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

![Graphs showing air temperature, wind speed, and FDOM with red circles highlighting anomalies.](Campbell_2013.png)
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

**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.

If test $t$ is passed: $\prob_t = 1 - \min(p\text{-value}, 0.5)$  \hspace{1cm} (1)

If test $t$ fails: $\prob_t = 0 + \min(p\text{-value}, 0.5)$  \hspace{1cm} (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: 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:

\[ P = \min(probt; t = 1, 2, ..., n) \]  \hspace{1cm} (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!
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

Diagram showing a progression from G0 to Gn with different stages and tests.
IMPLEMENTATION OF THE CONCEPT

Performing tests in G1
IMPLEMENTATION OF THE CONCEPT

Performing tests in G2

complexity increase

start

G0
G1
G2
Gn

range1
outlier
step1
negativity
range2
spike
step2
constant-value
q-test

0
1

colour}

Member of the Helmholtz Association
July 22, 2019
Slide 12
IMPLEMENTATION OF THE CONCEPT

Estimating the final probability (P)

| qc-tests name | range1 | outlier | step1 | negativity | range2 | spike | step2 | constant-value | q-test |...
|---------------|--------|---------|-------|------------|--------|-------|-------|----------------|--------|-------
| probability   | 1      | 0.7     | 0.7   | 0.7        | 0.7    | 0.7   | 0.7   | 0.7            | 0.7    | 0.7  |
| complexity increase | G0 | G1 | G2 | Gn |

final results
P = 0.7
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"
}
```
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 ESDE\textsuperscript{1} group members on 12 June 2019

\textsuperscript{1}Earth System Data Exploration

Member of the Helmholtz Association
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
QC Complexity in Env data

Figure: an example time series, derived from TOAR\textsuperscript{1} database.

\textsuperscript{1}Tropospheric Ozone Assessment Report
AutoQC4Env Performance for One-year Time Series Data