EasyChair Preprint

№ 1325

A New Tool for Automated Quality Control of Environmental Data in Open Web Services

Najmeh Kaffashzadeh, Felix Kleinert and Martin G Schultz

July 22, 2019

A New Tool for Automated Quality Control of Environmental Time Series (AutoQC4Env) in Open Web Services

Najmeh Kaffashzadeh[1], Felix Kleinert[1], and Martin G. Schultz[1]

Jülich Supercomputing Centre, Forschungszentrum Jülich GmbH, Jülich, Germany
{n.kaffashzadeh, f.kleinert, m.schultz}@fz-juelich.de

Abstract. We report on the development of a new software tool (AutoQC4Env) for automated quality control (QC) of environmental time series data. Novel features of this tool include a flexible Python software architecture, which makes it easy for users to configure the sequence of tests as well as their statistical parameters, and a statistical concept to assign each value a probability of being a valid data point. There are many occasions when it is necessary to inspect the quality of environmental data sets, from first quality checks during real-time sampling and data transmission to assessing the quality and consistency of long-term monitoring data from measurement stations. Erroneous data can have a substantial impact on the statistical data analysis and, for example, lead to wrong estimates of trends. Existing QC workflows largely rely on individual investigator knowledge and have been constructed from practical considerations and with a least theoretical foundation. The statistical framework that is being developed in AutoQC4Env aims to complement traditional data quality assessments and provide environmental researchers with a tool that is easy to use but also based on current statistical knowledge.

Keywords: AutoQC4Env tool · Quality control · Environmental time series

1 Introduction

Environmental monitoring is an essential component in humanity's quest to protect Earth and mitigate adversarial effects of climate change, water, and soil pollution, and other transformations that are directly or indirectly caused by human activities. Moreover, environmental monitoring drives the development of advanced information technology. This is due to the exponential growth of monitoring data, the open data attitude in major parts of the environmental science communities and of many governmental agencies. Also, a high degree of standardization has been achieved with respect to geo-data and metadata. Users of environmental data services need to be able to assess the fitness-for-purpose of a data set and they need to find information that allows them to use the data set correctly. While numerous research institutions have implemented open

data services for environmental data and metadata, less has been achieved with respect to facilitating the assessment of the data quality.

Although various software for checking the quality of time series have been developed by several environmental agencies and research institutions in the past, most of these have a rather specific application focus and they are generally deeply embedded in specific data processing workflows and thus not fully transparent to the data users. Examples of relatively well documented QC procedures include the QA/QC of Real-Time Oceanographic Data (QARTOD) [2], Carbon in the Atlantic Ocean (CARINA) [3], National Ecological Observatory Network (NEON) [1,5], US EPA Air Quality System (AQS), and European Air Quality Database Airbase. A common element of these QC procedures is the use of data quality flags as classifiers of the measured data values. The level of detail of the flagging schemes varies greatly, and this makes it difficult to automatically process the quality information and support user decisions on which data shall be accepted for a given analysis purpose, in particular, if data from different sources should be merged [4]. To provide one example: US air quality data can be flagged as "not-to-be-used-for-attainment-purposes" in the event that wildfires pushed air pollutant concentrations above the regulatory values. Thus, for an official air quality reporting, those data must be excluded. However, if these data are used in the evaluation of numerical air quality models, the omission of such events leads to substantial bias.

Here, we introduce a tool (AutoQC4Env) with a flexible software framework that allows users to configure the QC tests according to their needs. Moreover, the tool introduces a novel concept to environmental time series analysis based on statistical measures of testing uncertainty. While the concept and software framework may also be useful in other domains, we focus on environmental time series in this study in order to keep the problem tractable.

This paper is structured as follows: the methodology is presented in Section 2; Section 3 describes the software framework and implementation of the concept; Section 4 shows a case study of the tool's application, and Section 5 contains a short summary.

2 Methodology

The task of automated QC tests is to detect *abnormal* values. Statistically, this implies either that an individual value lies outside the expected distribution for a given variable at a specific time and location, or that some properties of a group of values are inconsistent with expectations. Although such *errors* are labeled according to their visual appearance on a time series graph, e.g. "outlier", "constant values", and "data out of range", or the result of a single test is typically categorized to "pass" or "fail", various statistical tests allow estimation of the uncertainty of a test result, for example via a *p-value* or from the probability density function (PDF) of an extreme value distribution (Figure 1). We estimate such uncertainty by using them as proxies to obtain a probability that a given value is "valid" depending on the test outcome. For example:

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

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

Effectively, test results with low *p-value* (i.e. low uncertainty) provide relatively strong confidence that a value is either "valid" or "invalid" with respect to the test's properties. If the uncertainty of the test result is large, the probability of a value being "valid" will approach 0.5 to indicate indifference.

