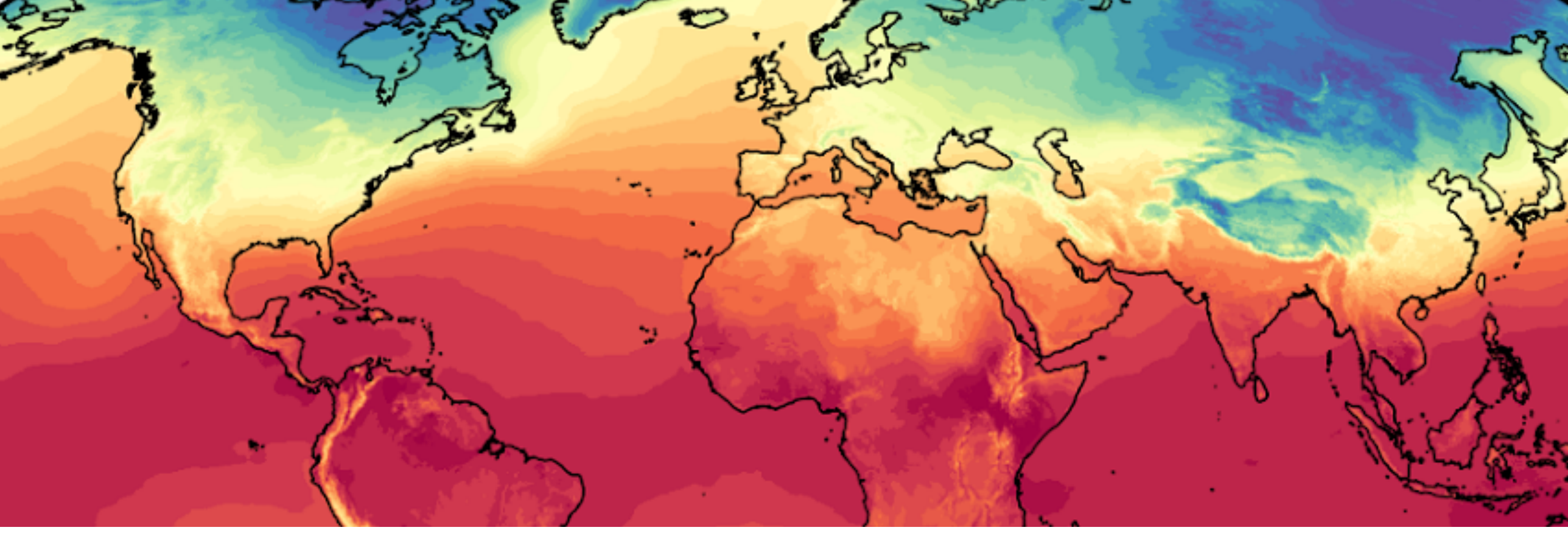


EMP²: Environmental Modelling and Prediction Platform

Christian Lessig, Ilaria Luise, Martin Schultz,
Michael Langguth, Alberto Di Meglio, Maria Girone et al.

KT Meeting - 3rd December 2024

CERN | CIPEA CERN Innovation Programme on Environmental Applications | ECMWF | JÜLICH Forschungszentrum

EMP2: Summary

Title: EMP2 - Environmental Modelling and Prediction Platform
Dates: March 2023 - Feb 2025
Status: Executing

Principal Investigators

- Ilaria Luise (CERN)
- Christian Lessig (ECMWF, Magdeburg University)
- Martin Schultz (JSC)

CERN Team

- Project Leader: Alberto Di Meglio
- Project employees: Ilaria Luise, funded (100% FTE)

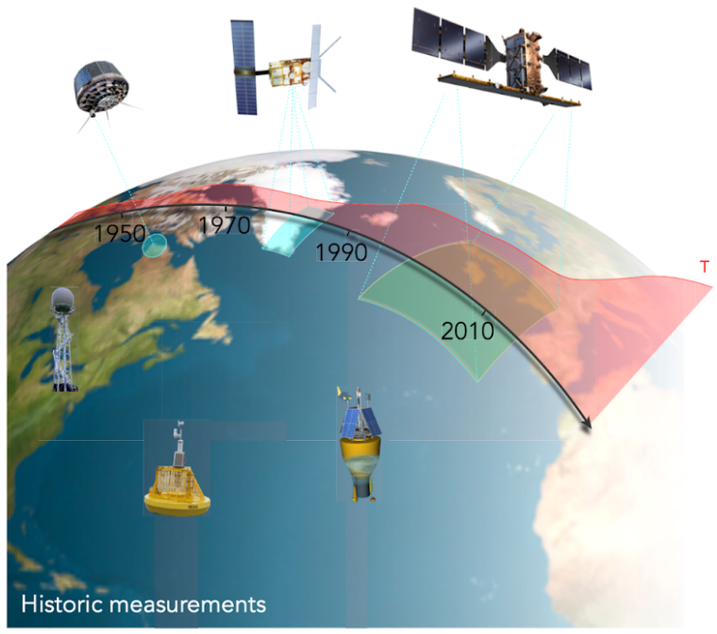
Partners

Fundings

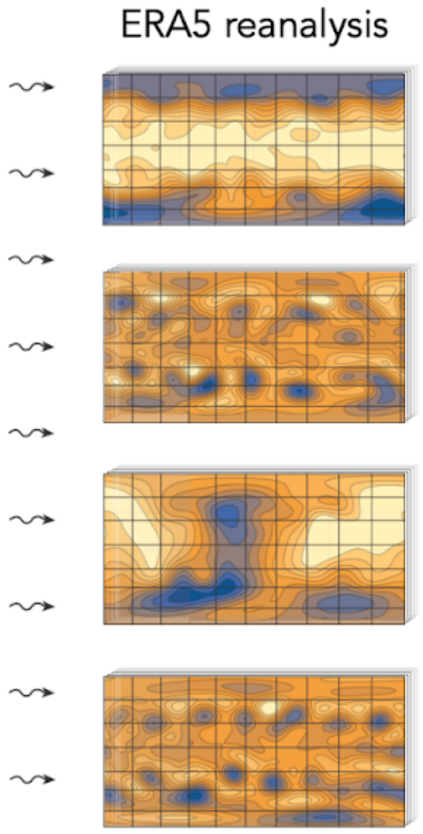
- Source: Funds provided by the Knowledge Transfer fund
- Allocated to: cover Ilaria Luise's Research Fellowship Grant
- Amount: 224.000 CHF for the period 2023-2025

The project in a nutshell

A machine-learning based global environmental model trained on terabytes of observational data



Observations

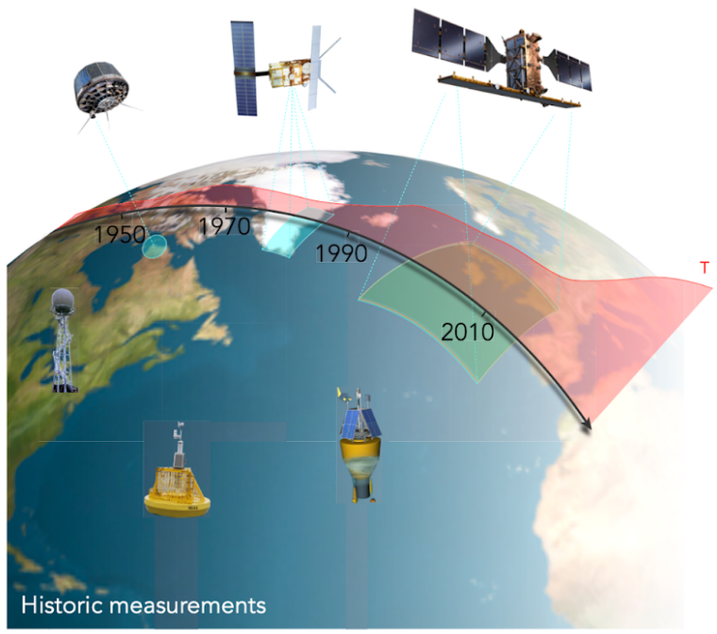


50 TB

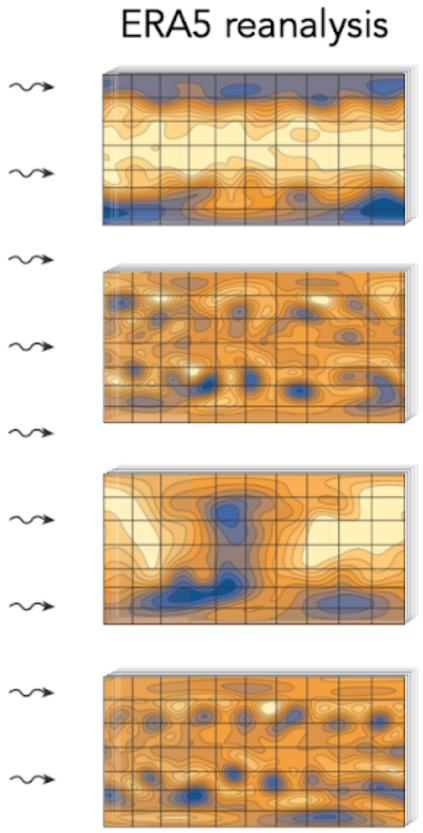
Ilaria Luise, CERN - ilaria.luise@cern.ch

