



Data Assimilation activities at CMCC-REMHI: current developments and ongoing work

Francesca Vittorioso, Alessandro Sebastianelli, Mario Raffa, Ilenia Manco, Otavio Feitosa, Oliver Carlo, Gabriella Ceci, Paola Mercogliano

francesca.vittorioso@cmcc.it
alessandro.sebastianelli@cmcc.it

CMCC – Institute for Climate Resilience
www.cmcc.it



14.09.2026



Reanalysis

Assimilation of historical data from ground stations and satellite sensors to generate consistent, continuous, and high-resolution climate datasets.



Weather Forecasting

Real-time integration of observations (*in-situ* and remote sensing) into NWP models to improve the accuracy of short-term monitoring, especially for extreme event analysis.

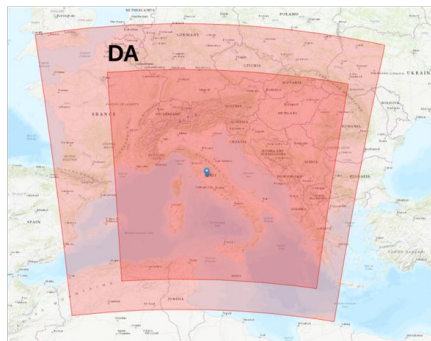


AI & data-driven approaches

Towards hybrid modelling and DA: integration of physical models, observations and ML; data-driven approaches for model enhancement and prediction.

- Horizontal resolution 9 km / 1.8 km
- Hybrid Vertical Levels 60
- Boundary Conditions ERA5
- Initial Conditions WRF-DA 3D-Var
- Version 3.7.1

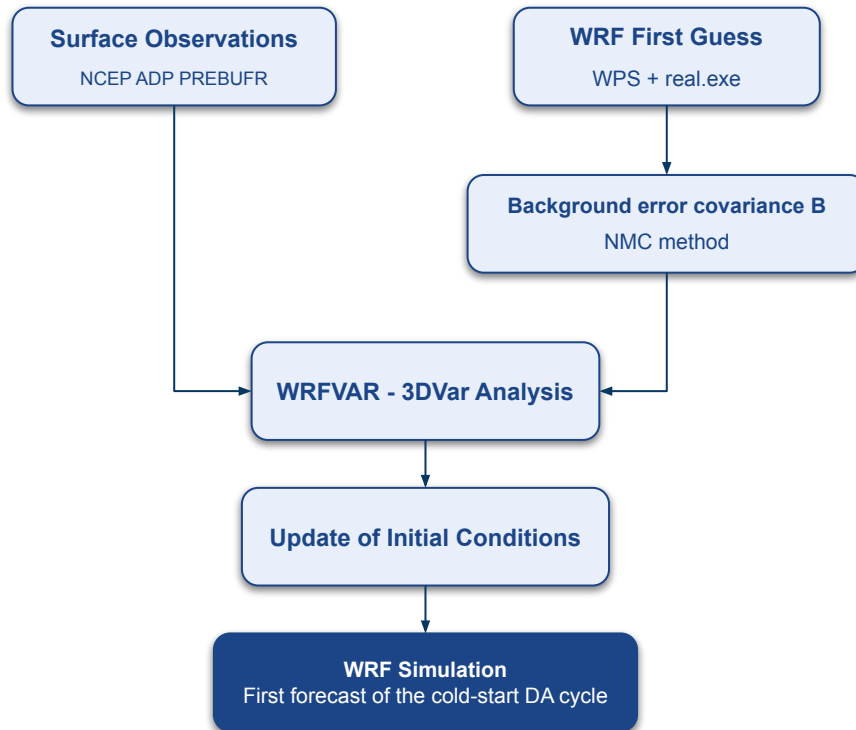
9km / 1.8 km



Convection-permitting configuration at kilometre scale
CONUS parametrizations set-up:

- PBL MYJ (Mellor-Yamada-Janjić, Ma Y. F., Jan., 2024)
- Surface (Monin-Obukhov (Janjic Eta), Janjić, Z. I., 1994)
- Surface Noah LSM (Chen, F., & Dudhia, J., 2001)
- RAD RRTMG (Iacono, M. J., 2008)

