



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

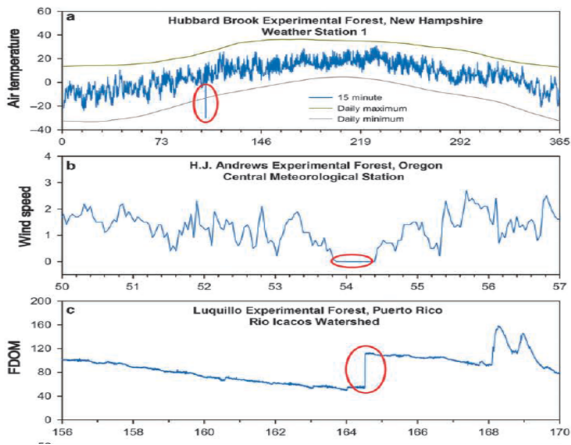


Figure: Campbell 2013

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		> 150 Qualifier Code in EPA				
Qualifier Code	Qualifier Description	Qualifier Type	Qualifier Code	MSL Active	Legacy Code	
1	Deviation from a CPE/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		
IV	Data received 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	QMOC Unknown.	Quality Assurance Qualifier	QA	NO		
9	Negative value detected - zero reported.	Quality Assurance Qualifier	QA	YES		
A	High Winds.	Informational Code	INFORM	NO		
AA	Sample Processor out of Limits.	Null Data Qualifier	NULL	YES	9967	
AB	Turbidity Unavailable.	Null Data Qualifier	NULL	YES	9968	
AC	Construction/Repairs to Area.	Null Data Qualifier	NULL	YES	9969	

EBAS

Group 1: Valid data

Flag: Validity Description
 255: 1/ ... Data measurement

Group 2: Suspicion flag for suspect, irregular data

Flag: Validity Description
 255: 1/ ... Suspicion flag checked and accepted by data originator. Data measurement
 255: 1/ ... Before theoretical detection limit or normal (20: 20: 20), but it is clear that these conditions are violated and it is considered that

Group 3: Flags for aggregated datasets

Flag: Validity Description
 255: 1/ ... Data compliance with flag 255
 255: 1/ ... Data compliance with flag 255
 255: 1/ ... Data compliance with flag 255

Group 4: Chemical problem

Flag: Validity Description
 255: 1/ ... Responder intervention or non-compliance. Not considered as
 255: 1/ ... Responder intervention or non-compliance

Group 5: Missing flag

Flag: Validity Description
 255: 1/ ... Missing measurement, compliance issue

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

Figure: QC flags from various environmental agencies

BACKGROUND

Requirements for environmental QC software

- Many heterogeneous datasets with different sampling times, statistical distributions, etc. → 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*.

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

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

- 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 $1 - \text{CDF}^1$ from the statistical distribution as a *proxy*

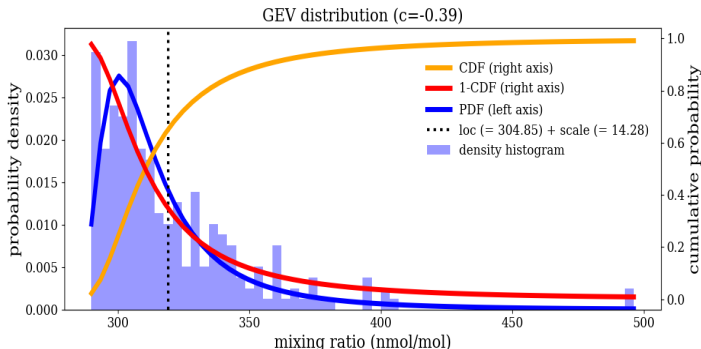


Figure: GEV² distribution derived from the 1000 largest ozone values measured after 1990 from the TOAR database.

A NEW METHODOLOGY

How to estimate the overall probability (P)?