The project in a nutshell

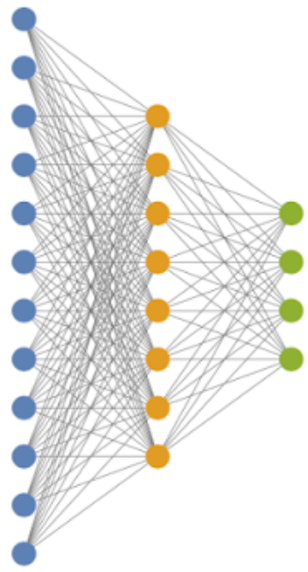
A machine-learning based global environmental model trained on terabytes of observational data



Observations



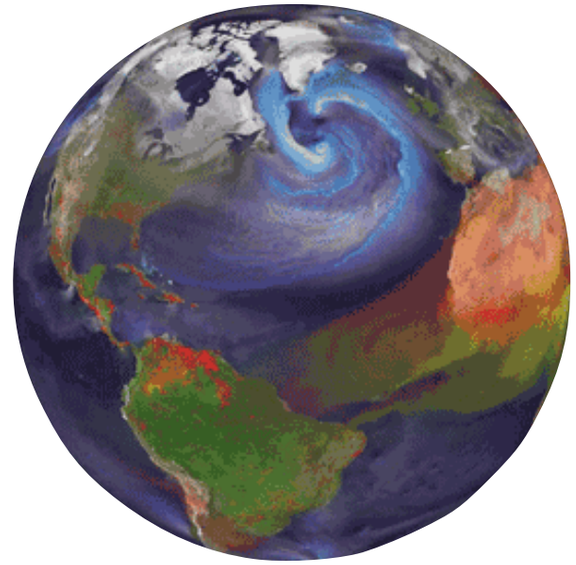
50 TB



large scale machine learning

Transformers architecture

Spatio-temporal representation of atmospheric dynamics



Model
given by the trained neural network

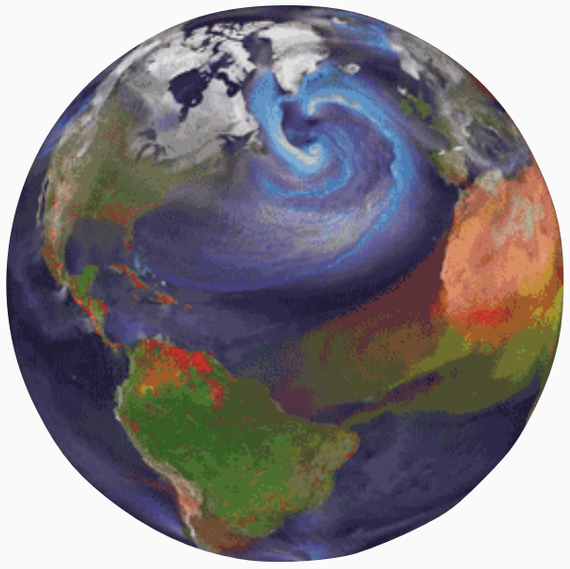
R&D at Juelich SSC:
4x10⁶ GPU hours granted in 2023

JÜLICH Forschungszentrum

Applications: one model for multiple purposes

Use the learned representation to improve the state-of-the-art of specific weather & climate-related scientific applications

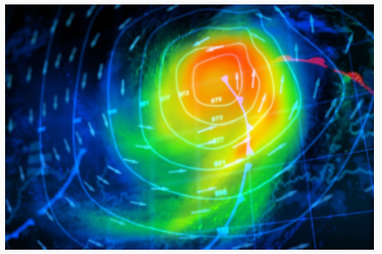
Spatio-temporal representation of atmospheric dynamics



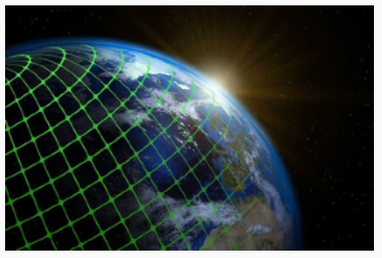
Model
given by the trained neural network

Task dependent Adaptable smaller networks

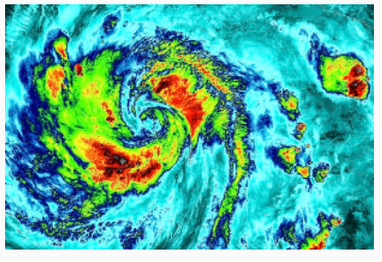
Adaptation



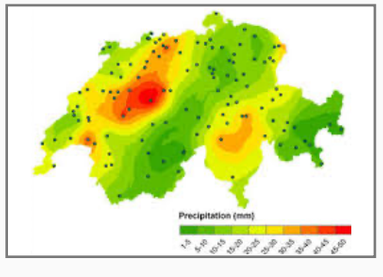
Weather predictions



Downscaling



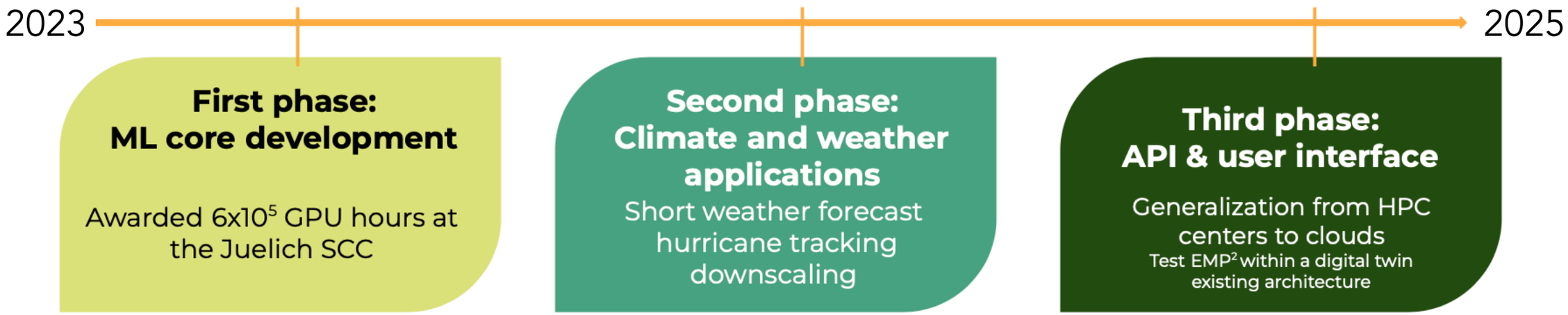
Bias corrections



Spatio-temporal Interpolations (WIP)

Ilaria Luise, CERN - ilaria.luise@cern.ch

Main Objectives



Completed

- First prototype of ML core model **released in August 2023**
- Model publicly **available on GitHub**
- More infos at: **www.atmorep.org**

