Data Assimilation & Machine Learning
Presentation: October 24

Machine Learning methodology for generating ensemble members in Data Assimilation of Earth Observations

Alessandro D'Ausilio (Arianet/SUEZ, researcher), Giorgia De Moliner (Politecnico di Milano, PhD cand), Camillo Silibello (Arianet/SUEZ, researcher), Giovanni Lonati (Politecnico di Milano, Professor)

Abstract

In this study, we present a methodology to generating ensemble spread for atmospheric modeling by integrating Random Forest (RF), a machine learning (ML) technique, into the perturbation process. This method manages high-dimensional data and captures complex nonlinear relationships through RF, aiming to produce multiple, realistic 3D concentration fields. The goal is to obtain a spread between ensemble members that provides a realistic estimate of the model uncertainties, enabling operational Data Assimilation (DA) without needing to run the model multiple times.

The methodology consists of four main steps. Initially, multiple inference datasets are constructed by perturbing emissions, meteorological conditions, and state vectors. A training dataset is then generated for each model level by clustering spatial grid cells with similar time series characteristics for the target variable, using PCA/K-means, reducing dimensionality while preserving essential patterns. Subsequently, multiple RF models are trained to predict the concentration of interest using the clustered time series data as inputs. These trained models are then employed to generate the perturbed 3D concentration fields.

This approach was tested for assimilating NO2 Sentinel-5p/TROPOMI observations in the FARM chemical transport model forecast system, QualeAria. DA was conducted using DART. EnAKF was used as an offline version of the EnOI method, under the hypothesis that uncertainties grow quickly, so the analysis step isn't used to update initial conditions for subsequent runs. The complete assimilation cycle will be further implemented and tested. Preliminary results show it can capture both diagonal and off-diagonal covariance matrix terms. This method has the potential to address the challenging computational costs of DA for air quality monitoring and forecasting by incorporating ML into ensemble generation. However, further work is needed to account for evolving model dynamics.

Contact information

Alessandro D'Ausilio (Arianet/SUEZ, researcher), Giorgia De Moliner (Politecnico di Milano, PhD cand), Camillo Silibello (Arianet/SUEZ, researcher), Giovanni Lonati (Politecnico di Milano, Professor)