Assuming non-independent tests:

$$P = \min(\text{prob}_t; t = 1, 2, \dots, n) \quad (3)$$

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)

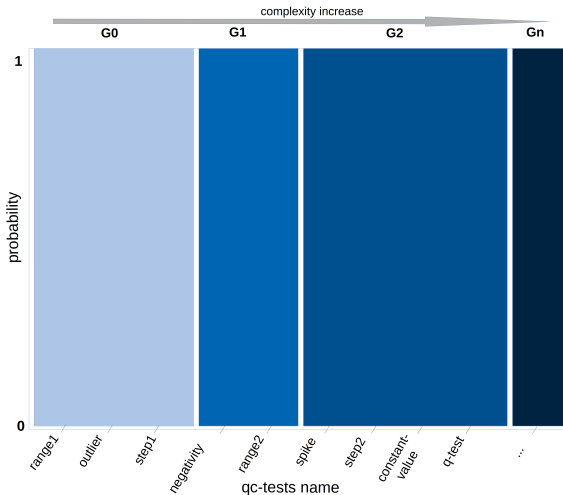
Group3: spatial consistency tests (statistical distributions)

Group4: internal consistency tests (correlation)

Group5: deep learning

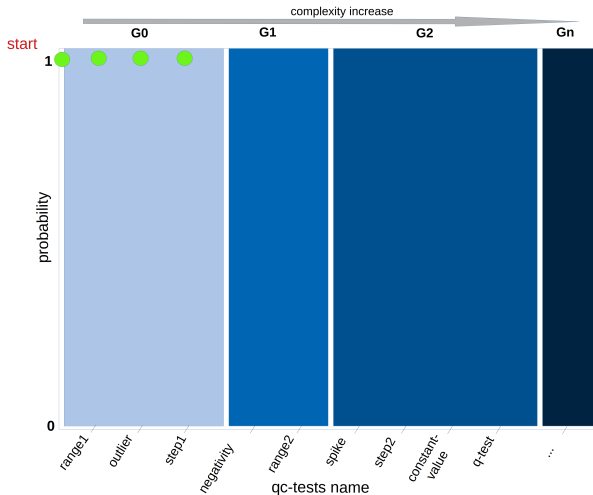
THE AutoQC4Env SOFTWARE FRAMEWORK

Implementation of the concept



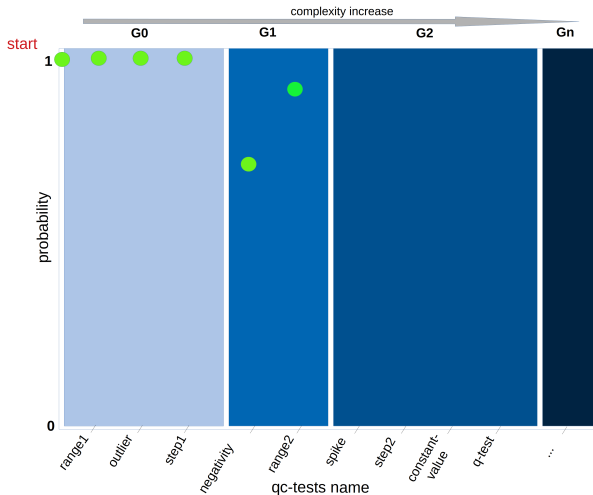
IMPLEMENTATION OF THE CONCEPT

Performing tests in G0



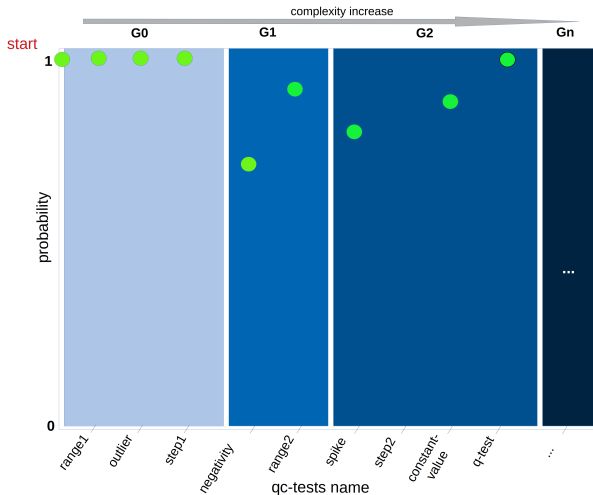
IMPLEMENTATION OF THE CONCEPT