WRF 3DVar - Cold Start Data Assimilation



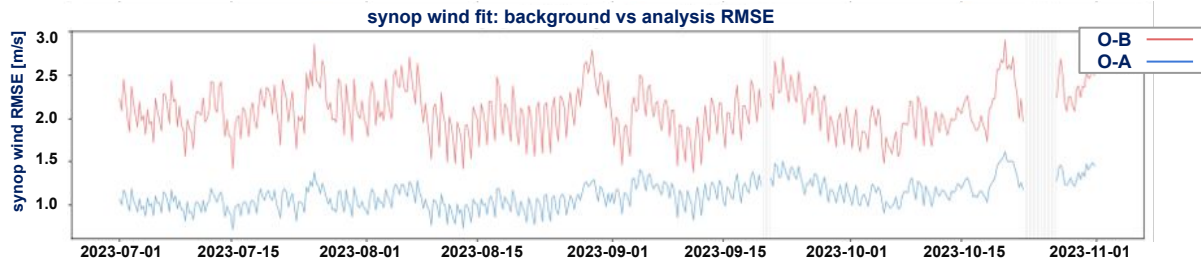
DA Impact in WRF model

→ Particularly pronounced impact on the wind field over the Po Valley.

→ For other meteorological variables, relying solely on NCEP station assimilation yields a limited effect due to low network density.

→ Integrating the HYRAD4CC dataset (approximately 5000 stations) leads to a clear and **widespread performance gain across all variables**.

→ These results underscore the critical importance of high-density spatial coverage for effective data assimilation.



*Courtesy of
I. Manco
and
O. Feitosa*

VARIABLE	RMSE (O-B)	RMSE (O-A)	REDUCTION (%)
U Wind (Zonal Component)	4.58 m/s	4.24 m/s	▲ 7.5%
V Wind (Meridional Component)	4.26 m/s	3.94 m/s	▲ 7.5%
Temperature	2.15 K	1.94 K	▲ 10.2%
Pressure	122.3 Pa	107.2 Pa	▲ 12.3%
Specific Humidity (Q)	2.08×10^{-3}	1.84×10^{-3}	▲ 11.7%



Building on the group's experience with ICON within the COSMO consortium (and the positive results achieved with WRF), we started developing **data assimilation activities** on **ICON** in limited-area model configuration **over Italy**, focusing on two main activities:

- Implementing “**Classical**” **Data Assimilation** in ICON model in LAM configuration over Italy with a specific **focus on Satellite Data class of observation**
- Implementing a multi-branch **Machine Learning framework** to combine NWP, satellite and ground-station data, **aimed at improving ICON's forecast accuracy**

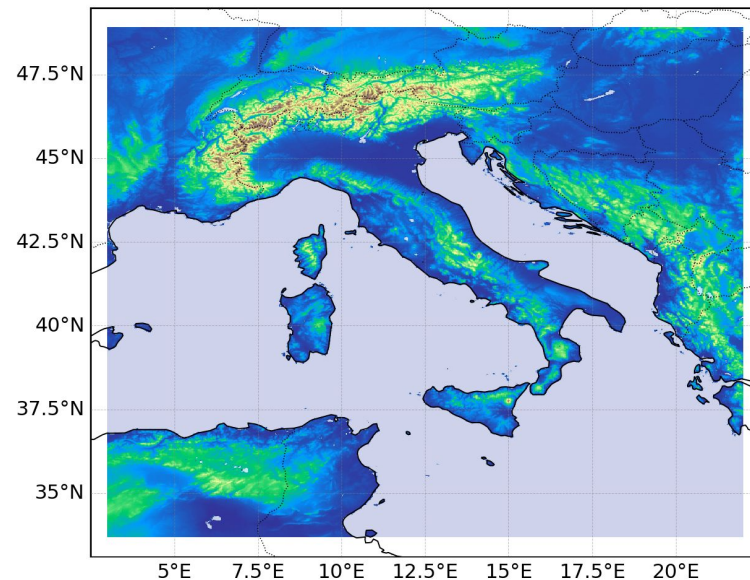
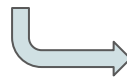
Building on the group's experience with ICON within the COSMO consortium (and the positive results achieved with WRF), we started developing **data assimilation activities** on **ICON** in limited-area model configuration **over Italy**, focusing on two main activities:

- Implementing “**Classical**” **Data Assimilation** in ICON model in LAM configuration over Italy with a specific **focus on Satellite Data class of observation**
- Implementing a multi-branch **Machine Learning framework** to combine NWP, satellite and ground-station data, **aimed at improving ICON's forecast accuracy**

- Regional configuration, as a limited area model (**LAM**) covering the **italian domain**

Model integration performed with:

- **Horizontal resolution** **2.2 km**
- **Hybrid Vertical Levels** **65**
- **Boundary Conditions** **ERA5**



- ★ Convection-permitting configuration at kilometre scale
- ★ Explicit representation of high-impact weather processes

- **Initial Conditions** **LeTKF/KENDA**

OUR SET-UP:

20 ensemble members:

number to be tested - is it enough to reproduce variability for long periods and high resolution?

3h assimilation cycle

Observations:

- Satellite data (in collaboration with DWD and ARPAE)
- Exploring other data class assimilation, like SYNOP

OUR SET-UP:

20 ensemble members:

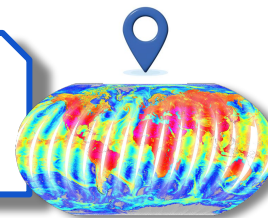
number to be tested - is it enough to reproduce variability for long periods and high resolution?

3h assimilation cycle

Observations:

- **Satellite data (in collaboration with DWD and ARPAE)**
- Exploring other data class assimilation, like SYNOP

Microwave Sounders
from MHS (microwave
Humidity Sounder - Metop)



- One of the most impactful observing systems for global NWP
- **Limited (but growing) adoption in convection-permitting regional models**
- Mainly needed to improve precipitation forecasting, but proven to have an impact also on humidity, temperature and wind

OUR SET-UP:

20 ensemble members:

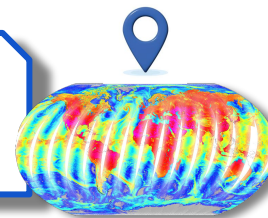
number to be tested - is it enough to reproduce variability for long periods and high resolution?

3h assimilation cycle

Observations:

- **Satellite data (in collaboration with DWD and ARPAE)**
- Exploring other data class assimilation, like SYNOP

Microwave Sounders
from MHS (microwave
Humidity Sounder - Metop)



In collab
with
ARPAE
and
DWD



- See T. Gastaldo's presentation (ARPAE). Their results come from the use of operational configuration ARPAE runs for the national chain with ICON-2I: different boundaries, different time-step, plus other classes of obs assimilated.
- However, the implementation is being carried out in collaboration, and as soon as possible we will also run it on our own machines, to test the more scientific/research-oriented impact it could have on REHMI's activities (e.g. on historical periods/reanalysis?).

OUR SET-UP:

20 ensemble members:

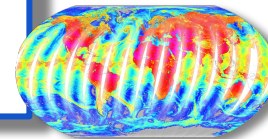
number to be tested - is it enough to reproduce variability for long periods and high resolution?

3h assimilation cycle

Observations:

- **Satellite data (in collaboration with DWD and ARPAE)**
- Exploring other data class assimilation, like SYNOP

Microwave Sounders
from MHS (microwave
Humidity Sounder - Metop)



**Geostationary Imagers
(VIS-IR)**
SEVIR-MSG / FCI-MTG



OUR SET-UP:

20 ensemble members:

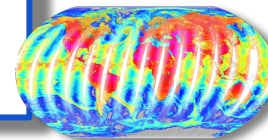
number to be tested - is it enough to reproduce variability for long periods and high resolution?

3h assimilation cycle

Observations:

- **Satellite data** (in collaboration with DWD and ARPAE)
- Exploring other data class assimilation, like SYNOP

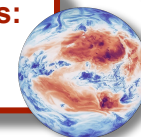
Microwave Sounders
from MHS (microwave
Humidity Sounder - Metop)



**Geostationary Imagers
(VIS-IR)**
SEVIR-MSG / FCI-MTG



Future Developments:
Infrared Sounder
IRS-MTG



OUR SET-UP:

