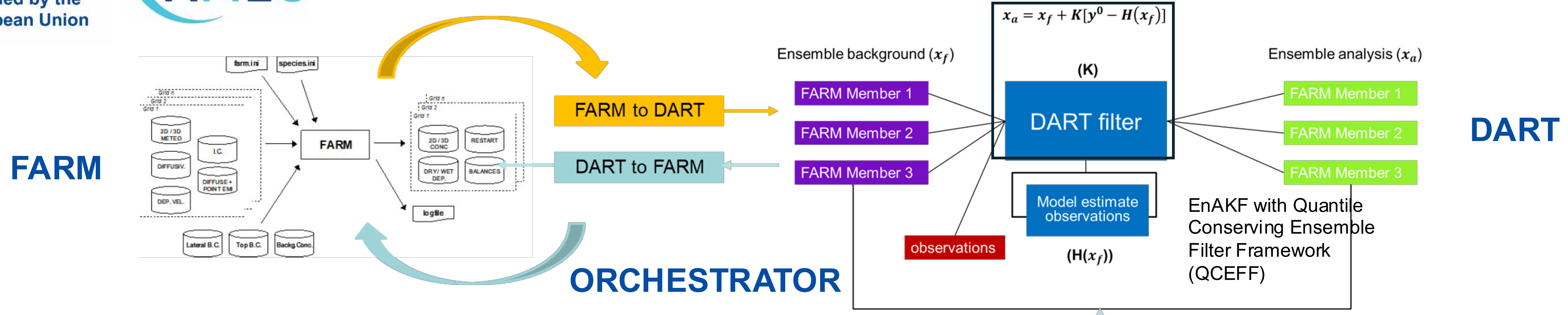
10th International Symposium on Data Assimilation
Kobe, Japan
October 21-25, 2024

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

Giorgia De Moliner (PhD Candidate, Politecnico di Milano), **Alessandro D'Ausilio** (ARIANET-SUEZ), **Camillo Silibello** (ARIANET-SUEZ), **Giovanni Lonati** (Professor, Politecnico di Milano)

Funded by the European Union



Localize the impacts of the variable only on the model variable state. In other words, NO_2 influences only itself.

NCAR | DART

Sat Observation in DART format

Input.nml
- controls behavior
DART executables

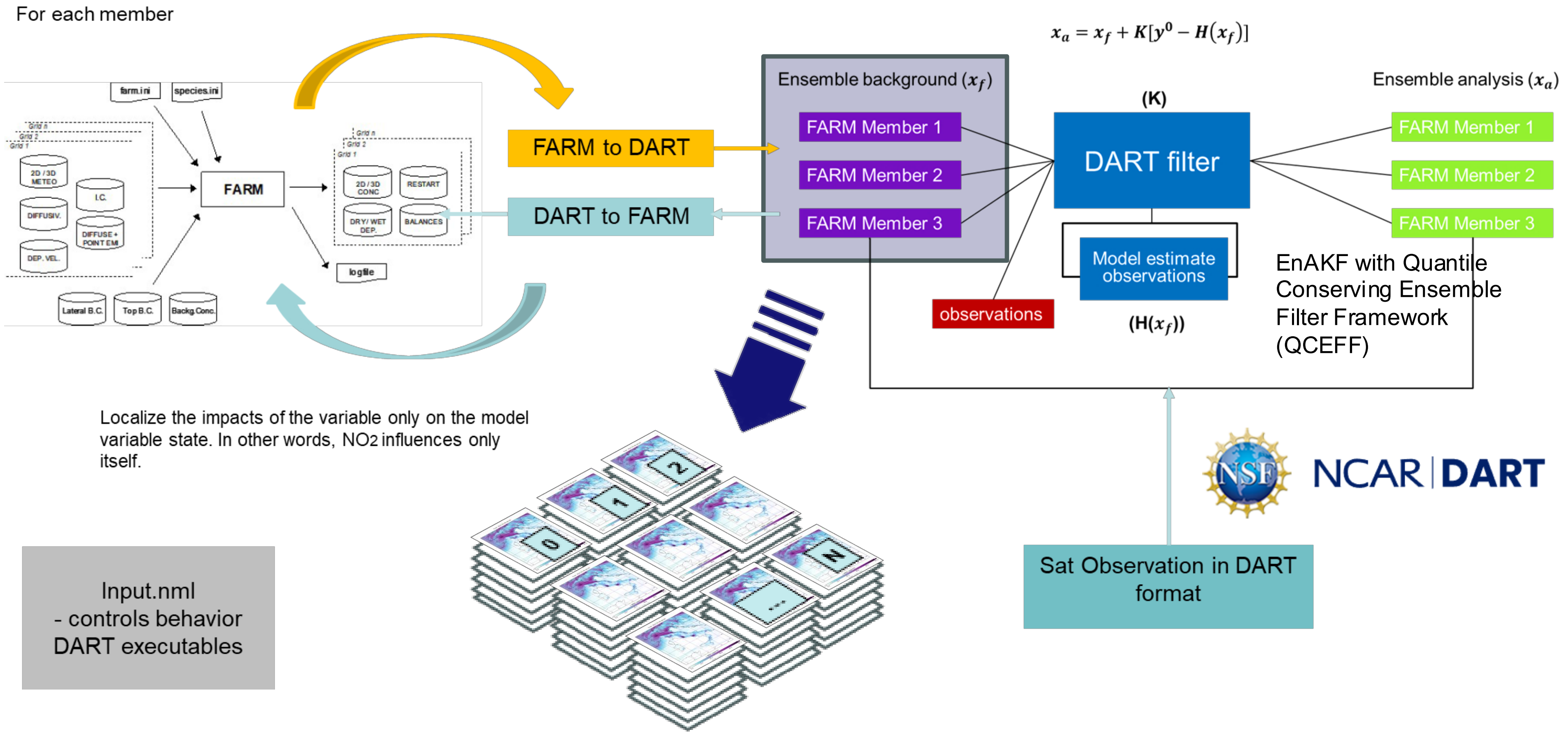
FARM-DART

https://www.cameo-project.eu/

Assessing the impacts of assimilating S5P-SO2-COBRA retrievals using FARM and DART

"The CAMEO project (grant agreement No101082125) is funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Commission. Neither the European Union nor the granting authority can be held responsible for them."