It is important to note that we do not use the *p-value* as a significance test, but

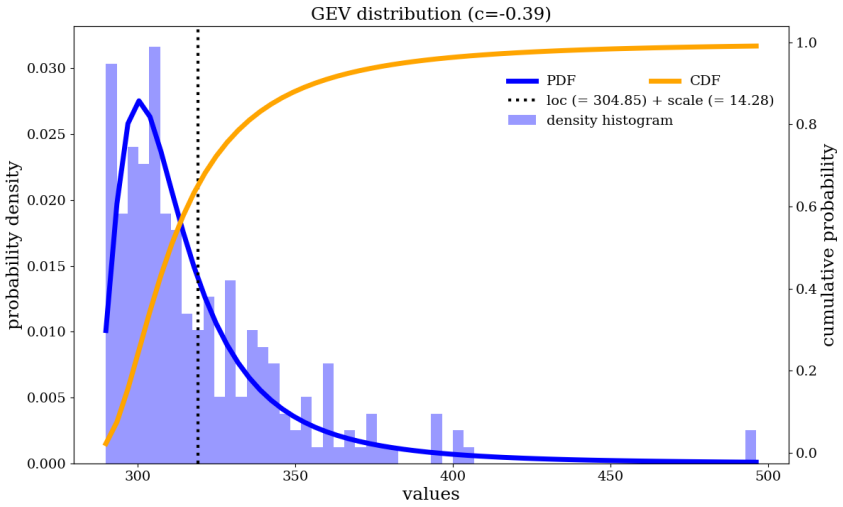

Fig. 1. Generalized extreme value (GEV) distribution and cumulative density function (CDF) derived from the 1000 largest ozone values measured after 1990 from the Tropospheric Ozone Assessment Report (TOAR) database [4]. The quantity $1-CDF$ is used as a proxy for the validity of ozone measurements in a one-sided extreme value test. The c, loc, and scale show shape, location and scale parameters of the GEV distribution, respectively.

only as a proxy. The overall likelihood of a value's validity is obtained by combining all the QC tests result. As the tests might not be independent, the final probability P approximates the conditional probability by using the minimum probability of all individual test result:

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

where n is the number of QC tests that have been performed. If all tests are passed, the P will lie between 0.5 and 1. Conversely, if at least one test fails,

the P will be between 0 and 0.5. Then the user can map the P into categorical flags. Thresholds can be defined according to the intended analysis, consequently balancing requirements for good quality data and sample size.
In closing this section, we emphasize that the statistical testing can only detect aberrations from expected data patterns and therefore does not constitute a judgment about the quality of a data value *per se*. Even though exceptional values often indicate some problem with the measurements or the data processing, they can also arise from exceptional measurement conditions, i.e. sampling of rare events. Therefore, a value with a low probability is not necessarily invalid, but may occasionally also point to an event of special significance. Detecting such events automatically will, if at all possible, require additional analyses with independent data or metadata. In future versions of AutoQC4Env we therefore plan to add consistency tests among multiple variables and multiple measurement locations as well as workflows to analyze numerical model results or metadata information.

3 AutoQC4Env Software Framework

The AutoQC4Env software implements a flexible chain of statistical QC tests with easily configurable settings, which are read from the *JSON* files. The tests are categorized into five groups (G0 - G4) with increasing test complexity (Table 1). The users are able to choose which tests shall be executed within each group and they can specify test parameters depending on their data set. All statistical tests are implemented as Python classes and they are derived from two base classes, which modify the probability and calculate the statistics, respectively. It is therefore easy to extend the tool or to modify existing test procedures. We note that the software development is work in progress and only a few tests have been fully implemented. Given the potentially huge amount of environmental data that may stream into a data processing center, it is important to follow a stateless concept so that several instances of the tool can be run in parallel.

Table 1. Definition of test groups in the AutoQC4Env tool framework and implementation status of specific tests.

Group label	Scope of the group	Available QC tests
G0	exclude the very gross errors for subsequent analysis (*sanity check*)	range1 test outlier test
G1	check a single value quality	negative value test, range2 test
G2	check the quality of a single value with adjacent data points	spike test, step test, q test, constant value test
G3	check consistency across multiple variables measured at the same site	not yet implemented
G4	check spatial consistency across nearby stations	not yet implemented

The current alpha version of the tool allows for a limited parallelization in that only full test suites can be run simultaneously. The software will be distributed via a git repository. The package includes automated documentation (Sphinx), unit-testing, and example applications. Functionality testing and the definition of new features are being conducted in the Digital Earth initiative of the German Helmholtz Association.

4 Case Study Ground-level Ozone Time Series

To demonstrate typical errors in ground-level ozone measurement time series and their *flagging* by the AutoQC4Env tool, we selected four time series of hourly ozone data from an arbitrarily data set of the TOAR database. Since these data had been already quality controlled (see [4]), we added some typical features of raw data sets for this demonstration. Figure 2 includes the probability from a run of the AutoQC4Env tool. The color shading clearly shows that the statistical tests capture several data artifacts and assign low probabilities to these values. Moreover, these figures highlight another feature of the AutoQC4Env tool: lower probabilities are not only assigned to the offending data values, but also to several neighbouring points. This reflects a common experience that certain data errors are usually indicative of a perturbation in a larger sub-sample of the data. Details about this will be provided in a subsequent paper.