100% completed

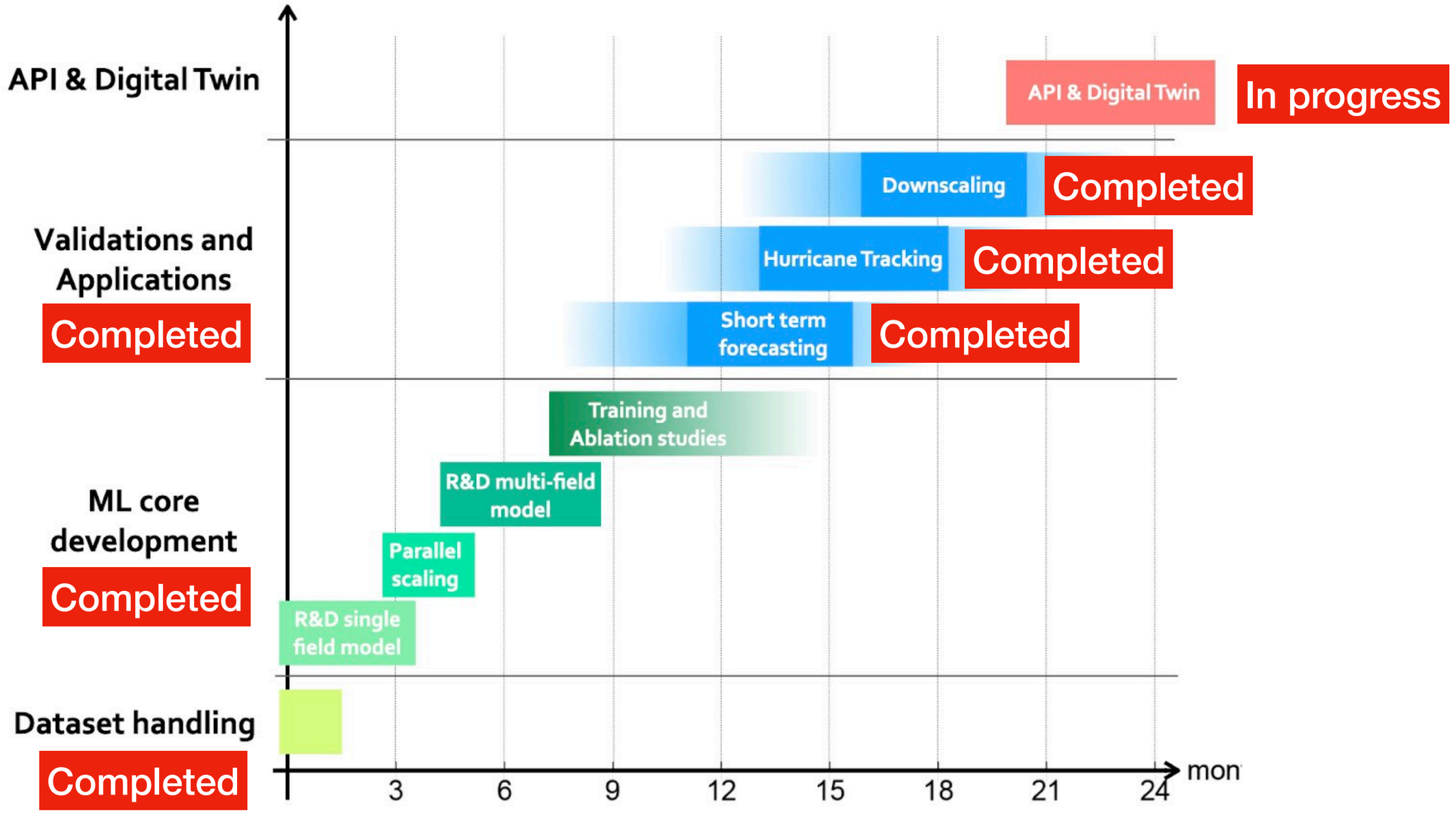
- **hurricane tracking** application
- Added **spatio-temporal interpolation** task.
- Interest towards **longer prediction windows** and **climate modelling** (see later)
- Interest towards **learning from raw data** tasks (ECMWF)

In progress

- Closely **following the interTwin developments** (weekly meetings)
- Integration of EURAC use case within interTwin (simpler use case)
- Definition of what the environmental platform concretely means

Ilaria Luise, CERN - ilaria.luise@cern.ch

Expected timeline



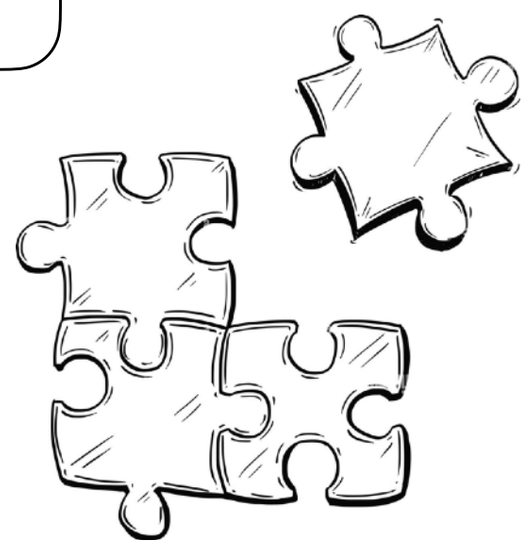
Ilaria Luise, CERN - ilaria.luise@cern.ch

Innovation factors

One of the first proof-of-concepts of a global machine-learning based environmental model trained on petabytes of observational data.

First example of **large-scale spatio-temporal representation learning (3D space+time)** of atmospheric dynamics.

Explore the potential of foundation models in fundamental science.
Apply representation learning to scientific data. Useful e.g. for HL-LHC



Applications:
Faster weather prediction models. Help shaping the next generation of weather prediction models

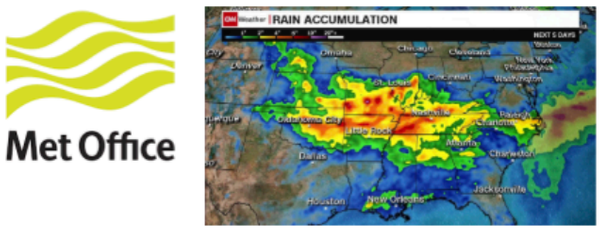
Innovative probabilistic approach implemented for this kind of models.
Leads to well calibrated ensemble outputs (no perturbative approach as currently used).

Applications:
First step towards developing **data-driven systems that learn directly from observations.**

Ilaria Luise, CERN - ilaria.luise@cern.ch

Expected impact

Slide from 2022 from the KT fund application



Integration into current weather forecast models

Potential first user:
EU-funded project MAELSTROM

Licensing to insurances and risk assessment companies

eg. assess floods/droughts/fires risks under climate change

Partnerships in future EU funded projects

prototype of a digital twin encapsulating information from a large dataset ⇒ expertise applicable to other domains

Renewables-based energy companies

e.g. optimize wind or solar farm placement, forecast the productivity of renewable-based power plants

Large high-tech companies

integration of machine learning-based models for weather and climate forecasting into their products

Ilaria Luise, CERN - ilaria.luise@cern.ch

Expected impact

It didn't go through because CERN is not a *national* organisation

Integration into current weather forecast models

Potential first user:
EU-funded project MAELSTROM

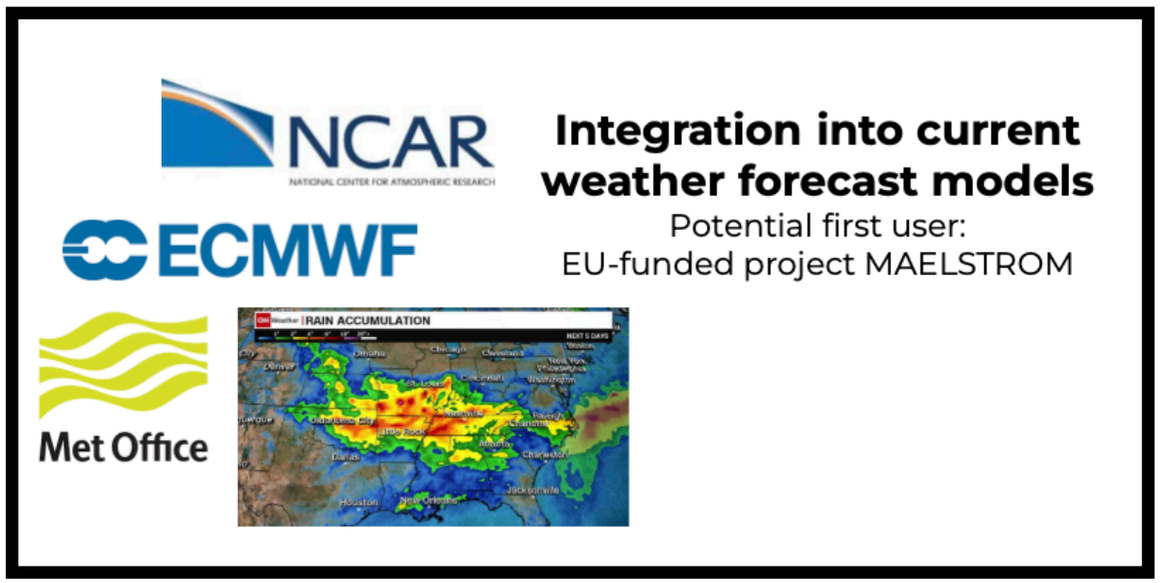
We tried a link with DestinE:

- Invitation to Tender **DE371:**

Large-scale atmospheric representation learning for ensemble generation and time interpolation

Need to check if the caveat was removed for DestinE phase 2

Final impact



Partnerships in future EU funded projects

prototype of a digital twin encapsulating information from a large dataset ⇒ expertise applicable to other domains

Projects that will be using AtmoRep as core model:

- **MAELSTROM:** Granted 4M core hours at JSC to use the EMP2 core model for the Downscaling (super-resolution) task within MAELSTROM

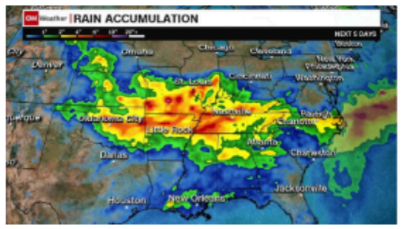
- **Call for Pilot Foundations: Helmholtz Foundation Model Initiative**
  - Partners: JSC (leader), AWI, KIT
  - External partners: ECMWF, CERN
  - HClimRep project proposal:
    - Use AtmoRep as core model for climatological predictions
    - No extra funds/FTE needed.