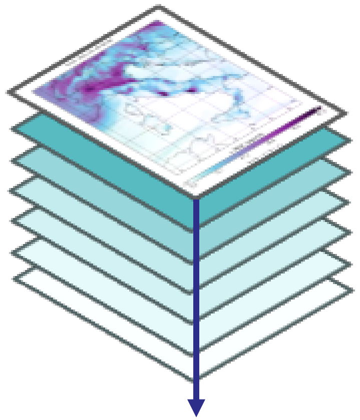
Problem Statement



Scope of the work

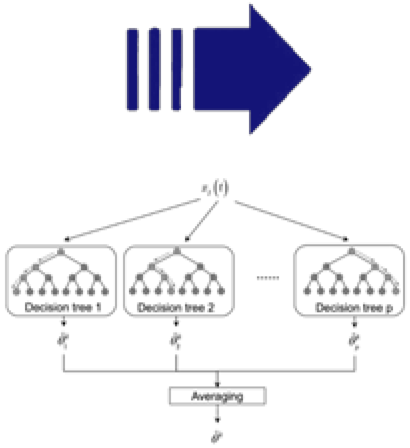
I. To generate a stationary ensemble of perturbated, realistic 3D concentration fields that provide a realistic estimate of the model uncertainties, without the need to run the model multiple times
II. To enable computationally-feasible, operational Data Assimilation (EnOI) of satellite data

One CTM run



Z-levels

ML



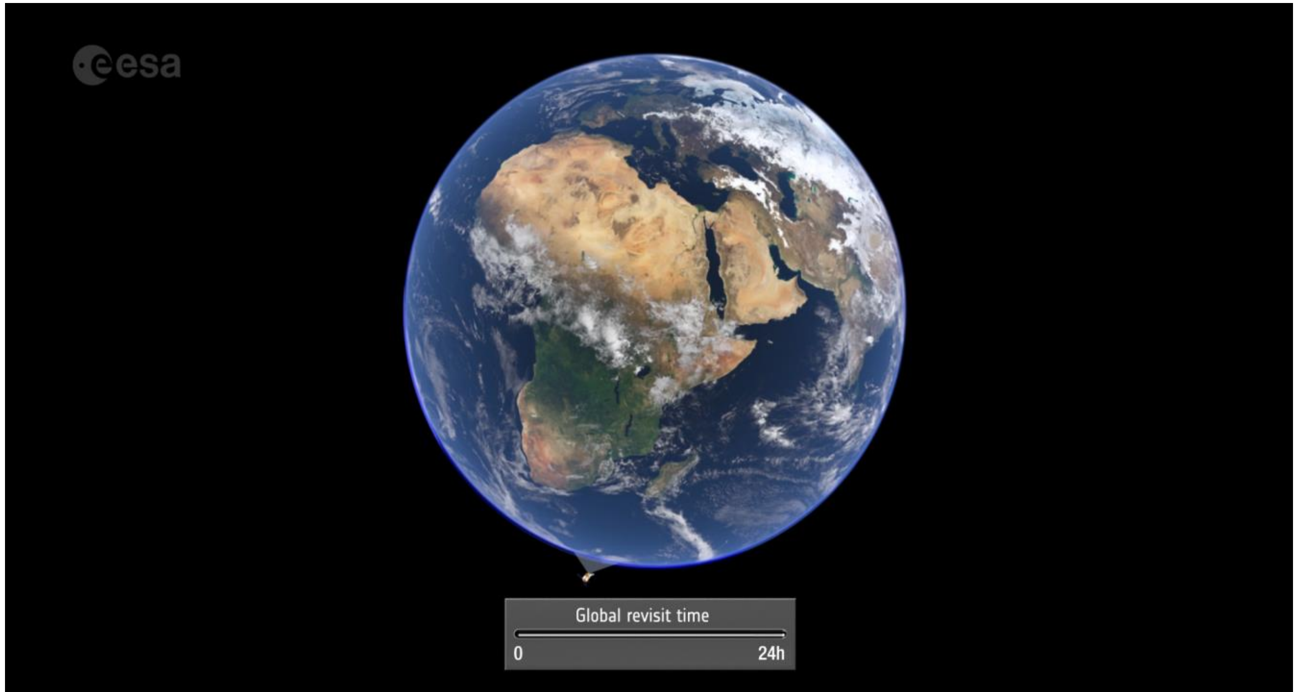
Ensemble

B $\begin{bmatrix} \sigma_{00}^2 & \sigma_{01}^2 & \sigma_{02}^2 \\ \sigma_{10}^2 & \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{20}^2 & \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix}$