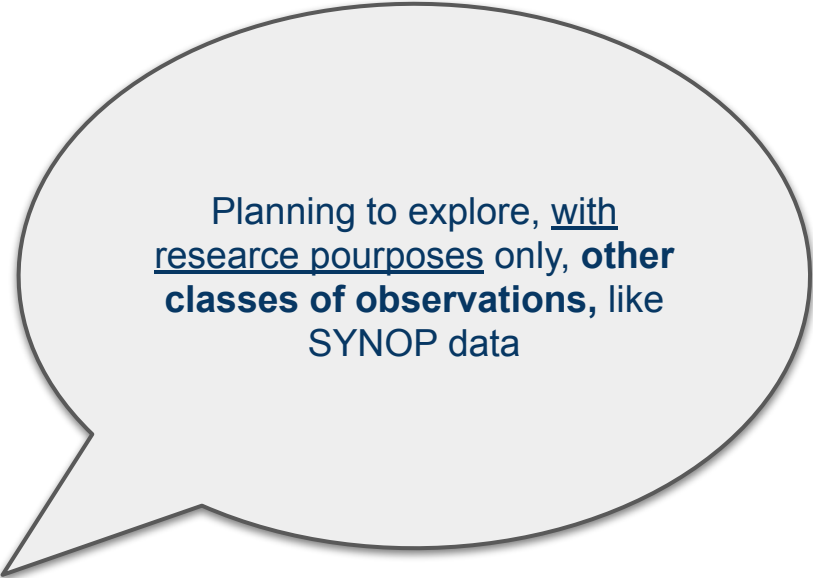
20 ensemble members:

number to be tested - is it enough to reproduce variability for long periods and high resolution?

3h assimilation cycle

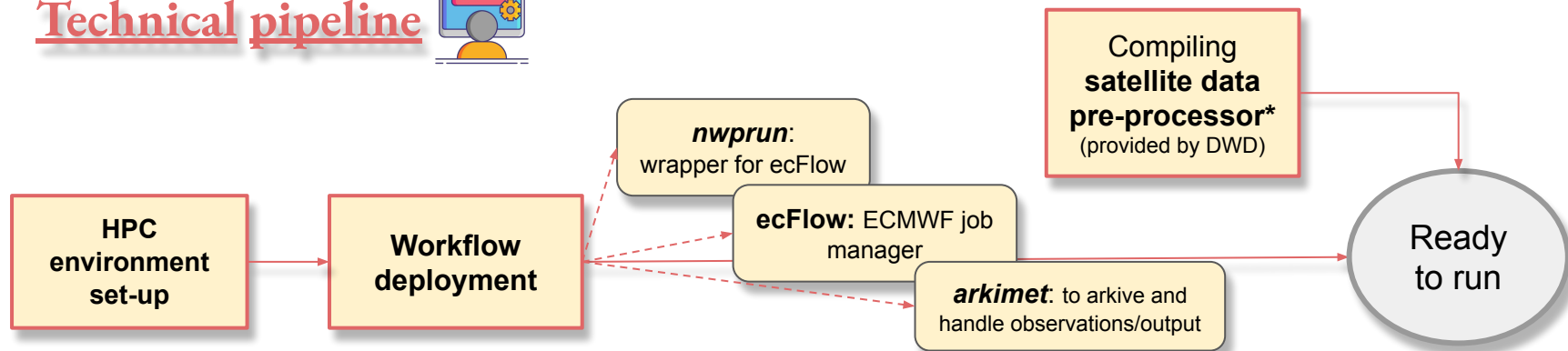
Observations:

- Satellite data (in collaboration with DWD and ARPAE)
- **Exploring other data class assimilation, like SYNOP**



Planning to explore, with
research purposes only, **other
classes of observations**, like
SYNOP data

Technical pipeline



*SAT_PP compiled and functioning for MHS, still ongoing compilation with HRIT for SEVIRI



→ Technical environment for assimilation over Italy set-up - independent experiments are about to start

MAIN FOCUS ON SATELLITE DATA

→ **MHS assimilation has already shown encouraging results through our collaboration with ARPAE**
- the next step is running it in our own environment, and then extending to SEVIRI

→ **SAT_PP** working for MHS; compilation for SEVIRI to be solved

→ **Active collaboration with ARPAE, ongoing exchange with DWD**
- moving faster by building on each other's work

Building on the group's experience with ICON within the COSMO consortium (and the positive results achieved with WRF), we started developing **data assimilation activities** on **ICON** in limited-area model configuration **over Italy**, focusing on two main activities:

- Implementing “**Classical**” **Data Assimilation** in ICON model in LAM configuration over Italy with a specific **focus on Satellite Data class of observation**
- Implementing a multi-branch **Machine Learning framework** to combine NWP, satellite and ground-station data, **aimed at improving ICON's forecast accuracy**

AI-based Post-Processing DA Pipeline — Overview

DATA SOURCES

ICON NWP

2km · hourly · 2023
619×675 grid

MSG/SEVIRI

EUMETSAT Data Tailor
hourly · 2023
11 channels + 9 BT

NOAA ISD Stations

Global hourly obs
~640 stations · 2023
2m air temperature

Domain

Italy & surroundings
lat 36°–48.36°N · lon 5.9°–19.38°E
Period: Jan–Dec 2023

Train/Val Split (by month)

Train: Jan Feb Apr May Jul Aug Oct
Nov (8 months)
Val: Mar Jun Sep Dec (1 per season)
Year: 2023

PREPROCESSING

preprocess_icon.py

Extract t2m per timestep
→ icon_YYYYMMDDHH.nc
8760 files · K units

regrid_seviri.py

Bilinear interp → ICON grid
→ seviri_YYYYMMDDHH.nc
8752 files · 619×675

regrid_stations.py

KD-tree nearest-neighbor
sparse 1D → 2D 619×675
station_temp + station_mask

compute_normalization_stats.py

ICON T2M: z-score · DEM: min-max
SEVIRI: z-score per channel
Stations: z-score (valid pixels only)

icon_static.nc

DEM (orography) — loaded once
land_fraction · lat · lon

DATASET & TRAINING

Dataset

Loads ICON + SEVIRI + Stations per timestep
Applies normalization
Random spatial crop 256×256
Random flip augmentation
Stations: ground truth only (not model input)

Train: 5832 · Val: 2928 samples

2-Branch U-Net

Branch 1: ICON T2M (1ch)
Branch 2: SEVIRI + DEM (19+1ch)

2 parallel encoders
Feature fusion at each depth level
Single decoder → ΔT
Output: T2M_corrected = ICON + ΔT (K)

Loss Function

MSE(T2M_corrected, station_K) @
station_mask
+ Spatial smoothness regularization
GT: NOAA ISD stations (not a model input)
Optimizer: AdamW · LR: 5e-4
Scheduler: CosineAnnealingLR

INFERENCE

Input

New ICON forecast
+ SEVIRI obs (real-time)
(no station data needed)

Model Forward Pass

2-branch encoder
→ fused features
→ decoder → Δ field

Output

T2M_corrected = T2M_ICON + Δ field
Full 619×675 grid
Resolution: ~2km

Evaluation Metrics

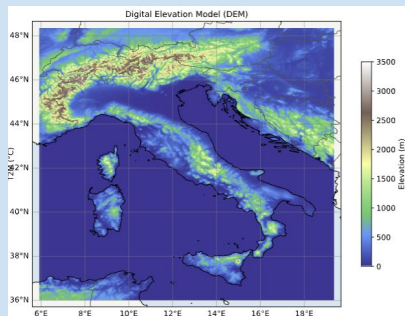
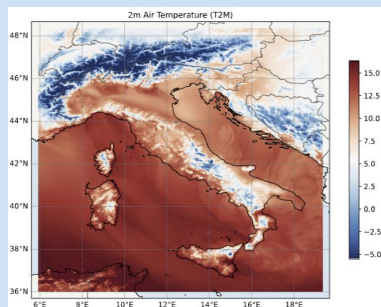
RMSE vs station obs
MAE vs station obs
Bias
Spatial correlation

The framework is modular.

With proper pre-processing, it can be used with ICON, COSMO and WRF.

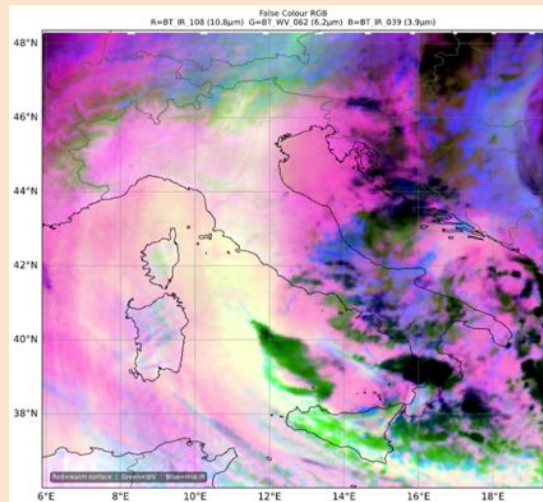
A model-based retrain is needed.