- **WarmWorld project**: AtmoRep will be adapted to the ICON grid (used by the German Weather Service, DWD) and used as core model for AI-based weather predictions.

Final impact

Integration into current weather forecast models

Potential first user:
EU-funded project MAELSTROM



Destination Earth

Our planet's digital twin

The role of ECMWF

Partnerships in future EU funded projects

prototype of a digit
encapsulating information from a large dataset ⇒ expertise applicable to other domains

Projects that will be using AtmoRep as baseline:

- **WeatherGenerator**

WeatherGenerator project aims to recast machine learning for Earth system modelling

2 October 2024

Share

ECMWF is coordinating an EU Horizon project called WeatherGenerator which aims to use machine learning in novel ways for weather forecasting and to model related Earth system processes.

The developments in the WeatherGenerator will feed into the digital twins implemented by ECMWF in the EU's Destination Earth (DestinE) initiative, in which we are one of three entrusted entities alongside ESA and EUMETSAT. It is anticipated that it can also be used to supplement the Centre's standard weather forecasts.

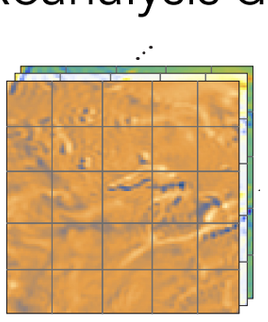
Link to full article

Ilaria Luise, CERN - ilaria.luise@cern.ch

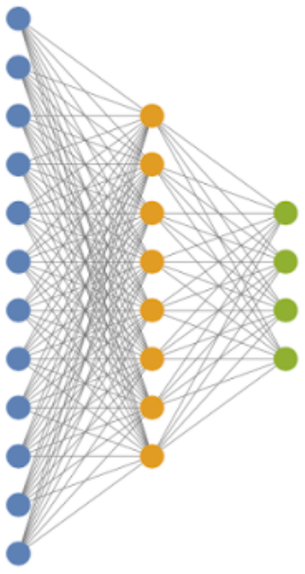
Present: The AtmoRep core model

Physics-related applications = uncertainties → *Need for a* ***stochastic approach***

Reanalysis data



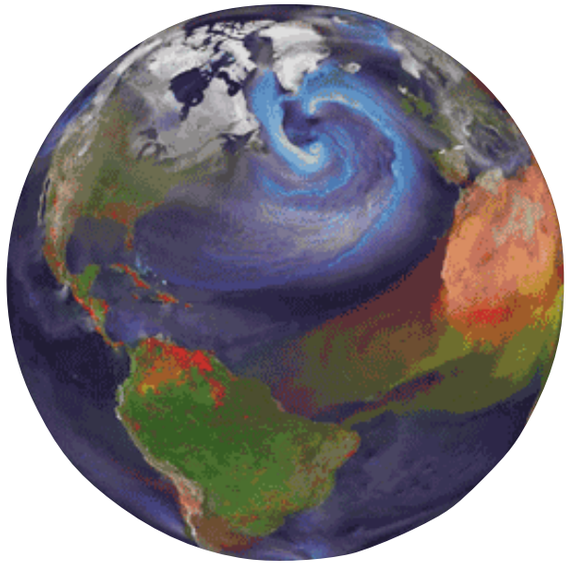
50 TB



large scale
machine learning

Transformers architecture

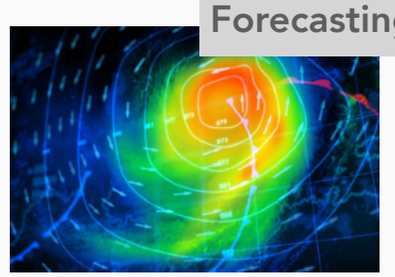
Spatio-temporal representation of atmospheric dynamics



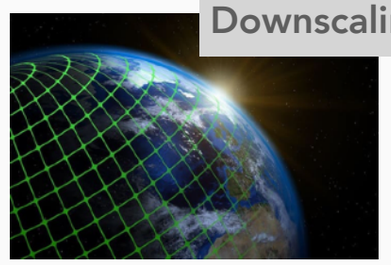
Model
given by the trained neural network

Adaptation

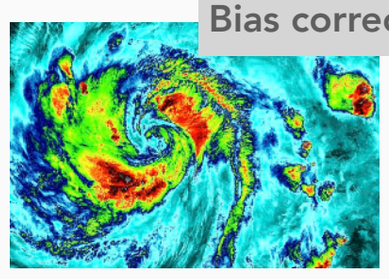
Forecasting



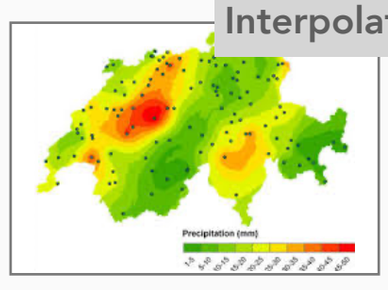
Downscaling



Bias corrections



Interpolations



C. Lessig, I. Luise et al. "AtmoRep: A stochastic model of atmosphere dynamics using large scale representation learning"

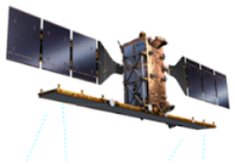
Ilaria Luise, CERN - ilaria.luise@cern.ch

Future: the WeatherGenerator

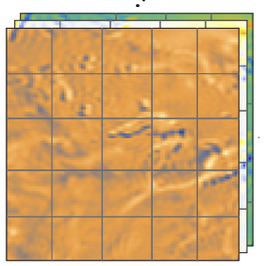
WeatherGenerator

Courtesy of the WeatherGenerator collaboration

Satellite data

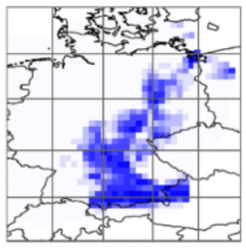


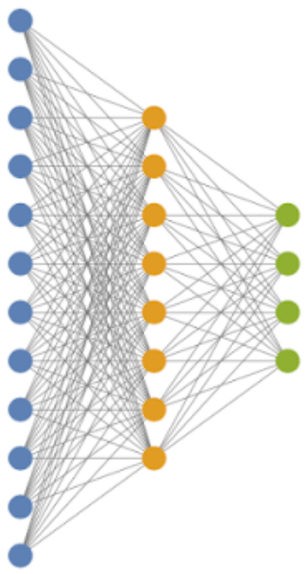
Reanalysis data



Simulations

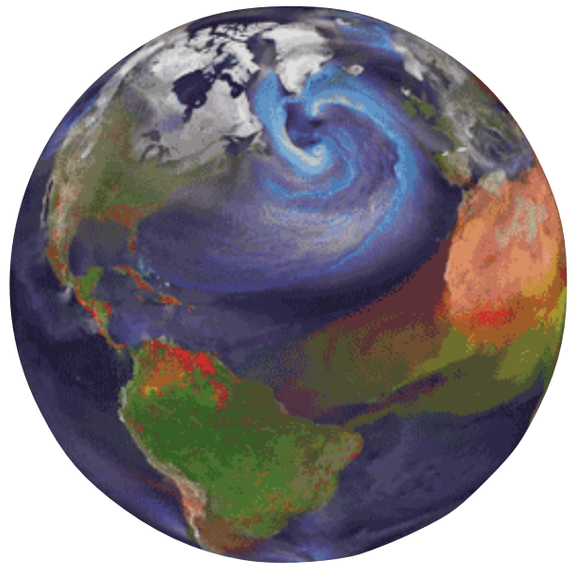
Other sources





large scale machine learning

Spatio-temporal representation of atmospheric dynamics



Model
given by the trained neural network

seamless prediction

- nowcasting
- medium-range
- decadal projections

best state estimate

- data-assimilation
- reanalysis
- ...

high-impact applications

- energy output
- extreme event risk
- ...