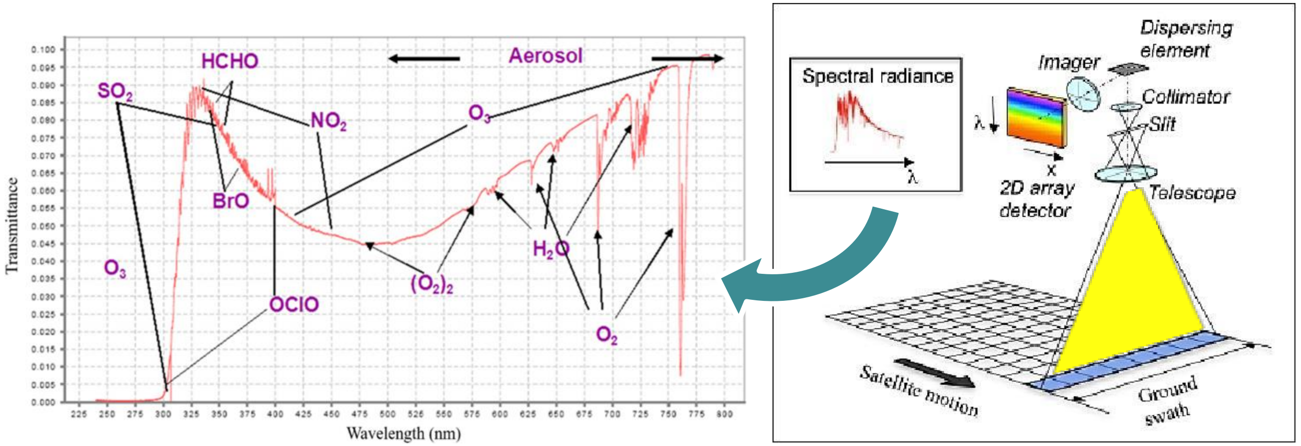
$\{\mathbf{0}, \dots, \boldsymbol{N}\}$

Obs: Sentinel-5p

Source: ESA



Source: ESA

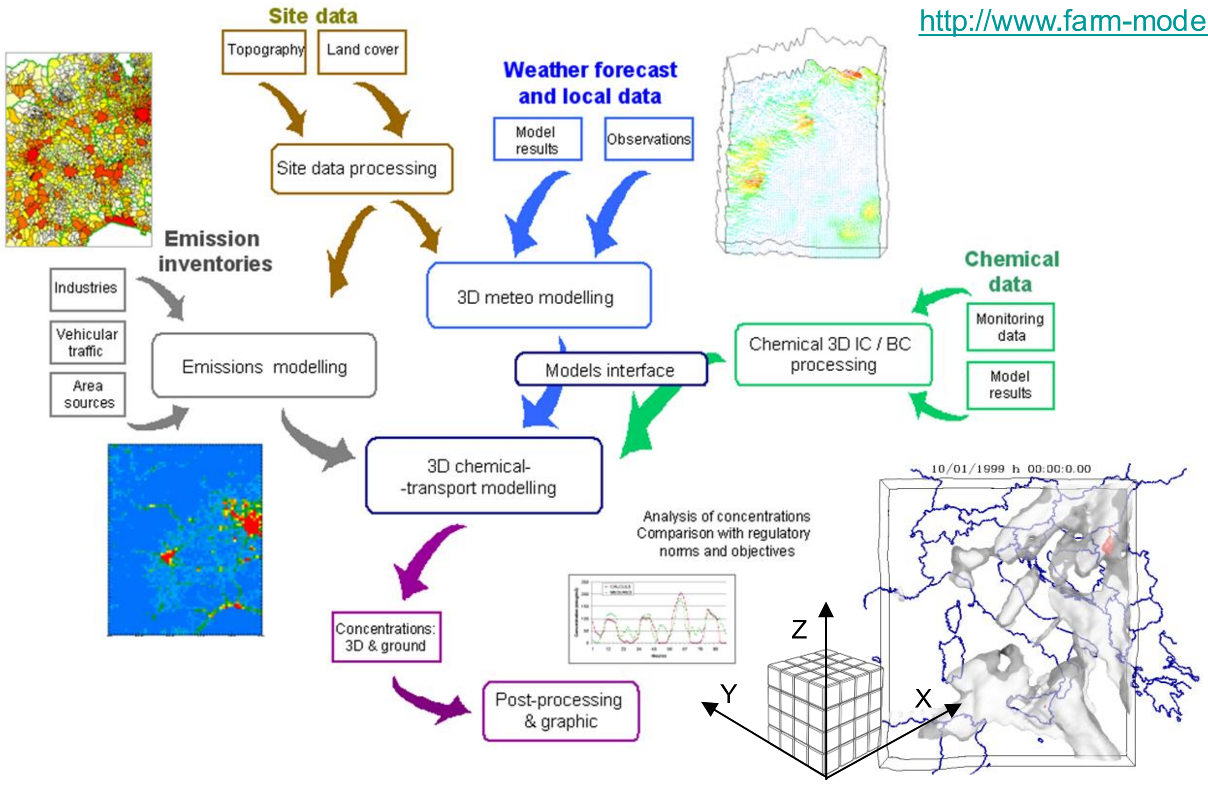


Model: CTM – FARM *(Flexible Air quality Regional Model)*

ARIANET
A suez company

- **Multi-grid Eulerian model** for dispersion, transformation and deposition of airborne pollutants in gas and aerosol phases

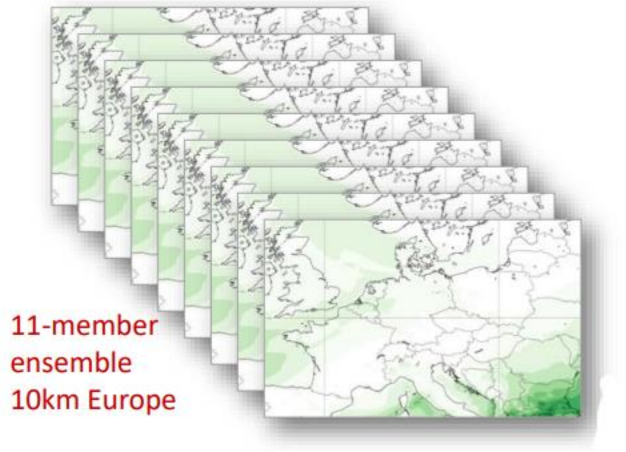
http://www.farm-model.org/



- Nested grids, from **continental** to **urban scales**;
- On- and off-line: **forecasts and re-analysis**

- Core of MINNI, operationally run to get **CAMS European air quality forecasts** in a multi-model approach

copernicus Europe's eyes on Earth
Atmosphere Monitoring Service atmosphere.copernicus.eu
ECMWF



https://atmosphere.copernicus.eu/european-air-quality-forecast-plots

IRIDE project

- A constellation of **Low Earth Orbit** (LEO) **satellites** to advance **environmental monitoring** and management from space

- Equipped with **optical**, **multispectral**, and **radar** technologies

- **High spatio-temporal resolution**, focused on the Mediterranean region

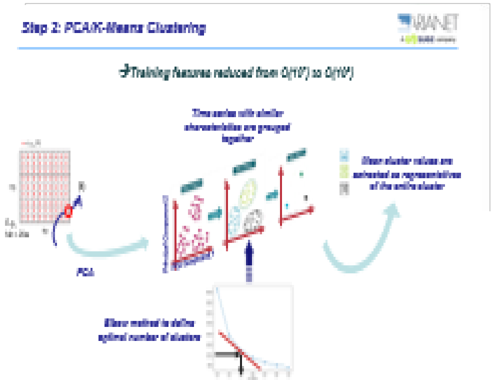
Funded by the European Union
NextGenerationEU

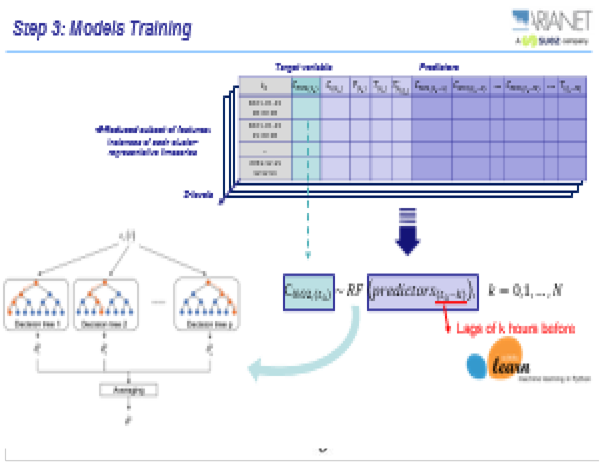
MEEO Meteorological Environmental Earth Observation