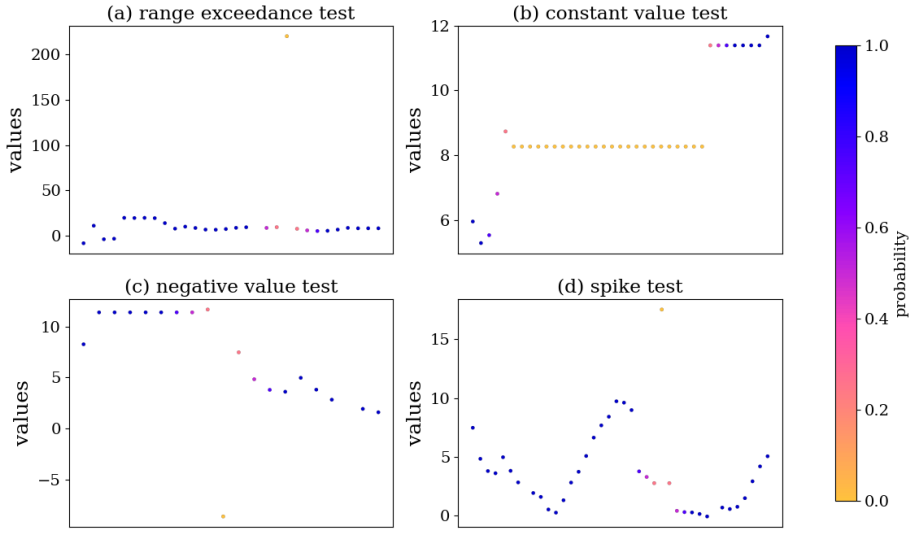

Fig. 2. Demonstration of typical environmental time series errors and their detection by the AutoQC4Env tool. The four panels show individual sub-samples of an arbitrarily selected ground-level ozone measurement series with added error features.

5 Conclusion

The AutoQC4Env tool is still at an early development stage, but it has already attracted the attention of several researchers, because of its modern code design and its novel concept to estimate probabilities for the validity of measured data points. While existing QC tools have mostly been developed for specific data sets and applications, the AutoQC4Env provides a generic and self-consistent software framework, which can easily be adapted to specific user needs. While the tool may also work for time series data from other domains, we are currently concentrating on environmental measurement time series, which often show some commonalities, such as auto-correlation or the fact that a large fraction of the variability arises from more or less regular cycles, e.g. diurnal or seasonal. The AutoQC4Env combines a theoretical concept based on statistical test results with a flexible software toolkit that can easily be adapted to different workflows and user needs. By adapting the test sequences and parameters, the tool can be used during various stages of environmental data management: (i) initial checks in a real-time data transmission system, (ii) data review before ingestion into a database, and (iii) use case-specific data selection procedures, for example as part of open web services.
The present development status of the tool is a demonstrator version at alpha stage, where most of the statistical tests still need to be implemented and "calibrated" with respect to their uncertainty measures.

Acknowledgements: This work has been performed and funded as part of the IntelliAQ project under ERC-2017-ADG#787576 grant at the Jülich Super Computing Centre, Forschungszentrum Jülich. The TOAR community and various national environmental agencies are gratefully acknowledged for providing data and collaborating on the development of the TOAR database. Sabine Schröder and Lukas Leufen helped with the data analysis and software infrastructure.

References

1. Durre, I., Menne, M.J., Vose, R.S.: Strategies for evaluating quality assurance procedures. Journal of Applied Meteorology and Climatology **47**(6), 1785–1791 (2008). https://doi.org/10.1175/2007JAMC1706.1
2. IOOS-US: U.S. Integrated Ocean Observing System: A Blueprint for Full Capability. Version 1.0. Tech. Rep. November (2010)
3. Key, R.M., Schirnick, C., Velo, A., Tanhua, T., van Heuven, S., Olsen, A.: Quality control procedures and methods of the CARINA database. Earth System Science Data **2**(1), 35–49 (2010). https://doi.org/10.5194/essd-2-35-2010
4. Schultz, Martin G., e.a.: Tropospheric Ozone Assessment Report: Database and Metrics Data of Global Surface Ozone Observations. Elem Sci Anth **5**(0), 58 (2017). https://doi.org/10.1525/elementa.244
5. Taylor, J.R., Loescher, H.L.: Automated quality control methods for sensor data: a novel observatory approach. Biogeosciences **10**(7), 4957–4971 (2013). https://doi.org/10.5194/bg-10-4957-2013