EGU2020-20777
https://doi.org/10.5194/egusphere-egu2020-20777
EGU General Assembly 2020

On the potential and challenges of using machine-learning for automated quality control of environmental sensor data

Lennart Schmidt[1], Hannes Mollenhauer[1], Corinna Rebmann[2], David Schäfer[1], Antje Claussnitzer[3], Thomas Schartner[3], and Jan Bumberger[1]

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

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

[3]Deutscher Wetterdienst, Offenbach, Germany

With more and more data being gathered from environmental sensor networks, the importance of automated quality-control (QC) routines to provide usable data in near-real time is becoming increasingly apparent. Machine-learning (ML) algorithms exhibit a high potential to this respect as they are able to exploit the spatio-temporal relation of multiple sensors to identify anomalies while allowing for non-linear functional relations in the data. In this study, we evaluate the potential of ML for automated QC on two spatio-temporal datasets at different spatial scales: One is a dataset of atmospheric variables at 53 stations across Northern Germany. The second dataset contains timeseries of soil moisture and temperature at 40 sensors at a small-scale measurement plot.

Furthermore, we investigate strategies to tackle three challenges that are commonly present when applying ML for QC: 1) As sensors might drop out, the ML models have to be designed to be robust against missing values in the input data. We address this by comparing different data imputation methods, coupled with a binary representation of whether a value is missing or not. 2) Quality flags that mark erroneous data points to serve as ground truth for model training might not be available. And 3) There is no guarantee that the system under study is stationary, which might render the outputs of a trained model useless in the future. To address 2) and 3), we frame the problem both as a supervised and unsupervised learning problem. Here, the use of unsupervised ML-models can be beneficial as they do not require ground truth data and can thus be retrained more easily should the system be subject to significant changes. In this presentation, we discuss the performance, advantages and drawbacks of the proposed strategies to tackle the aforementioned challenges. Thus, we provide a starting point for researchers in the largely untouched field of ML application for automated quality control of environmental sensor data.