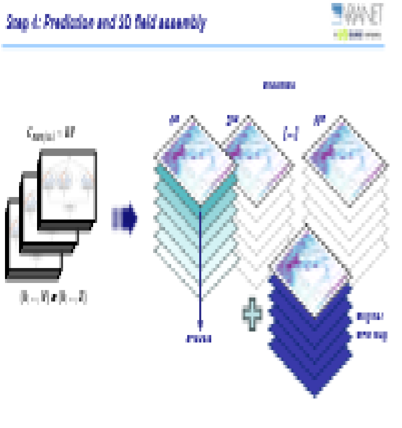
Agenzia Spaziale Italiana

Methodology

1. Inference Datasets Generation
2. Dimensionality reduction
3. Models Training
4. Prediction and 3D field assembly





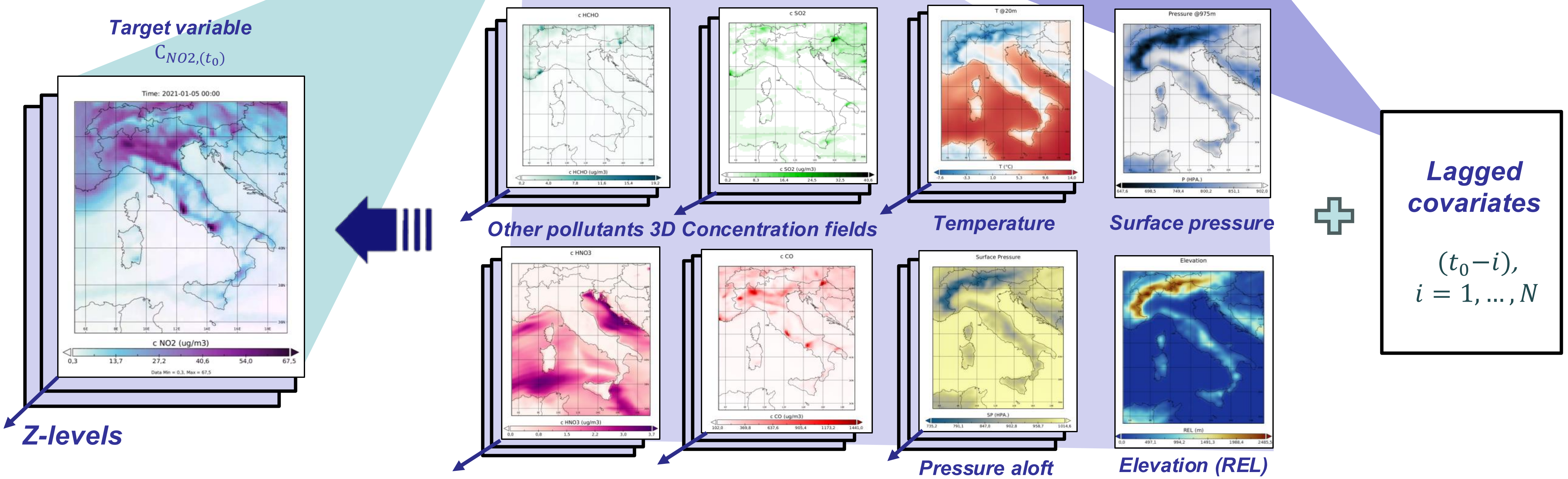


Step 1 : Inference Dataset Generation

Target variable — *Predictors*

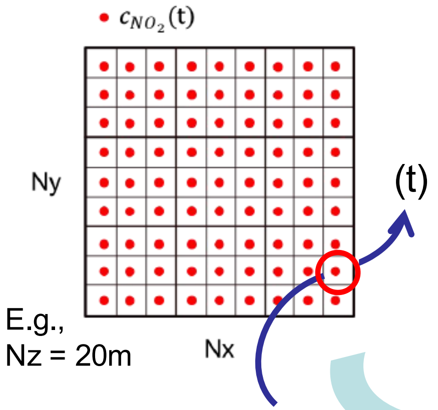
t_0	$C_{NO2,(t_0)}$	$C_{i,(t_0)}$	$P_{(t_0)}$	REL	$Sp_{(t_0)}$	$T_{(t_0)}$	$C_{NO2,(t_0-1)}$	$C_{NO2,(t_0-2)}$	…	$C_{NO2,(t_0-N)}$	…	$T_{(t_0-N)}$
2021-01-01 00:00:00												
2021-01-01 01:00:00												
…												
202x-xx-xx xx:xx:xx												