Ilaria Luise, CERN - ilaria.luise@cern.ch

Final impact

CERN teams up with Luxembourg and the World Food Programme to tackle global hunger

The new partnership aims to leverage state-of-the-art AI technology to speed up progress towards a world with zero hunger

15 OCTOBER, 2024 | By Antonella Del Rosso

Representatives from WFP, LIST and CERN met today in Luxembourg to sign a new strategic partnership that aims to reduce world hunger. (Image: MAE Luxembourg)

WP1: the AtmoRep model will be used to forecast seasonal predictions of atmospheric variables for crop yield predictions

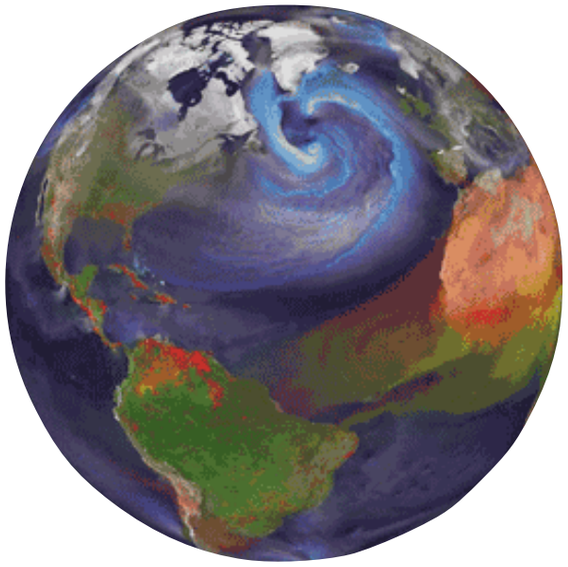
Ilaria Luise, CERN - ilaria.luise@cern.ch

Relationships with other projects

Future: develop the API & the user interface

Integration within the digital twin engine being developed by the interTwin project

Challenging part:
Close collaboration with the members of InterTwin & CS4OD projects at CERN

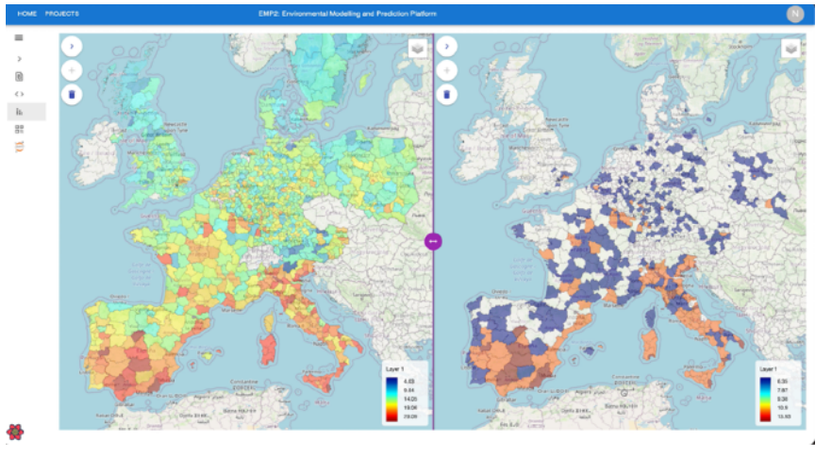


Onboard

(New CS4OD)

interTwin

Final product:
Prototype of a user oriented platform for environmental applications



1) **integration of a much simpler environmental use case** from EURAC (Trento) to understand the pipelines
2) Definition of the **concept of platform** and definition of the **roadmap** to achieve it
3) **Summer students** from NMS program + OpenLab summer program will work on the project this summer

Ilaria Luise, CERN - ilaria.luise@cern.ch

The platform

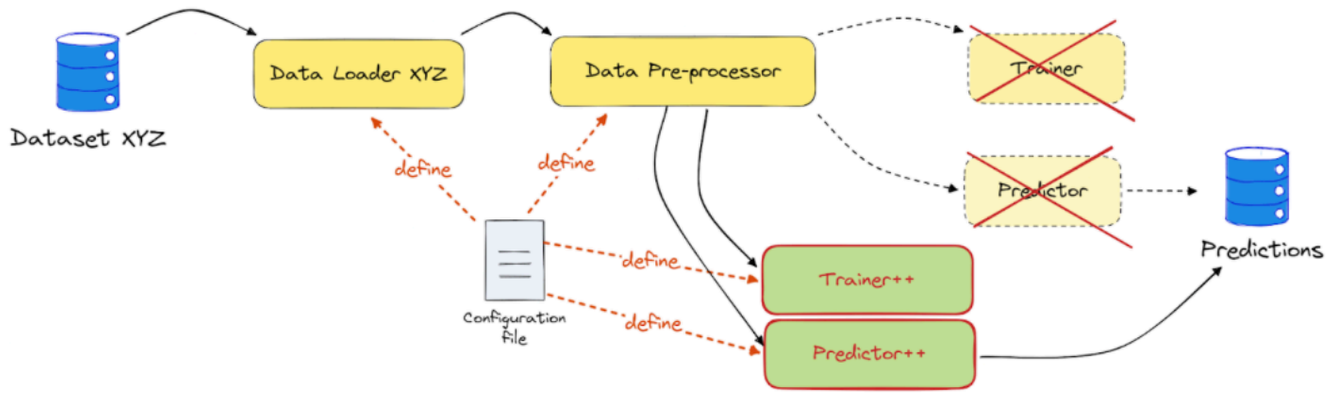
itwinai Plugins (6)

Plugins allow *anyone* to reuse existing components at low cost, without reinventing the wheel, and improving over them.

How does it work, *in practice*?

- Install **`itwinai`** core library (toolkit):
  - **`pip install itwinai`**
- Install use case **plugins**:
  - **`pip install itwinai-cern-gan`**
  - **`pip install itwinai-hydromt`**
  - **`pip install itwinai-virgo`**
  - **`...`**

Yellow boxes are pre-existing modules which are reused



Slide from Matteo Bunino

Results: Role of CERN

What has been done at CERN:

- Implemented from scratch the **analysis chain** to ensure a robust evaluation of the training protocol after each iteration
- working on the **ensemble uncertainty definition** and quantification.
- **coordinating the CERN-JSC analysis team** for the downstream applications.
- leading role in the **writing process of the first AtmoRep/EMP2 paper**, which is foreseen to be published by the **end of summer**.
- **Supervision of the OpenLab summer student** David Hidary in co-supervision with Christian Lessig from the University of Magdebug (2023), supervision of NMS programs summer student Dorcas Mulaye (2024).
- **Definition of environmental platform** for benchmarking within interTwin and **integration of the EURAC use case** for ML-based droughts predictions within the interTwin pipeline.
- **Ablation studies and definition fo AtmoRep 1.0,** which is more efficient and more accurate
- GitHub maintainance of both model and analysis repos.
- External Partner in the HClimRep project, supervision and follow up of technical requests for HClimRep

Ilaria Luise, CERN - ilaria.luise@cern.ch

Public Talks - 2024

- I. Luise, **Incontri di Fisica**, INFN Frascati, invited lecturer, November 2024.
- I. Luise, **University of Bern**, invited seminar, October 2024
- I. Luise, **ELLIIT Symposium,** invited visitor scientist, October 2024.
- I. Luise, **ICES biennial workshop**, invited speaker, October 2024.
- I. Luise, **AI4Science Forum**, invited seminar, June 2024.
- I. Luise, **University of Clermont-Ferrand**, invited seminar, June 2024.
- I. Luise, **Helmholtz AI conference**, satellite event, invited talk, June 2024.
- I. Luise C. Lessig, **PASC 2024**, invited talk + organisers of a mini-symposium on "Foundation models in Earth System Science", June 2024.
- I. Luise, **ENES HPC Workshop on High resolution modeling in Earth system science**, invited talk, May 2024.
- I. Luise, **ESA-ECMWF Machine Learning Workshop**, talk and panelist, May 2024.
- C. Lessig, **Karlsruhe Institute of Technology**, invited seminar, May 2024.
- C. Lessig, **EGU24**, solicited talk, April 2024.

