EGU2020-13357
https://doi.org/10.5194/egusphere-egu2020-13357
EGU General Assembly 2020

A Statistical Model for Automated Quality Assessment of the TOAR-II

Najmeh Kaffashzadeh[1], Kai-Lan Chang[2], Sabine Schröder[1], and Martin G. Schultz[1]
[1]Forschungszentrum Jülich GmbH, Jülich Supercomputing Center, Jülich, Germany
[2]NOAA Earth System Research Laboratory, Boulder, CO, USA

The Tropospheric Ozone Assessment Report, phase 2, (TOAR-II) database is a collection of global ground-level ozone in-situ measurements from various locations. It also holds data of selected ozone precursors and meteorological variables. TOAR-II assembles air quality data from many different sources and thus requires a common data quality assessment (QA) to ensure the data meet the quality required for globally consistent analyses. The large volume of this database (more than 100,000 data series) enforces the use of automated, data-driven QA procedures.

Accordingly, we have developed a statistical model for automated QA. This model consists of several statistical tests that are classified into several sub-groups. In this model, a QA-score (an indicator ranging from 0 to 1) was assigned to each individual data point to estimates the value's plausibility. The foundation of this concept is statistical hypothesis testing and the probability theory. This model was implemented in a Python package and is called AutoQA4Env.

One application of AutoQA4Env is the data ingestion workflow of TOAR-II. The tool generates a data quality report which is then sent back to the data provider for inspection. Since AutoQA4Env is easily configurable, it allows the users to set quality thresholds and thus filter data according to their use case. While we primarily develop AutoQA4Env for air quality data, the same concept and model might be applicable to other databases and the software framework is flexible enough to allow for other use cases.