Z-levels



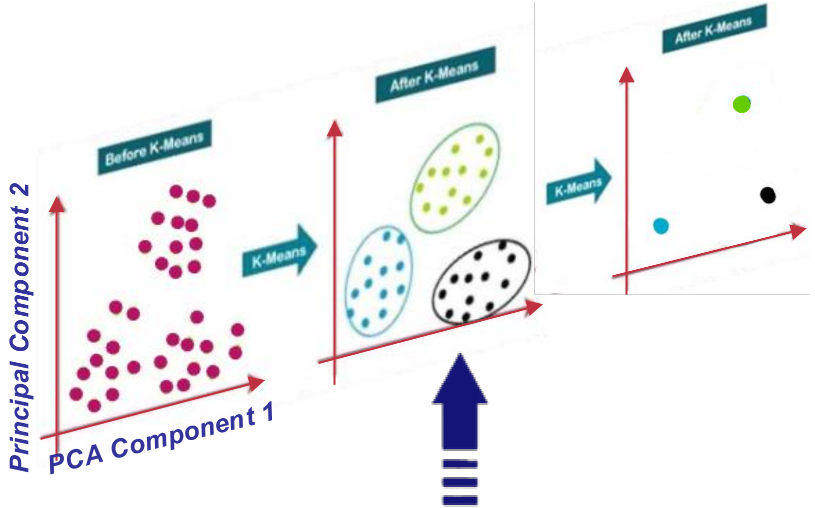
Step 2: PCA/K-Means Clustering

→ *Training features reduced from $O(10^7)$ to $O(10^5)$*



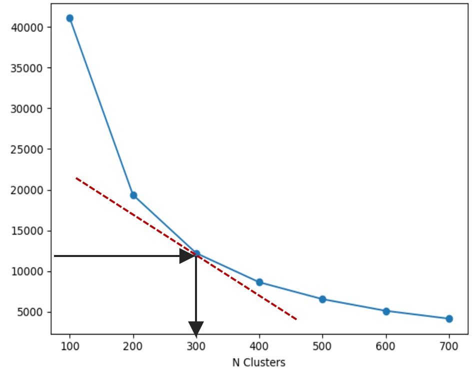
PCA

Time series with similar characteristics are grouped together



Mean cluster values are extracted as representatives of the entire cluster

Elbow method to define optimal number of clusters



Step 3: Models Training

Target variable — **Predictors**

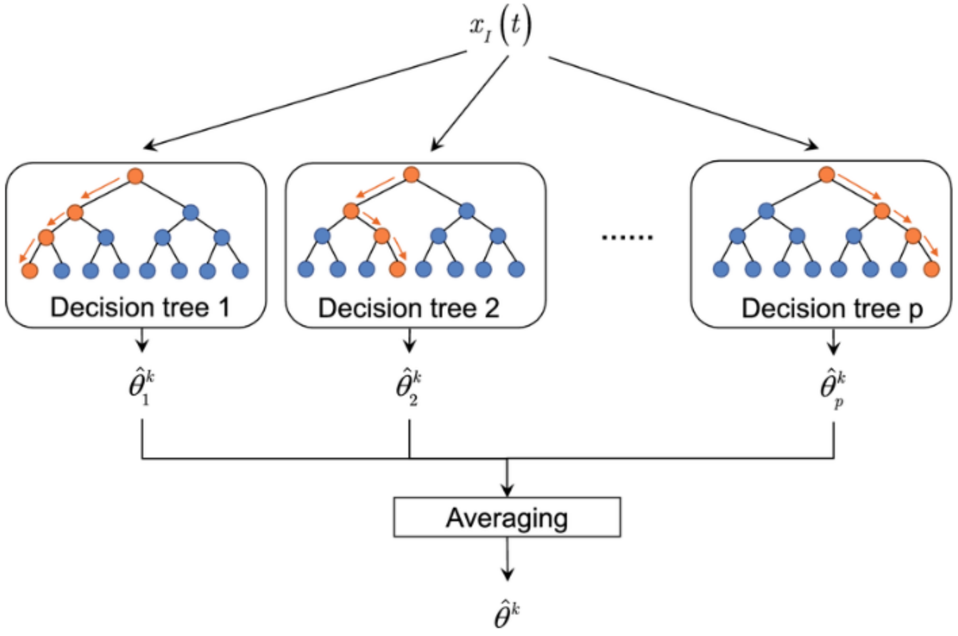
t_0	$\bar{C}_{NO2,(t_0)}$	$\bar{C}_{i,(t_0)}$	$\bar{P}_{(t_0)}$	$\bar{T}_{(t_0)}$	$\overline{Sp}_{(t_0)}$	$\bar{C}_{NO2,(t_0-1)}$	$\bar{C}_{NO2,(t_0-2)}$	…	$\bar{C}_{NO2,(t_0-N)}$	…	$\bar{T}_{(t_0-N)}$
2021-01-01 00:00:00											
2021-01-01 01:00:00											
…											
202x-xx-xx xx:xx:xx											

→ Reduced subset of features:
Instances of each cluster-representative timeseries

Z-levels

$$\boxed{C_{NO2,(t_0)}} \sim RF\,\boxed{\left(predictors_{(t_0-k)}\right)}, \quad k = 0,1,\dots,N$$

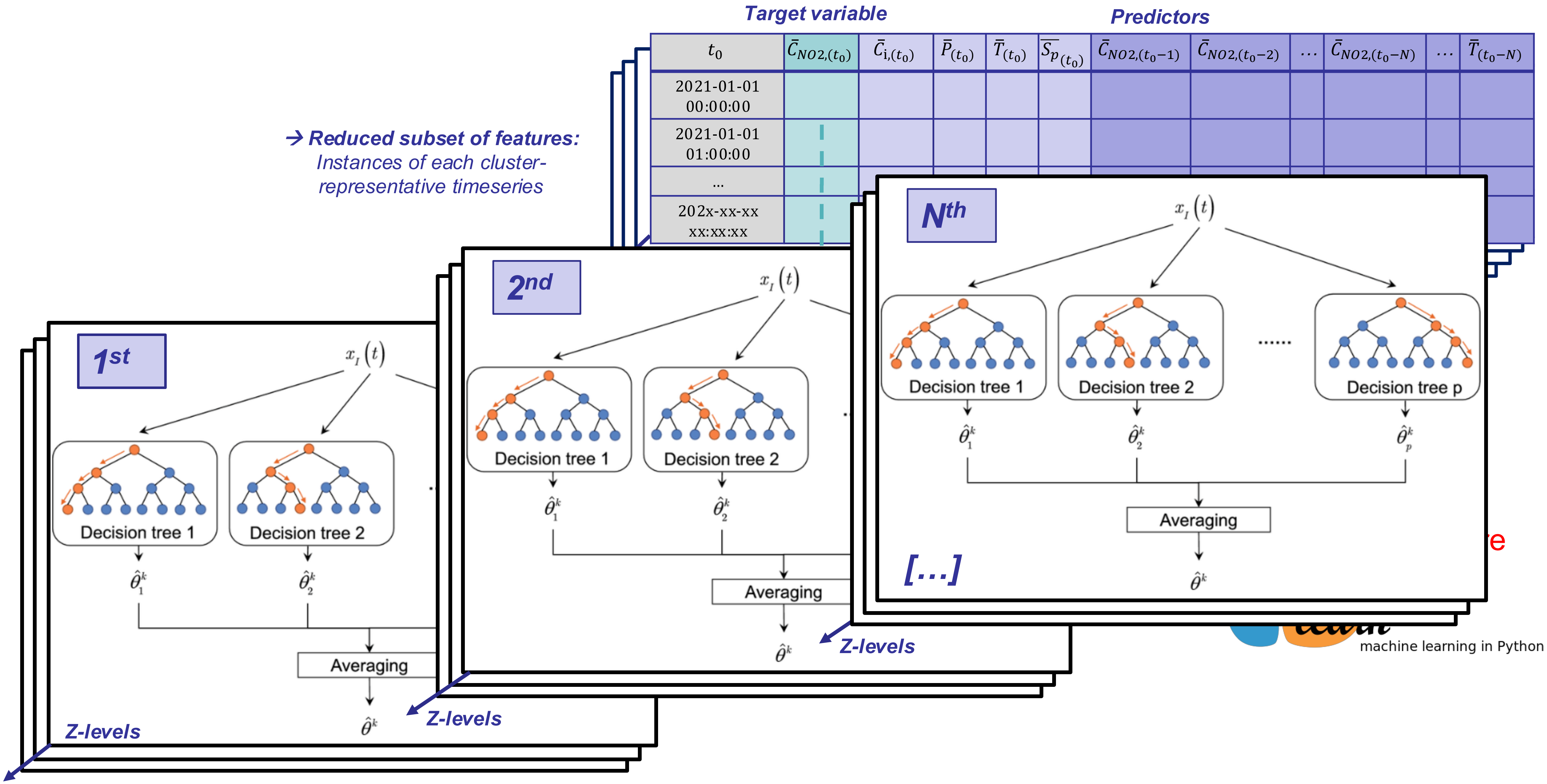
Lags of k hours before



machine learning in Python

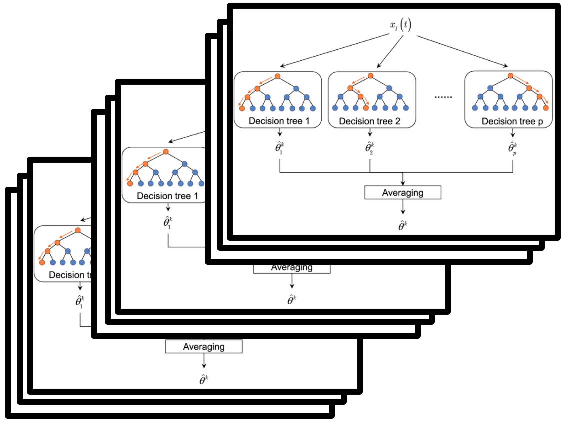
Step 3: Models Training

→ Reduced subset of features:
Instances of each cluster-representative timeseries