Public Talks - 2023

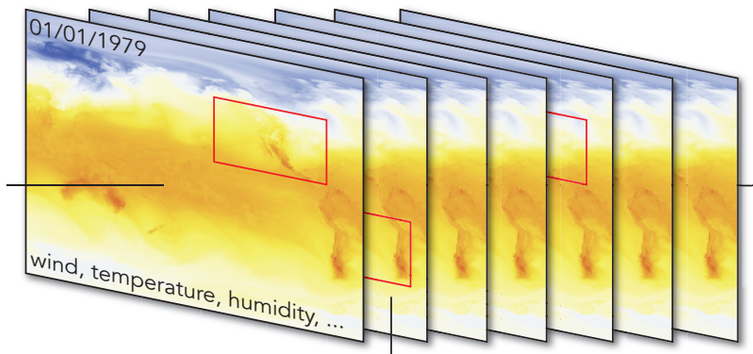
- I. Luise, **CERN OpenLab technical workshop**, April 2023.
- I. Luise, **NEURIPS 2023**, poster, November 2023.
- C. Lessig, **Google Geo Sustainability**, invited talk, August 2023.
- C. Lessig, **NVIDIA Reseach,** invited talk, May 2023; Slides.
- C. Lessig, ECWMF, invited talk, May 2023; Slides.
- I. Luise, **CERN-Microsoft Workshop**, April 2023.
- C. Lessig, EGU, April 2023; Slides.
- I. Luise, CERN OpenLab technical workshop, March 2023; Slides.
- C. Lessig, IntelliAQ workshop, March 2023; Slides.
- C. Lessig, NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography, February 2023; Slides.
- C. Lessig, **Google Research**, October 2022; Slides.
- I. Luise, **DWD KI Forum,** October 2022; Slides.

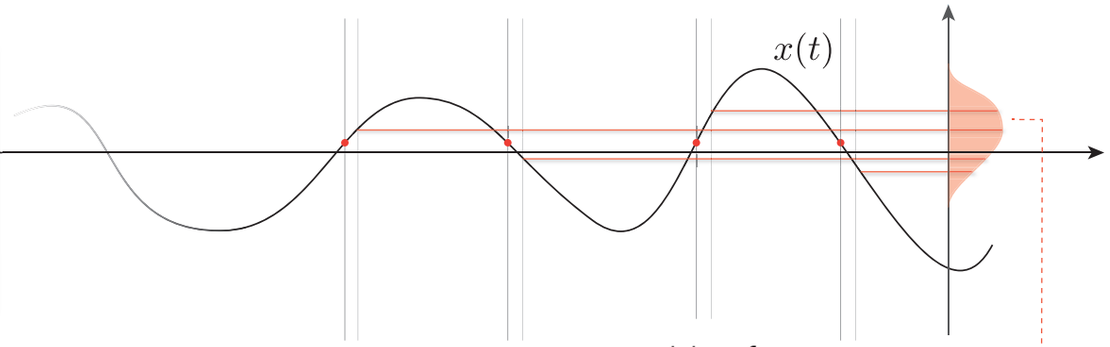
Ilaria Luise, CERN - ilaria.luise@cern.ch

Results

The AtmoRep workflow

pre-processed historical observational record $x(t)$ (ERA5 reanalysis)





sampling

36h x 5 levels x 1800 km x 3600 km

Large masking ratios above 80%

local space-time neighborhood

t

AtmoRep

$p_\theta\left(y|x,\alpha\right)$

ensemble of tail networks

ensemble prediction

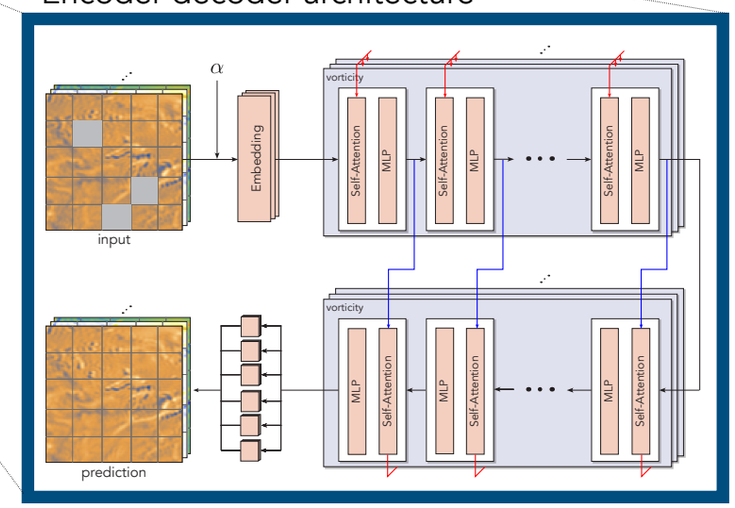
statistical loss

New stochastic approach

ensemble predictions with 16 members

Approximate the 4-Dim PDF of the process using a Transformers-based network with 3.5 billion parameters

