Contribution ID: **7** Type: **Oral presentation**

AtmoRep: Large Scale Representation Learning for Atmospheric Data

Tuesday, 7 March 2023 10:20 (40 minutes)

The AtmoRep project asks if one can train one neural network that represents and describes all atmospheric dynamics. AtmoRep's ambition is hence to demonstrate that the concept of large-scale representation learning, whose principle feasibility and potential was established by large language models such as GPT-3, is also applicable to scientific data and in particular to atmospheric dynamics. The project is enabled by the large amounts of atmospheric observations that have been made in the past as well as advances on neural network architectures and self-supervised learning that allow for effective training on petabytes of data. Eventually, we aim to train on all of the ERA5 reanalysis and, furthermore, fine tune on observational data such as satellite measurements to move beyond the limits of reanalyses.

We will provide an overview of the theoretical formulation of AtmoRep, of our transformer-based network architecture, and of the training protocol for self-supervised learning that allows for unlabelled data such as reanalyses, simulation outputs and observations to be used for training and re-fining the network. We will also present the performance of AtmoRep for applications such as downscaling and forecasting and, furthermore, demonstrate that AtmoRep has substantial zero-short skill, i.e., it is capable to perform well on tasks it was not trained for. Although not specifically designed for air quality forecasting and analysis, we will also explain why AtmoRep provides a powerful basis for it and how a pre-trained AtmoRep network can be adapted for the task with limited computational costs.

ML method

Transformer

Main air pollutant of interest

Other

Primary authors: LESSIG, Christian (Otto-von-Guericke-Universität Magdeburg); LUISE, Ilaria (OpenLab, CERN); SCHULTZ, Martin (JSC)

Presenter: LESSIG, Christian (Otto-von-Guericke-Universität Magdeburg)

Track Classification: Machine learning core