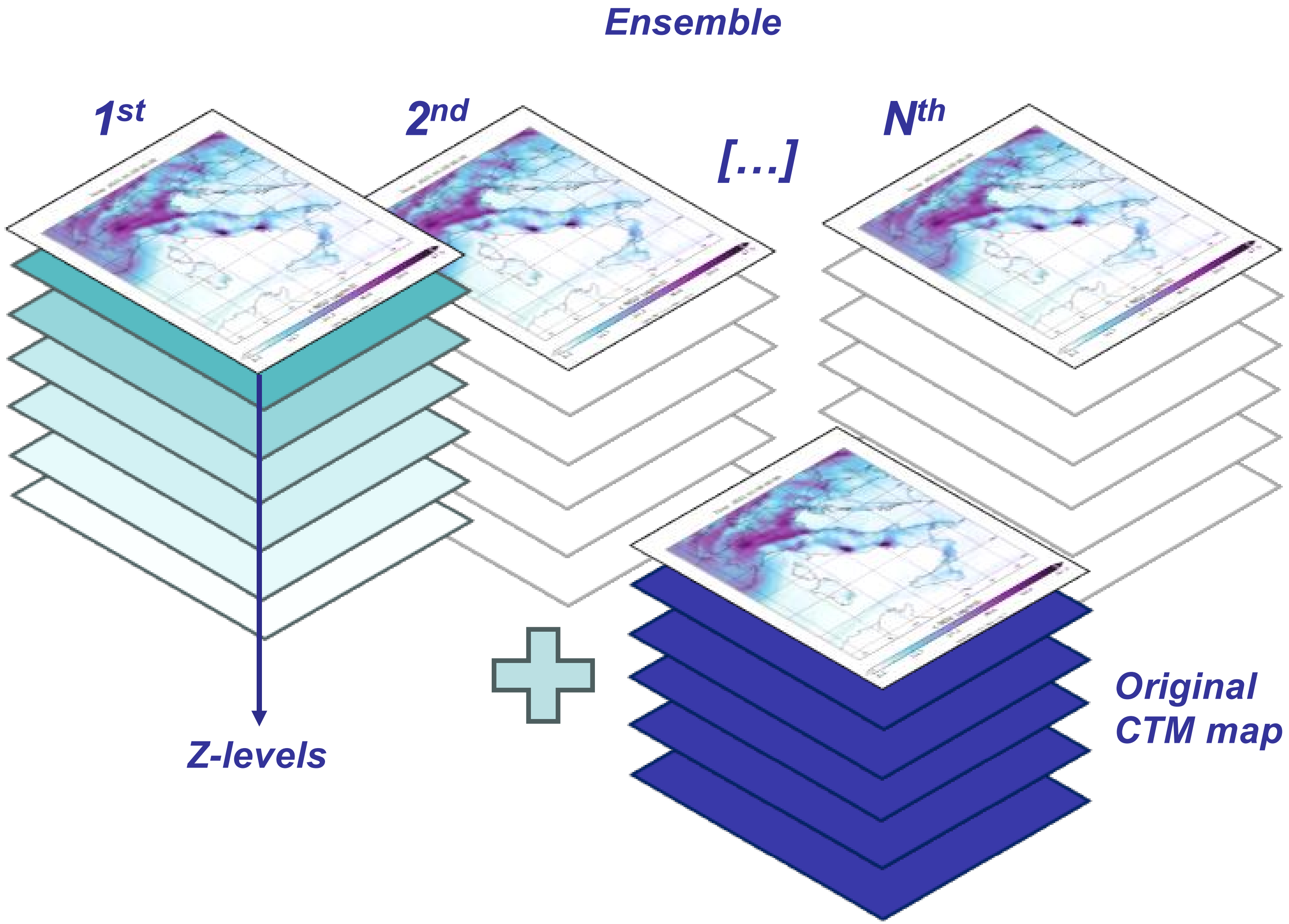


Step 4: Prediction and 3D field assembly

$$C_{NO2,(t_0)} \sim RF$$



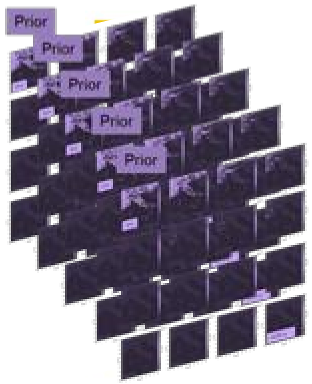
$$\{0, \dots, N\} \; x \; \{0, \dots, Z\}$$



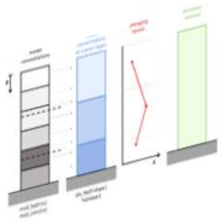
Data Fusion experiment

→ The ML-generated ensemble was tested in an offline Data Assimilation of S5p **tropospheric NO_2 total columns** observations in a FARM model simulations over the Italian domain during a **two-month winter period**.

→ **Data Fusion/OI** as the analysis isn't used to update initial conditions for subsequent runs, viable under the hypothesis that **model uncertainties grow quickly → Stationary B**

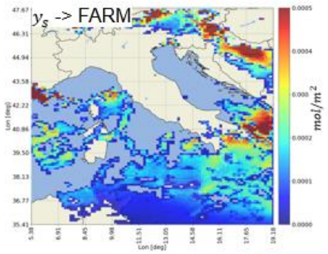


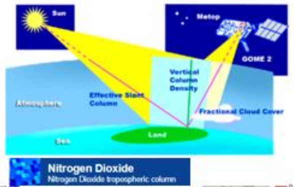
NCAR | DART

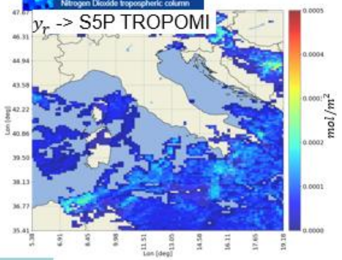


$H(x)$

Forward Operator
Function that maps model variables to observed quantities.



