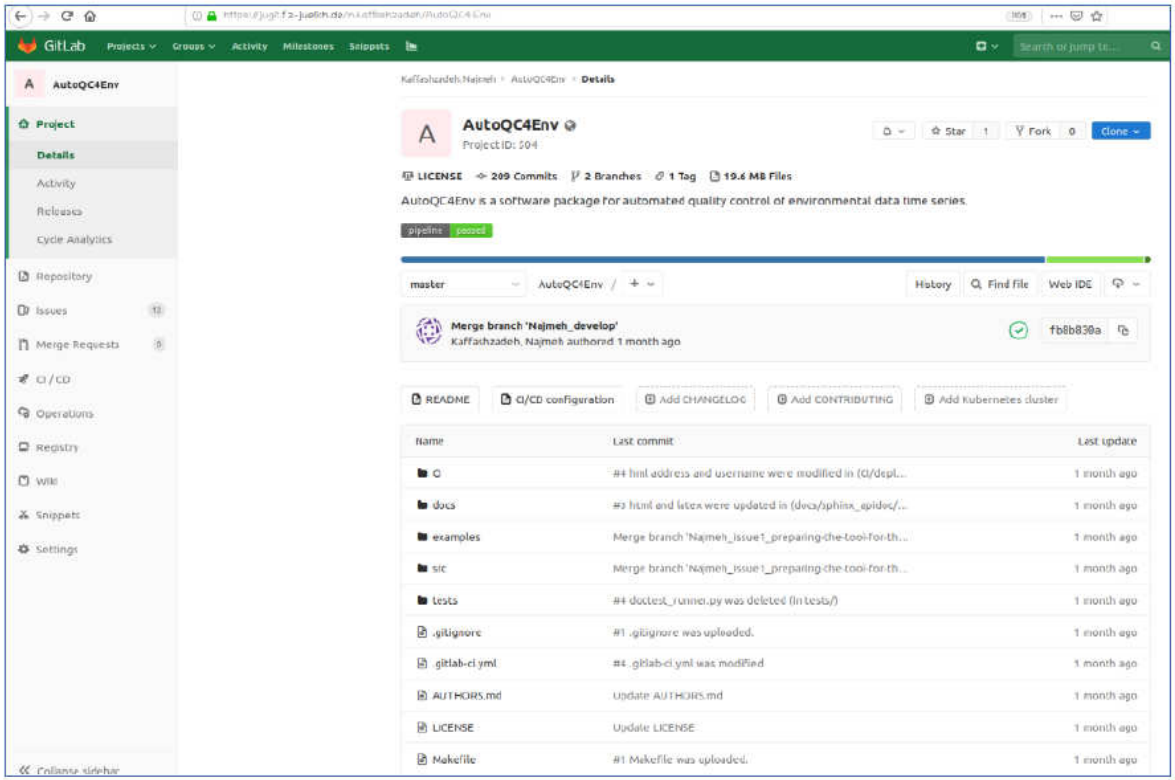


Figure: A screenshot of the repo at the
`https://jugit.fz-juelich.de/n.kaffashzadeh/AutoQC4Env`

Scan me

Appendix

QC Complexity in Env data

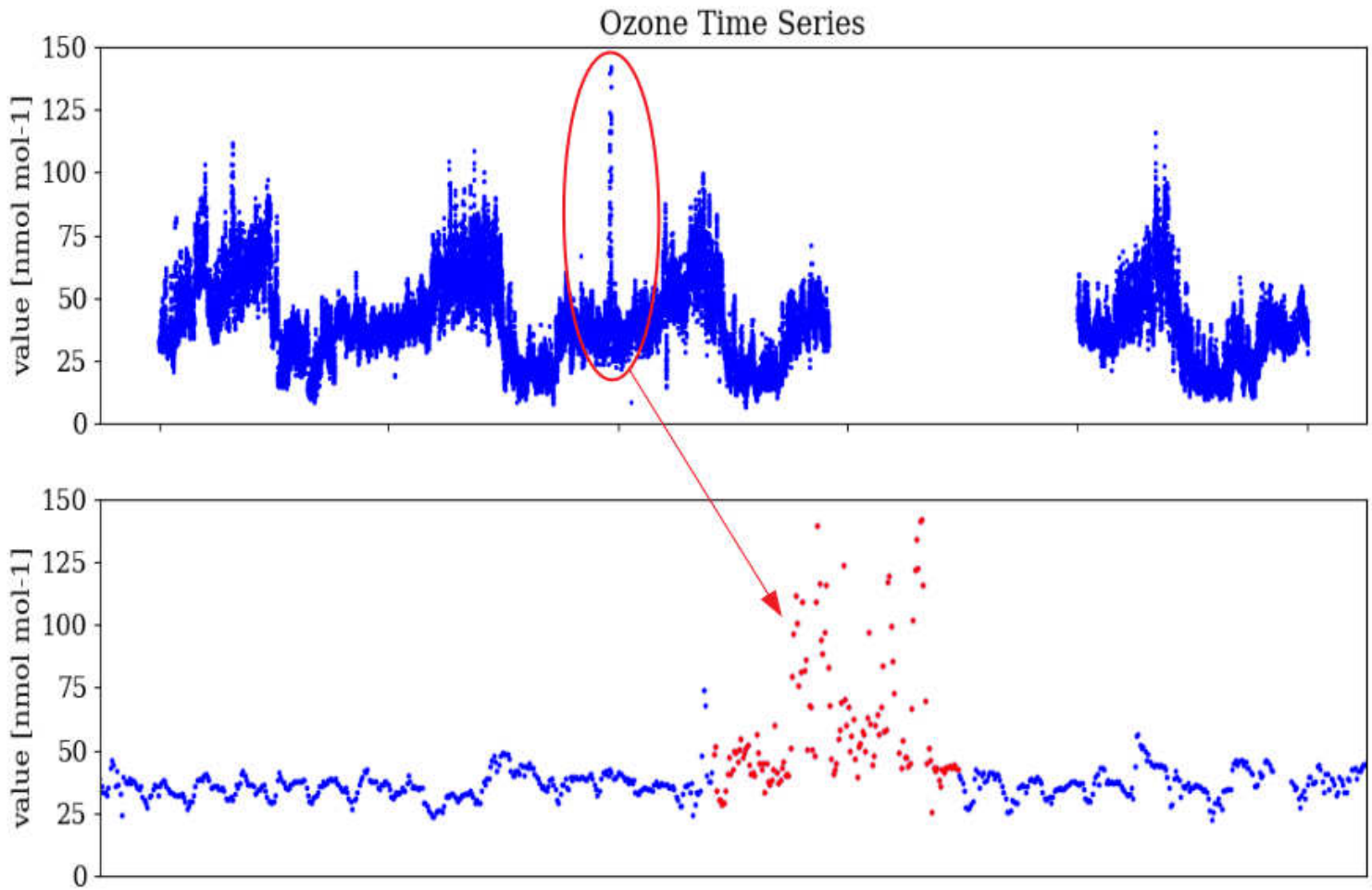


Figure: an example time series, derived from TOAR[1] database.

[1] Tropospheric Ozone Assessment Report

AutoQC4Env Performance