Performing tests in G1



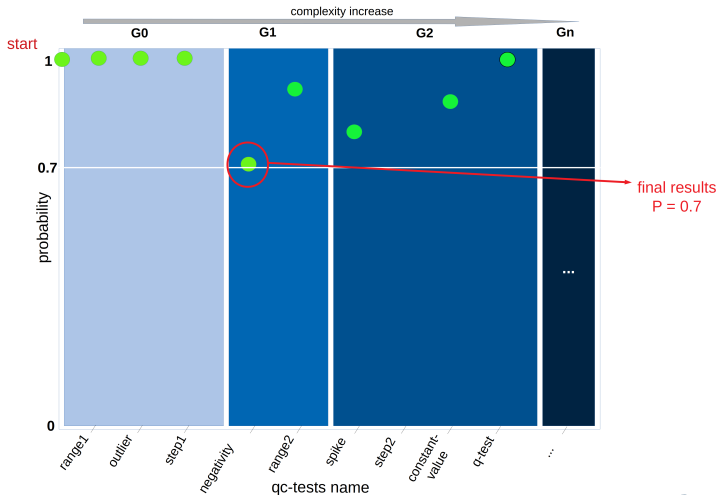
IMPLEMENTATION OF THE CONCEPT

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

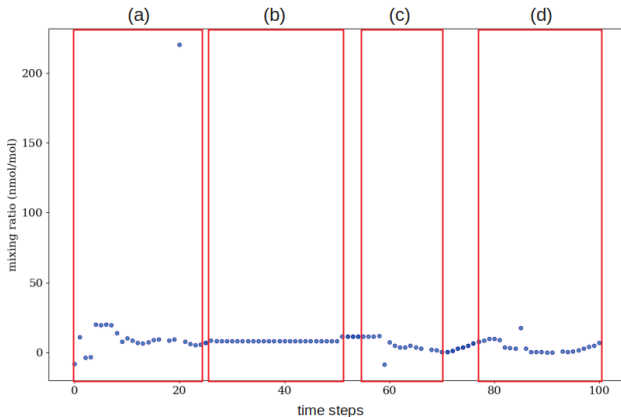


Figure: An arbitrarily selected ground-level ozone measurement series with typical error features

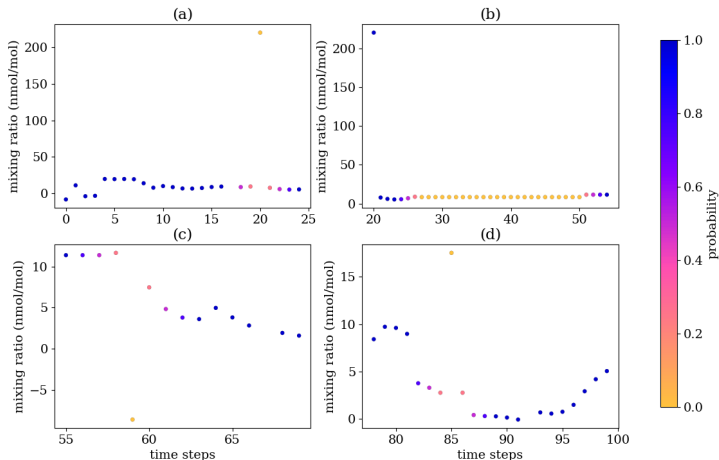
RESULTS OF A CASE STUDY

Sample user configuration settings

```
1 "RangeTest" : {  
2     "range_min" : -10,  
3     "range_max" : 150,  
4     "range_neighboring_size" : 3,  
5     "range_neighboring_side" : "both",  
6     "qc_group_name" : "g0"  
7 }
```