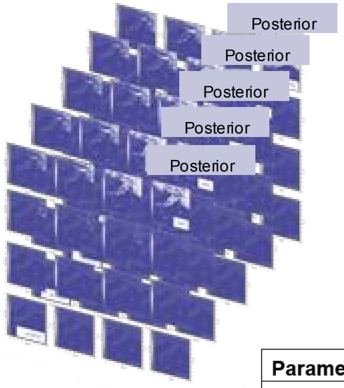




Increment

FARM-DART

Analysis



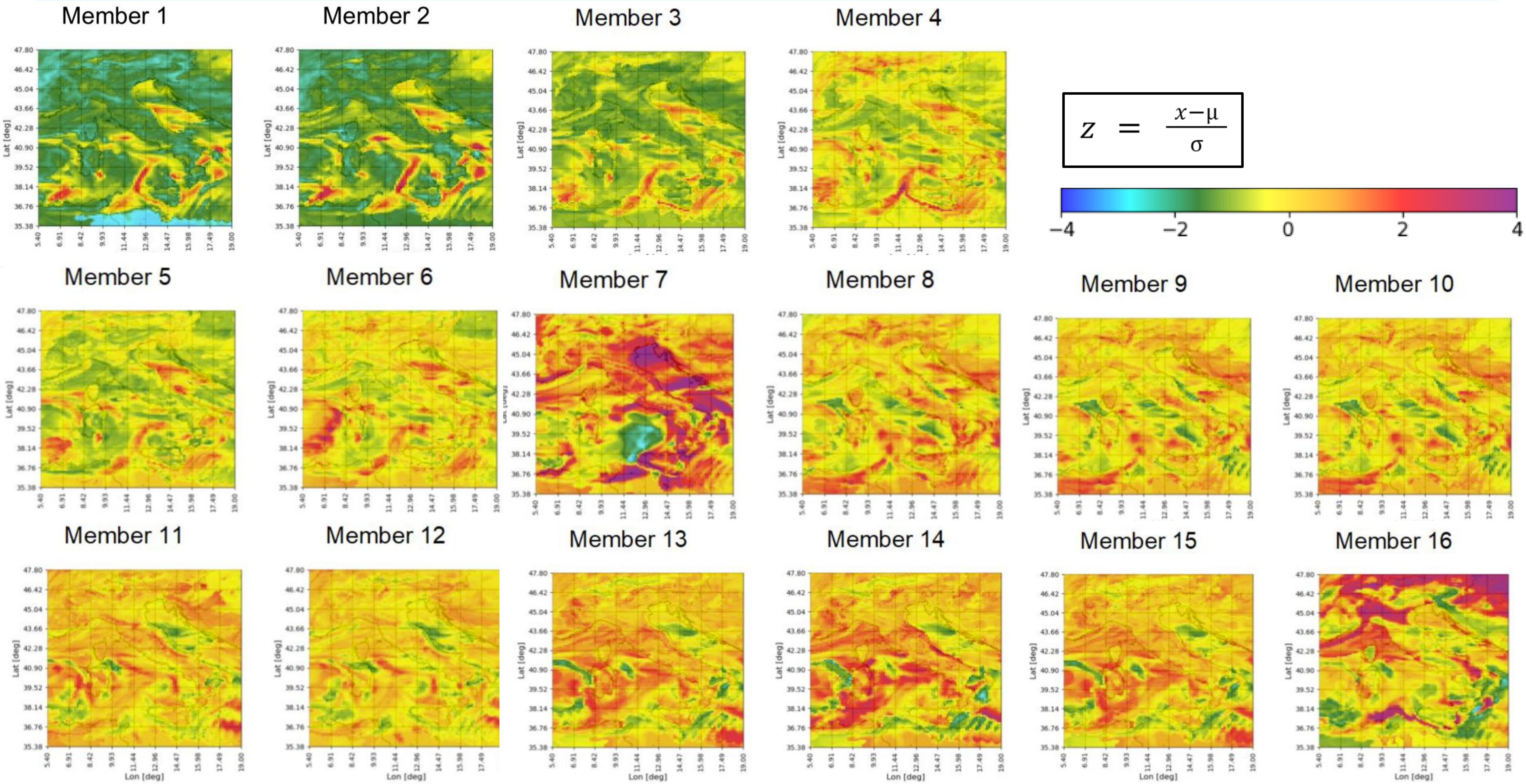
$$X_a = X_p + K[y_0 - Hx_f]$$

K: Filter

Parameter	Description
Model used	FARM
Vertical Levels	16 levels rangin from 20m near the surface to 9290 m at the top level
Horizontal Resolution	0.104° x 0.15°
Assimilation Tool	DART
Ensemble	17 members provided as netCDF files
Ensemble Spread Management	Not applied
Localization Function	Gaspari-Cohn
Cutoff Distance / Horizontal Influence	0.0025 (~30km / 0.25°)
Vertical Influence	3.6 km
Handling Negative Concentrations	Quantile-Conserving Ensemble Filter (QCEFF)
Filtering Method	Ensemble Adjustment Kalman Filter (EAKF)
Approach	Data Fusion: The analysis does not influence the next timesteps. Background Covariance Matrix considered as static: EnOI

Figure 19. Summary Data Fusion experiment.

