Geophysical Research Abstracts
Vol. 21, EGU2019-14392, 2019
EGU General Assembly 2019

A novel concept for Automated Quality Control of Atmospheric Time Series

Najmeh Kaffashzadeh, Sabine Schröder, and Martin G. Schultz
Forschungszentrum Jülich, Jülich Supercomputing Center, Jülich, Germany (n.kaffashzadeh@fz-juelich.de)

Measurements of atmospheric physical and chemical parameters are essential for atmospheric model evaluation, trend analysis, climate prediction, and other applications. Particularly when the time series from various measurement instruments or data providers are merged together, assessing the quality of the data presents a major challenge and often relies on subjective screening. The quality of the time series can be affected by several error types, such as random error, systematic error due to calibration errors, and gross error from malfunctioning instruments, or data processing errors, such as mistyped values and improper date-time formats. Some of these errors may have a considerable impact on the statistical analysis of the time series. Thus, identifying the quality of the data, i.e. quality control (QC), is an essential step for any data analysis.
Here, we present a software package for the automated QC of the atmospheric time series based on the use of several algorithms that are in use at various environmental agencies and research initiatives. The tool can either be embedded in automated workflows to process real-time data or be applied to a second-level analysis of archived multi-year data. Several statistical tests are grouped in categories with increasing complexity. Any number of tests can be defined and run sequentially. The set of statistical tests and any user arguments can easily be configured with variable-specific control files in the JSON format. This allows for easy integration into an automated workflow software and distributed data processing services.
For expressing the quality of a measured data series, we introduced a probability concept which assigns each value a likelihood of being "good" data. Here, "good" is interpreted in a statistical sense as belonging to an expected probability distribution. Some of the tests influence not only the probability of a single point but may also impact on the probability of its neighboring points.
We tested the software with multi-annual hourly ozone and temperature data from the database of the Tropospheric Ozone Assessment Report (TOAR). Preliminary results indicate that the concept works well and is able to deal with a large and heterogeneous dataset such as the global collection of ozone data in the TOAR database.