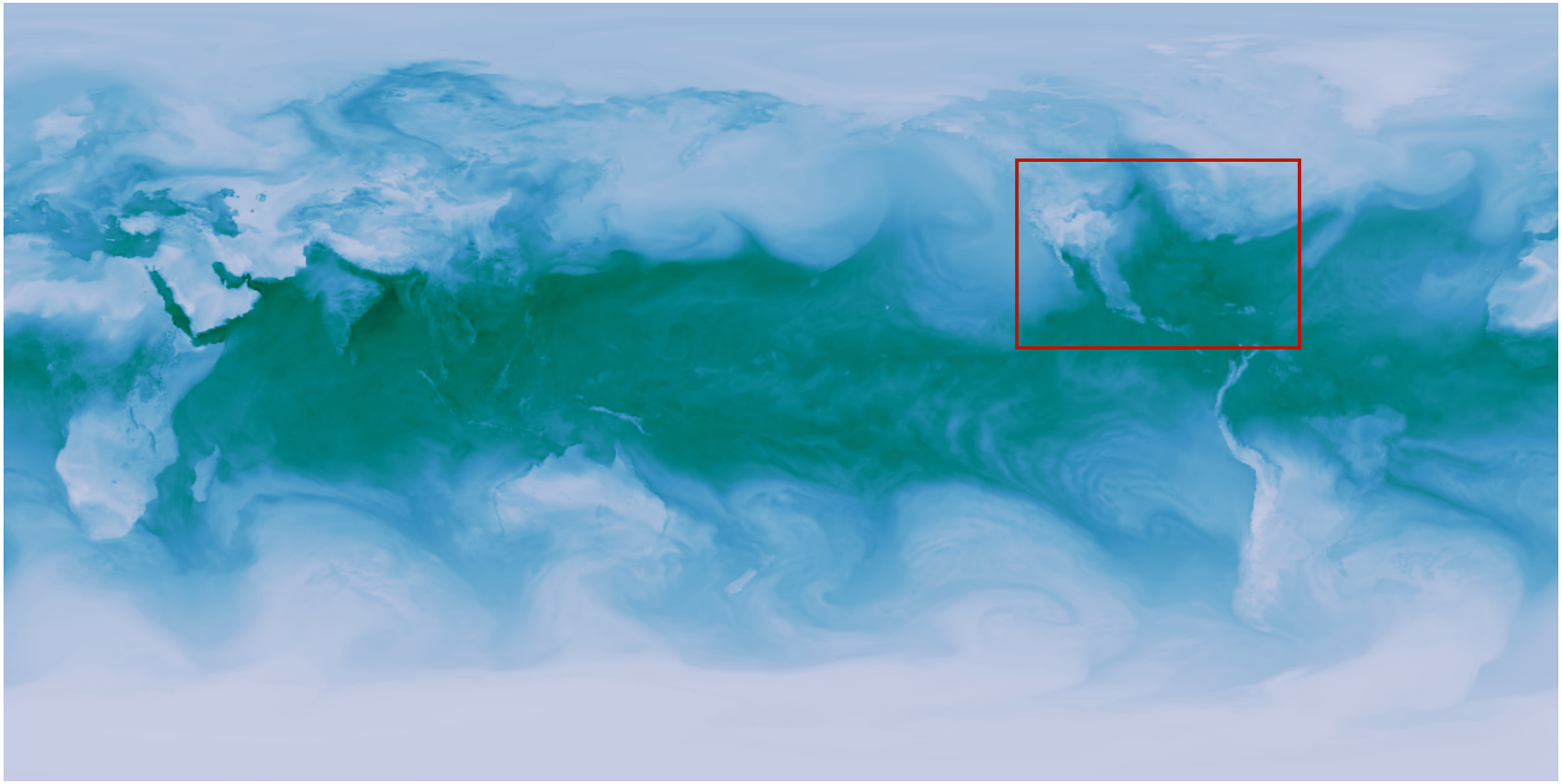
Encoder decoder architecture



Ilaria Luise, CERN - ilaria.luise@cern.ch

Results: Target - ERA5

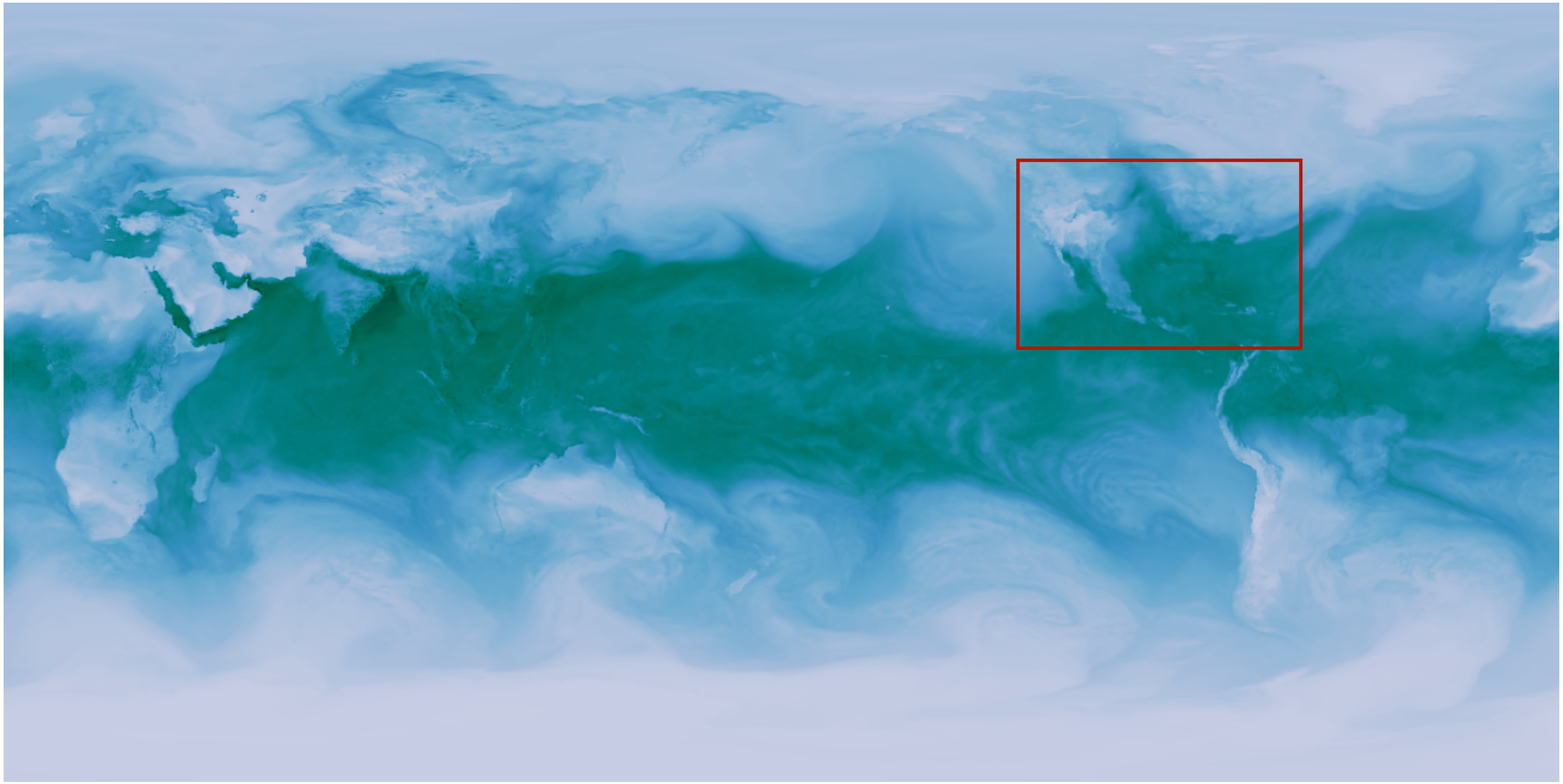
specific humidity, June 15th 2018 13:00 UTC



Ilaria Luise, CERN - ilaria.luise@cern.ch

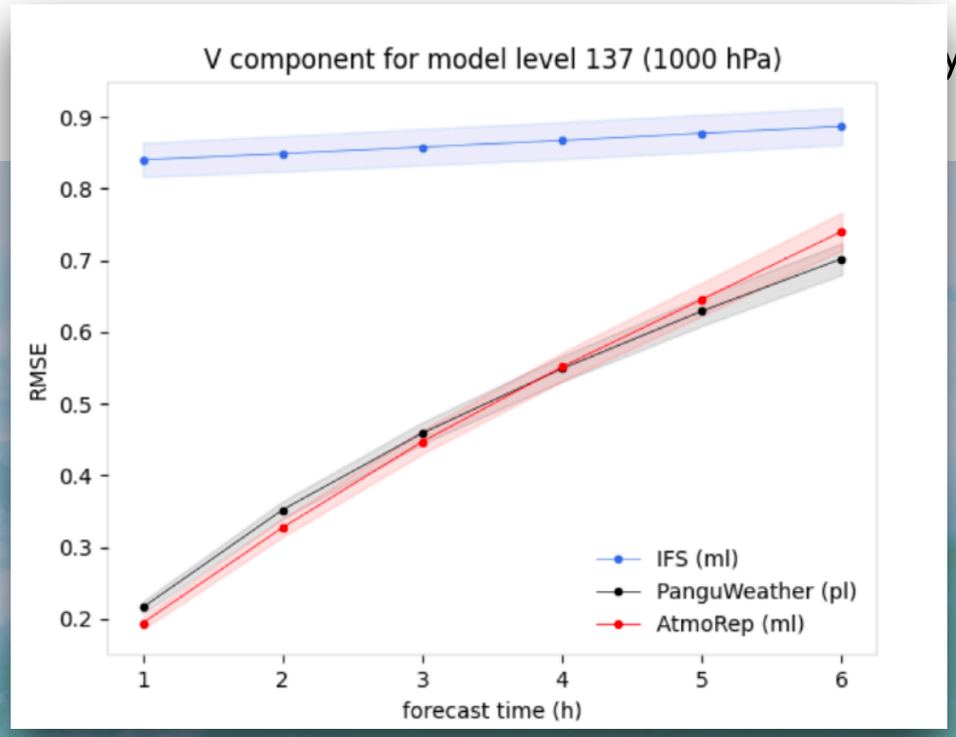
Results: Prediction

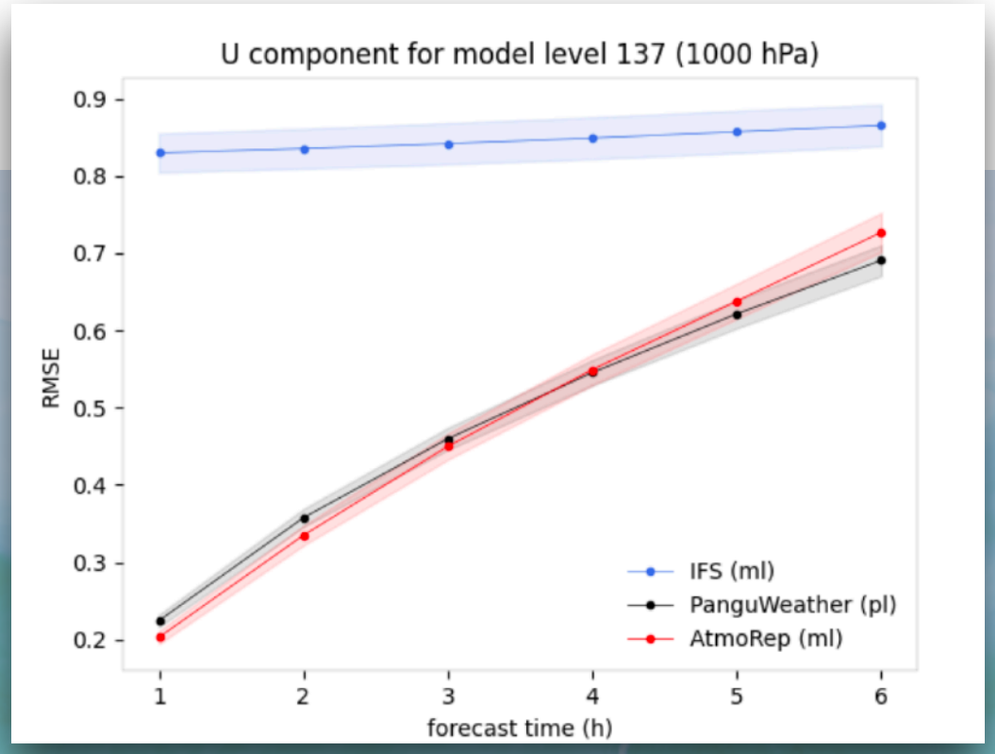
specific humidity, June 15th 2018 13:00 UTC