Ensemble insight – z-scores



$$z = \frac{x - \mu}{\sigma}$$

Data Fusion Results – 1

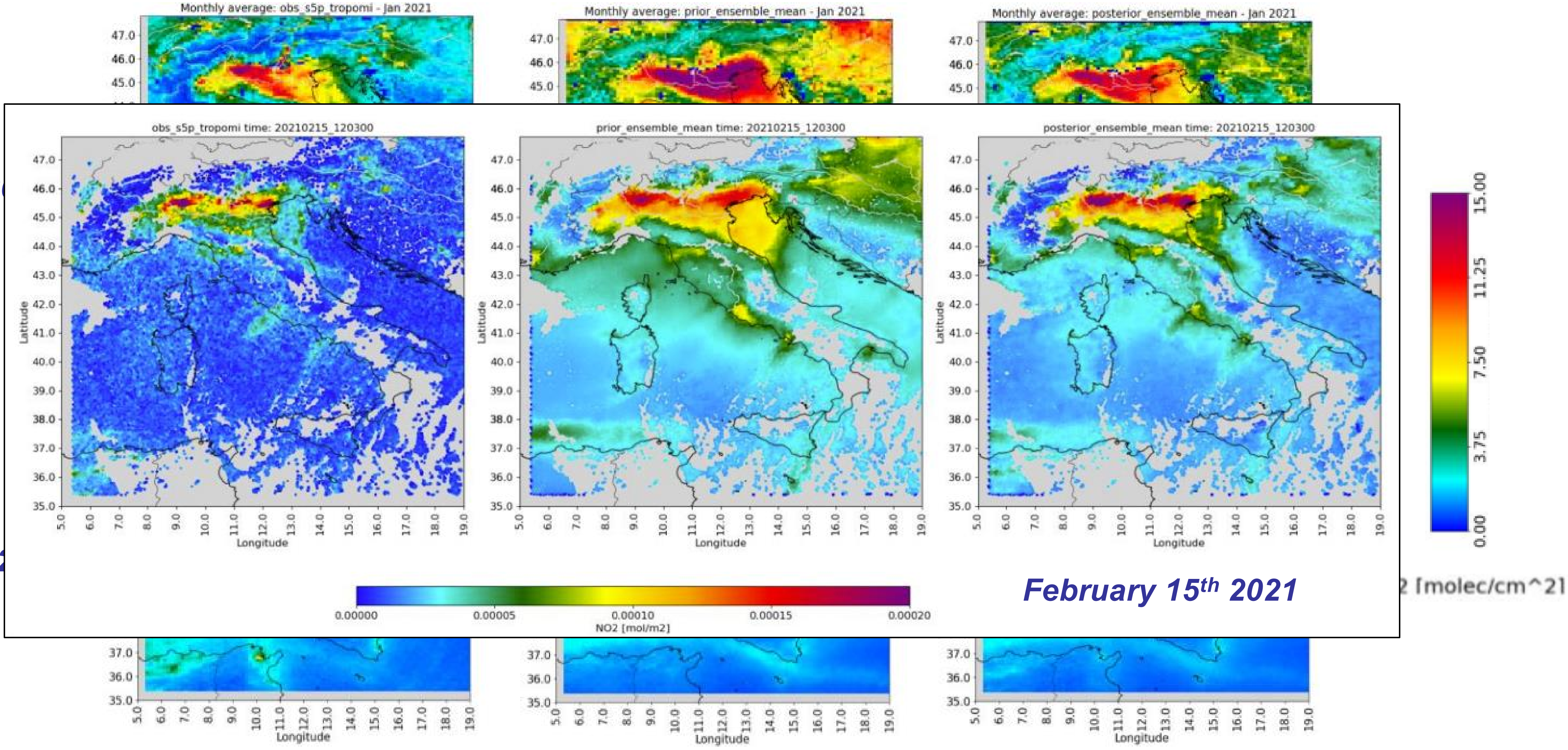
S5p NO2 vcd observations

Prior NO2 vcd ensemble mean

Posterior NO2 vcd ensemble mean

January 2

February 2

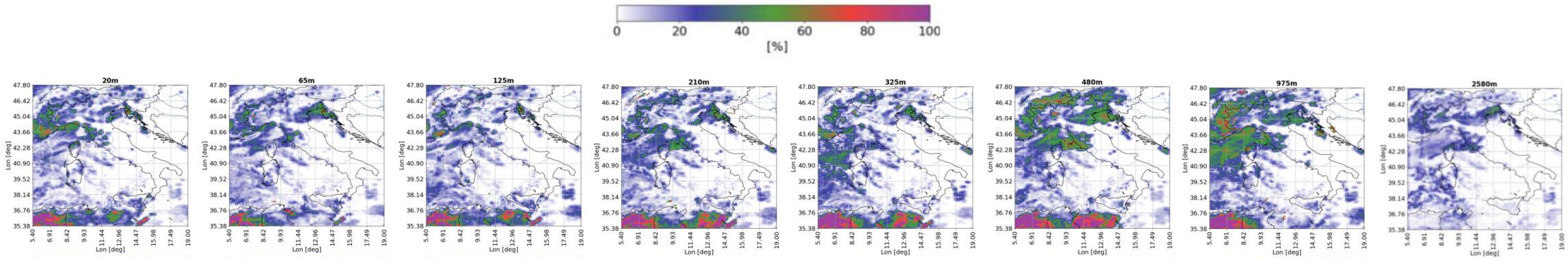


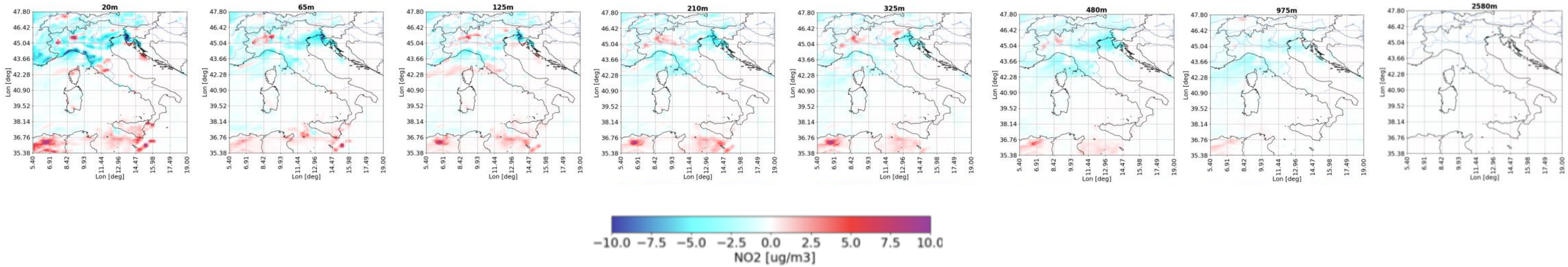
February 15th 2021

Monthly means

Data Fusion Results – 2

(prior-posterior)/prior





prior-posterior

Conclusions

- A **computationally efficient** alternative to conventional **ensemble generation** methods has been presented
- Leveraging on **multiple RF regressors**, an ensemble of CTM runs was synthetized; a statistical representation of the **uncertainties in the model's forecast** can be drawn
- Preliminary results show the presented methodology can capture **both diagonal and off-diagonal covariance matrix terms**
- By avoiding the computational burden of starting N perturbated runs, **operational DA** of satellite data for air quality applications can be feasible for national level's PAs → *towards geostationary Sentinel-4*
- Further work will focus on enabling **online DA**, as well as on anchoring the ensemble mean to the original CTM run

ARIANET

A SUEZ company

10th International Symposium on Data Assimilation
Kobe, Japan
October 21-25, 2024

ISDA Kobe 2024

POLITECNICO MILANO 1863

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

Giorgia De Moliner (PhD Candidate, Politecnico di Milano), **Alessandro D'Ausilio** (ARIANET-SUEZ), **Camillo Silibello** (ARIANET-SUEZ), **Giovanni Lonati** (Professor, Politecnico di Milano)

iride

GOVERNO ITALIANO | ASI | ESA

MEEO
Meteorological Environmental Earth Observation