

autoQC: An AI based online app for ocean data quality control

Thursday 4 April 2024 13:50 (15 minutes)

Marine data quality control (QC) is crucial to provide robust data products for climate analyses, monitoring, process- and model studies and much more. However, the QC of marine measurements of e.g. temperature, salinity, nutrients (phosphate, nitrate, ...), oxygen etc. is challenging. Measurements are prone to errors due to external forcing (sun, wind, currents, ...), internal variability (e.g. extremes), biogeochemical processes, instrument errors or failures and more. Ocean data QC is an international effort and large marine data infrastructures, like SeaDataNet (<https://www.seadatanet.org/>), EMODnet Chemistry (<https://emodnet.ec.europa.eu/en/chemistry>), Argo (<http://www.argodatamgt.org/>) or IQuOD (<https://www.iquod.org/>) have created sophisticated QC processing schemes. Typically, ocean data QC is a semi automatic process, whereas the ocean experts use algorithms to identify potentially “bad” data, which are accordingly often visually inspected to make a final decision and to give the data sample a quality flag, i.e. an indicator such as “good”, “bad”, “probably bad”, etc. One widely used tool for the QC is the Ocean Data View (ODV, <https://odv.awi.de>) software, which is also available as the online version webODV (<https://webodv.awi.de>). Because of the diverse nature of errors in the data, fully automated QC without expert visual checks is still less skillful and yields to too many misclassifications. However, visual QC is highly time demanding and skillful algorithmic support is needed, especially with the increase of fully automated sensors like the Argo buoys, where currently more than 3.800 buoys are drifting through the oceans and producing immense amounts of data.

To support the visual QC on marine data we have trained a deep neural network with the knowledge of an ocean QC data expert to mimic the human visual QC. The training of the ML algorithm is based on arctic ocean temperature data from UDASH (Unified Database for Arctic and Subarctic Hydrography). The ML algorithm improves the results of the classical checks significantly, hence increasing the data quality and reducing the experts workload.

For user friendly and easy access we have developed an online app at <https://mvre.autoqc.cloud.awi.de/>, where users can upload their data and let the data be quality controlled on our servers. The app provides detailed documentation and processed data are exported as simple .csv files or ODV Spreadsheet, which can be used directly in ODV or webODV (<https://hifis.webodv.cloud.awi.de>). The algorithm is written in Python (using Keras and Sklearn) and we provide two GitHub repositories, one which includes the sources of the algorithm, which can be used for further research or for training on other datasets. The other repository includes the fully trained model and provides an easy way to include it into other processing environments.

Currently the algorithm is limited to arctic temperature data and to two types of errors in the data, the so-called “Spikes” and “Suspect Gradients”. Next planned steps are to include salinity as well as another important error type named “Statistical Screening”.

Primary author: MIERUCH-SCHNUELLE, Sebastian (AWI)

Co-authors: Mr KREPS, Gastón (AWI); Dr CHOUAI, Mohamed (AWI)

Presenter: MIERUCH-SCHNUELLE, Sebastian (AWI)

Session Classification: Session 1: Capacity Building and AI

Track Classification: Capacity building and AI