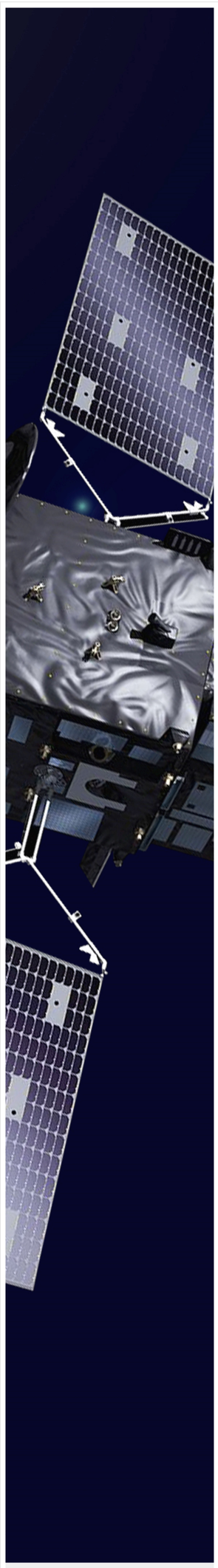
Bridging the Gap between Data and End-to-End ML4Weather and Climate Prediction

Proposed activities and open questions on EUMETSAT role supporting in observation-based ML weather forecasting

R. Tervo, J. Schulz, P. Ruti, P Albert, D. Puechmaille, V. John, R. Huckle, A. **Lattanzio**

[1] EUMETSAT, Germany - roope.tervo@eumetsat.int



WHICH DATA TO USE FOR END-TO-END NWP MODEL TRAINING?

- Best possible quality is important
- Homogenous time-series needed
- The data need to explain a high degree of spatiotemporal variance
- Data produced for global and regional re-analysis should be very good starting point

- What are the priorities in the data content?
- How long time-series should be for different activities?

HOW AND WHERE DATA HAS TO BE MADE AVAILABLE?

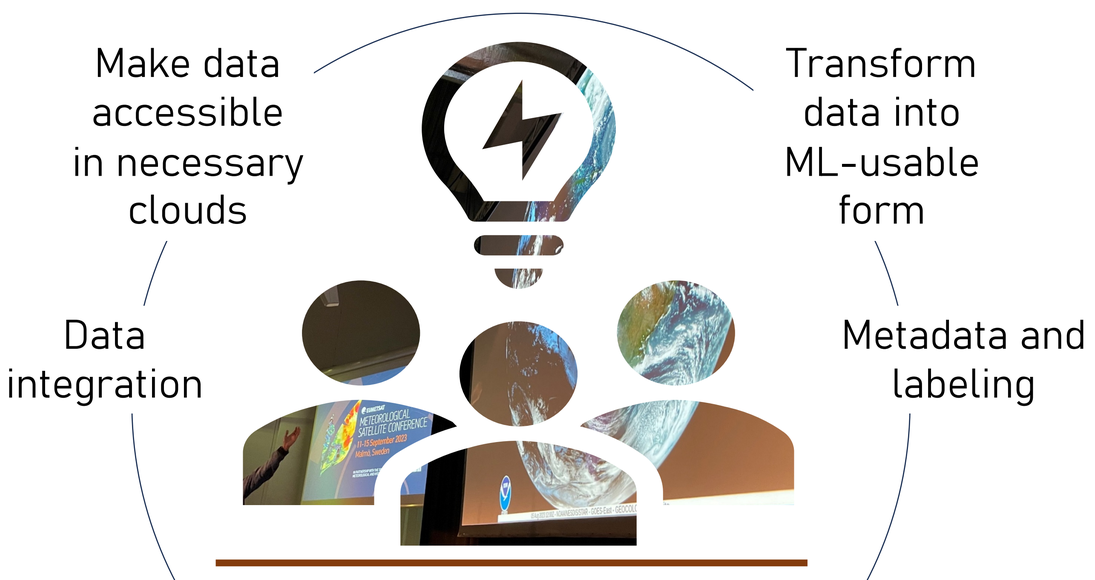
- Data must be accessible from various cloud platforms with sufficient performance
- Data proximate processing is needed for many use cases
- Many use cases need gridded data

- Can we find a common denominator across use cases that enables preparing data in terms of accessibility, structure, and format or do we have to modify on-demand?

KEY POSSIBILITIES FOR EUMETSAT OPERATIONS

- Supporting operationalization of the nowcasting algorithm is needed
- Several ML methods provides excellent basis for running feature detection for both NRT data and the whole EO archive
- ML provides possibilities in retrieval algorithms
- Large language models provides great potential in information retrieval

EUMETSAT HAS A CRUCIAL ROLE IN DATA CURATION



Workshop with EUMETSAT member states was organized to discuss ML/AI possibilities and EUMETSAT contribution to the development.

KEY ACTIVITY LINES

USER ENGAGEMENT

CLOUD BASED COLLABORATION

- Constructing robust training datasets initially based on satellite and radar data for ML forecast
- Ensuring and maintaining the high quality of these datasets
- Expanding training datasets with supplementary data
- Deploy currently available nowcasting models at European scale
- Evaluating existing and assess gaps for data sets suitable for assimilation in regional reanalyses
- Ensure satellite data is accessible through easy-to-use portals or APIs
- Providing infrastructure to allow MS to retrain ML-based Nowcasting tools currently available
- Tailoring data services: develop advanced functionalities
- Customizing data lake management to better serve ML needs

ML TRAINING DATA SETS

Tools

Data accessibility & infra

PILOTS

- Expand ML nowcasting using cloud environment for members
- Implement ML labelling, feature detection, to support forecasting for extremes
- GPU accessibility and performance for ML user's needs
- Utilize LLM for user assistance in exploring data
- Develop models for targeted satellite instrument retrievals