EGU21-14485
https://doi.org/10.5194/egusphere-egu21-14485
EGU General Assembly 2021

Supervised and unsupervised machine-learning for automated quality control of environmental sensor data

Julius Polz[1], Lennart Schmidt[3], Luca Glawion[1], Maximilian Graf[1], Christian Werner[1], Christian Chwala[1,2], Hannes Mollenhauer[3], Corinna Rebmann[4], Harald Kunstmann[1,2], and Jan Bumberger[3]

[1]Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Campus Alpin, Garmisch-Partenkirchen, Germany

[2]Institute of Geography, University of Augsburg, Augsburg, Germany

[3]Department of Monitoring and Exploration Technologies, Helmholtz-Centre for Environmental Research (UFZ), Leipzig, Germany

[4]Department of Computational Hydrosystems, Helmholtz-Centre for Environmental Research (UFZ), Leipzig, Germany

We can observe a global decrease of well maintained weather stations by meteorological services and governmental institutes. At the same time, environmental sensor data is increasing through the use of opportunistic or remote sensing approaches. Overall, the trend for environmental sensor networks is strongly going towards automated routines, especially for quality-control (QC) to provide usable data in near real-time. A common QC scenario is that data is being flagged manually using expert knowledge and visual inspection by humans. To reduce this tedious process and to enable near-real time data provision, machine-learning (ML) algorithms exhibit a high potential as they can be designed to imitate the experts actions.

Here we address these three common challenges when applying ML for QC: 1) Robustness to missing values in the input data. 2) Availability of training data, i.e. manual quality flags that mark erroneous data points. And 3) Generalization of the model regarding non-stationary behavior of one experimental system or changes in the experimental setup when applied to a different study area. We approach the QC problem and the related issues both as a supervised and an unsupervised learning problem using deep neural networks on the one hand and dimensionality reduction combined with clustering algorithms on the other.

We compare the different ML algorithms on two time-series datasets to test their applicability across scales and domains. One dataset consists of signal levels of 4000 commercial microwave links distributed all over Germany that can be used to monitor precipitation. The second dataset contains time-series of soil moisture and temperature from 120 sensors deployed at a small-scale measurement plot at the TERENO site “Hohes Holz”.

First results show that supervised ML provides an optimized performance for QC for an experimental system not subject to change and at the cost of a laborious preparation of the training data. The unsupervised approach is also able to separate valid from erroneous data at reasonable accuracy. However, it provides the additional benefit that it does not require manual

flags and can thus be retrained more easily in case the system is subject to significant changes.

In this presentation, we discuss the performance, advantages and drawbacks of the proposed ML routines to tackle the aforementioned challenges. Thus, we aim to provide a starting point for researchers in the promising field of ML application for automated QC of environmental sensor data.