RESULTS OF A CASE STUDY

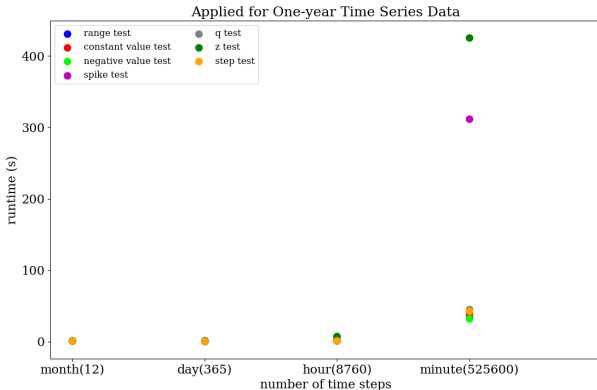
QC tests output



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

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 ESDE¹ group members on 12 June 2019



European Research Council
Established by the European Commission



How to access to the AutoQC4Env?

It is available in a jugit repository as:

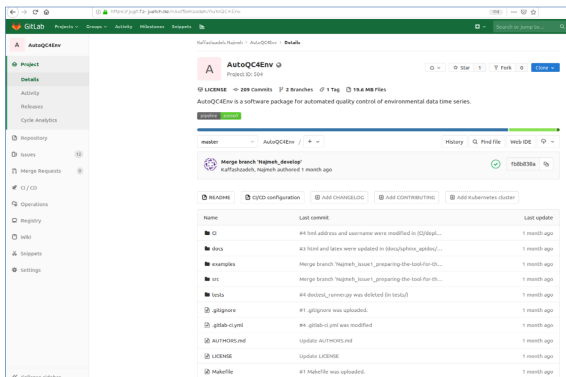


Figure: A screenshot of the repo at the <https://jugit.fz-juelich.de/n.kaffashzadeh/AutoQC4Env>

How to access to the AutoQC4Env?

It is available in a jugit repository as:

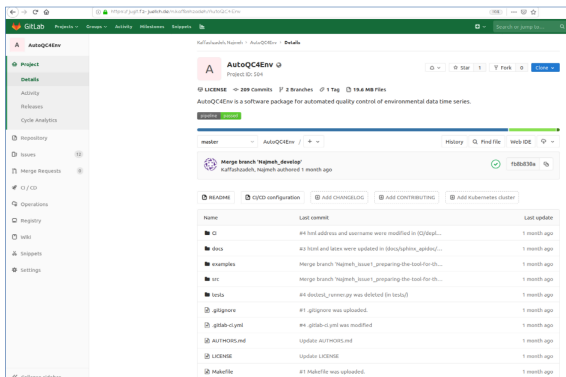


Figure: A screenshot of the repo at the <https://jugit.fz-juelich.de/n.kaffashzadeh/AutoQC4Env>



Appendix

QC Complexity in Env data

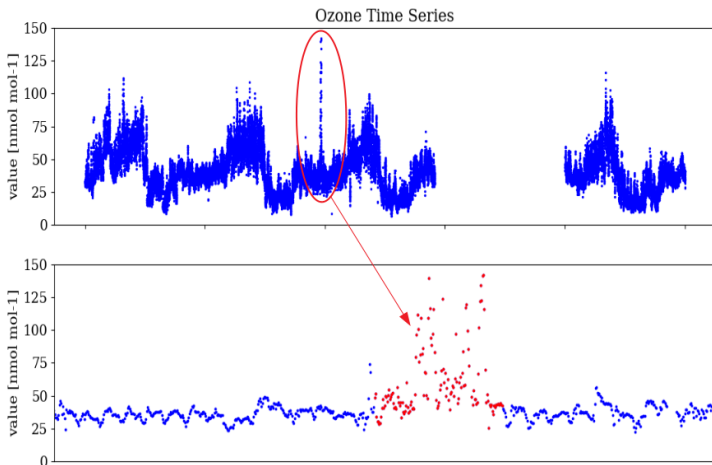


Figure: an example time series, derived from TOAR¹ database.

¹Tropospheric Ozone Assessment Report

AutoQC4Env Performance