NWP



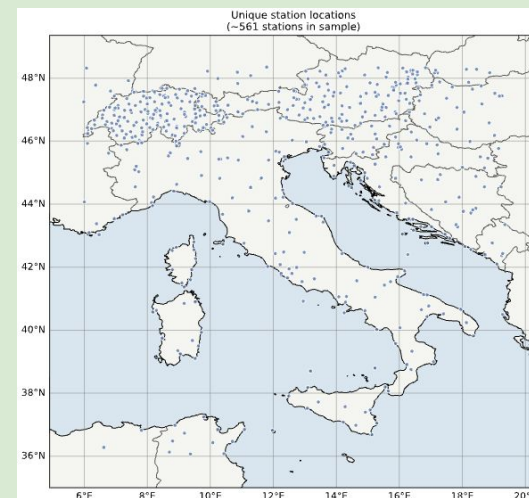
ICON/WRF/COSMO

EO



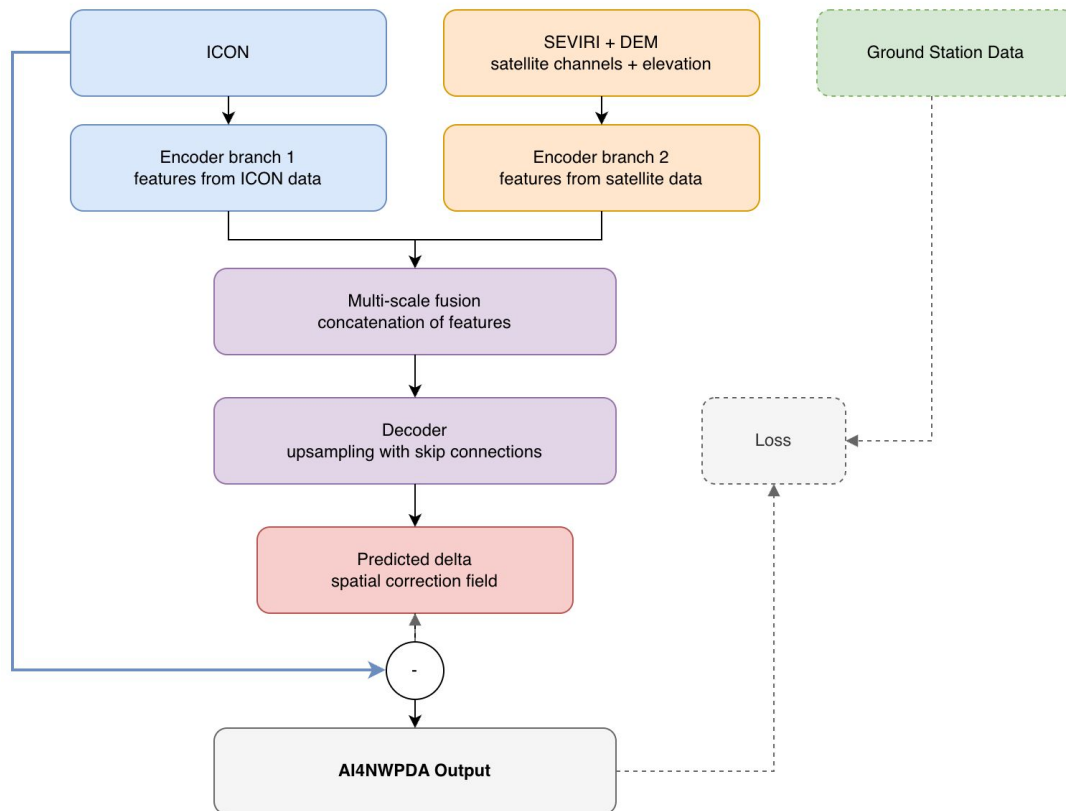
MSG SEVIRI

Ground Stations



NOAA ISD

Multi-branch U-Net



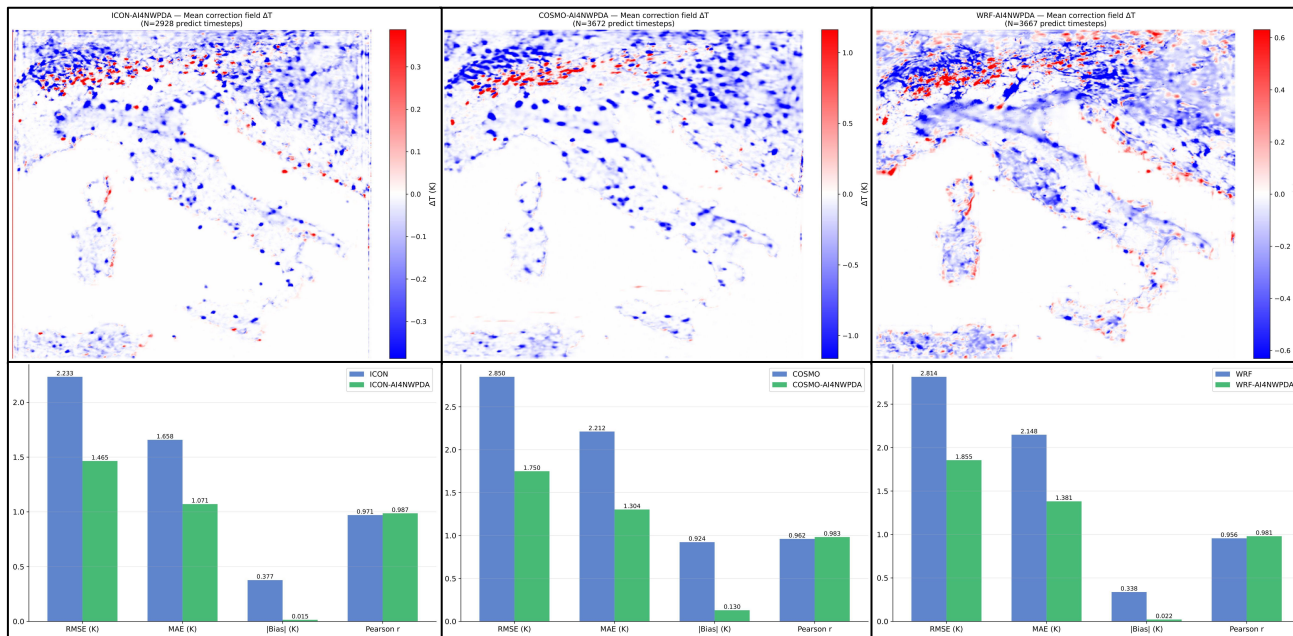
→ The model is modified and custom U-Net.

→ The model can accommodate heterogeneous data via multi-branching.

The structure is configurable, for now it includes the NWP and EO branch.

→ Initial tests are on T2M, but this can be changed with config file.

Results



NWP	RMSE base	RMSE AI4NWPDA	Δ RMSE %	MAE base	MAE AI4NWPDA	Δ MAE %	Bias base	Bias AI4NWPDA	Δ Bias (K) ^a
ICON	2.233	1.465	-34.4%	1.658	1.071	-35.4%	+0.377	+0.015	-0.362
COSMO	2.850	1.750	-38.6%	2.212	1.304	-41.0%	+0.924	+0.130	-0.794
WRF	2.814	1.855	-34.1%	2.148	1.381	-35.7%	+0.338	+0.022	-0.316

- Performance measured against the stations
- The 3 models have been tested and results are encouraging
- Next important step, multivariate approach

Drafting a paper

Multivariate NWP Input
(e.g. t2m, humidity)

Add more aux data
(e.g. *doy*, *encode time*)

Physics Loss
(e.g. *orography and seasonal weighting or classic DA oriented*)

Attention Mechanism on
U-net bottleneck
(*adaptive weighting of EO and NWP*)

Expand the training period

Run a proper
hyperparameters search

Add other sources to be
assimilated

Test resilience w.r.t. missing
sources



cmcc

Centro Euro-Mediterraneo
sui Cambiamenti Climatici

www.cmcc.it



Thanks for your attention!