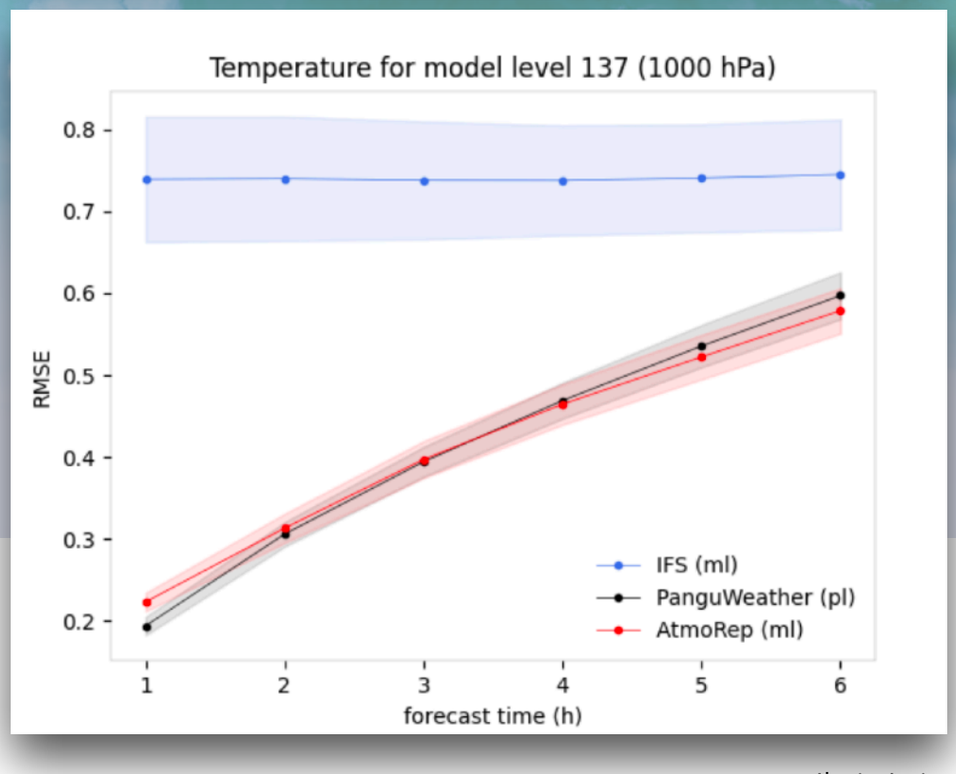


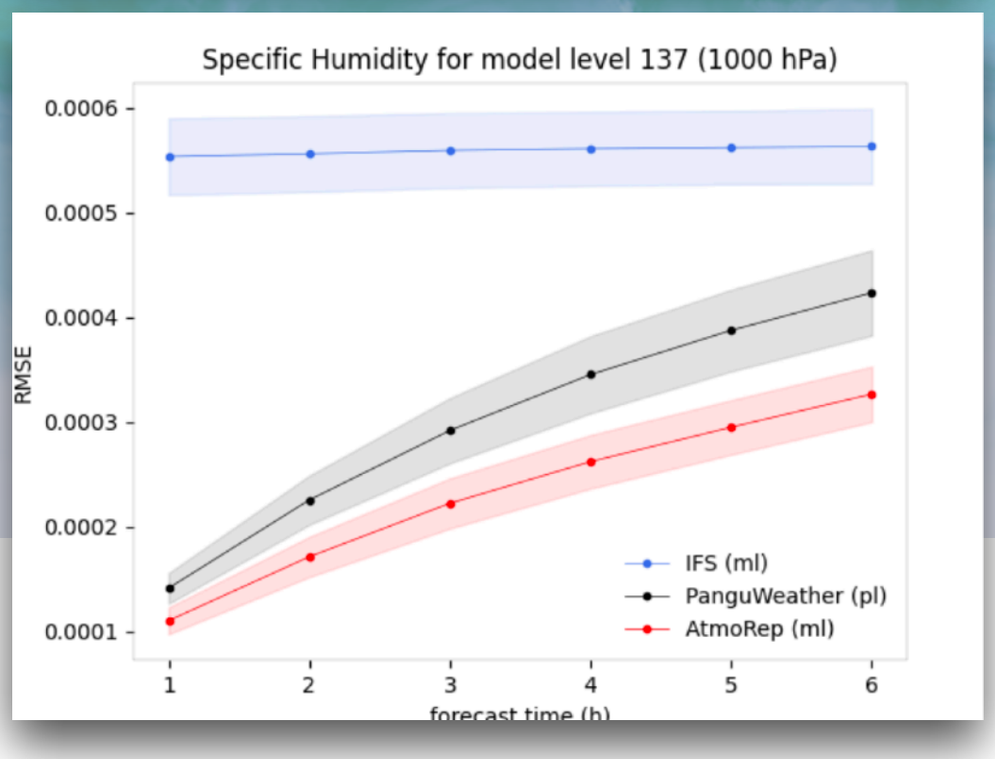
Ilaria Luise, CERN - ilaria.luise@cern.ch

Results: Prediction





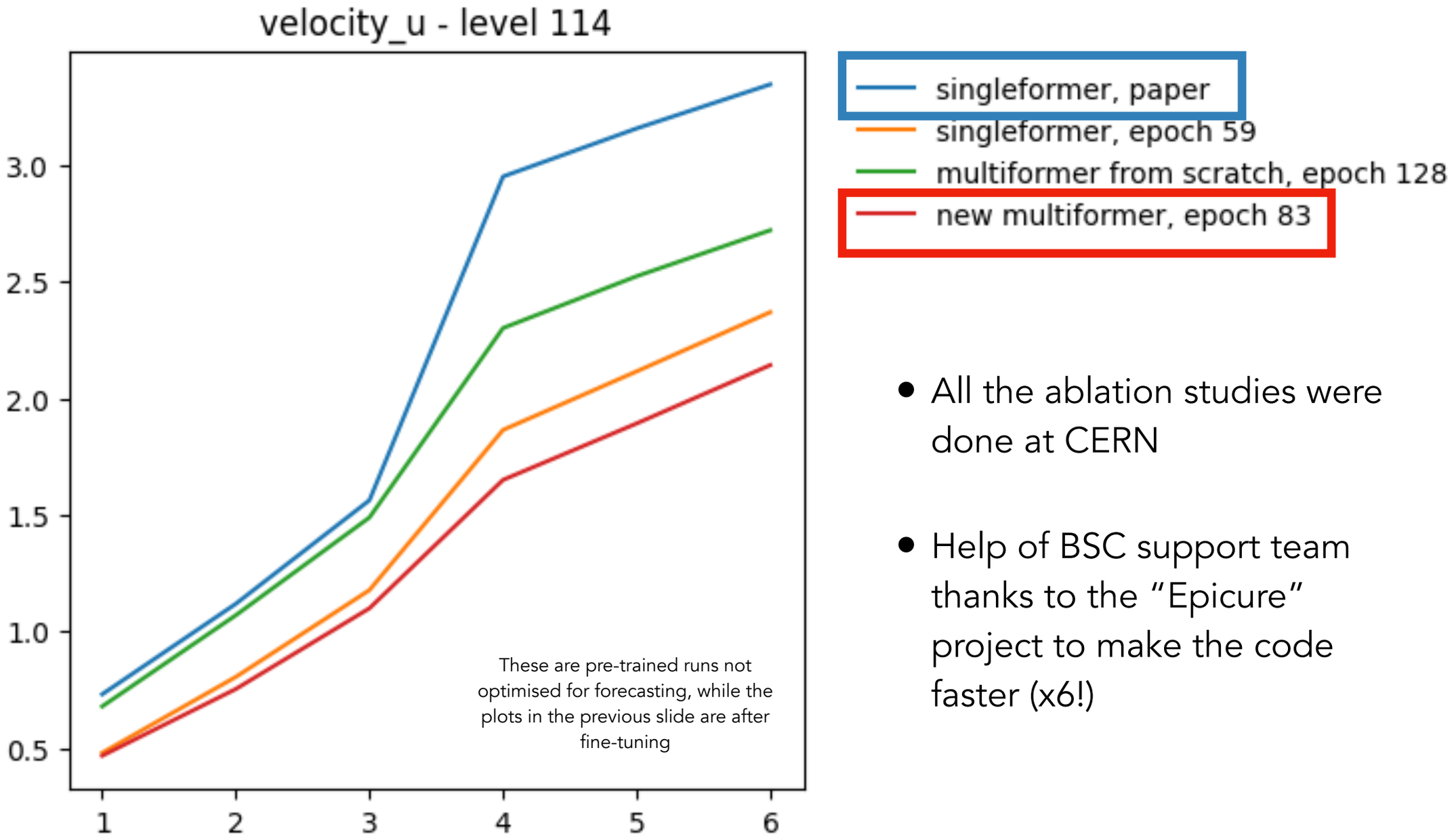




Results: AtmoRep 1.0

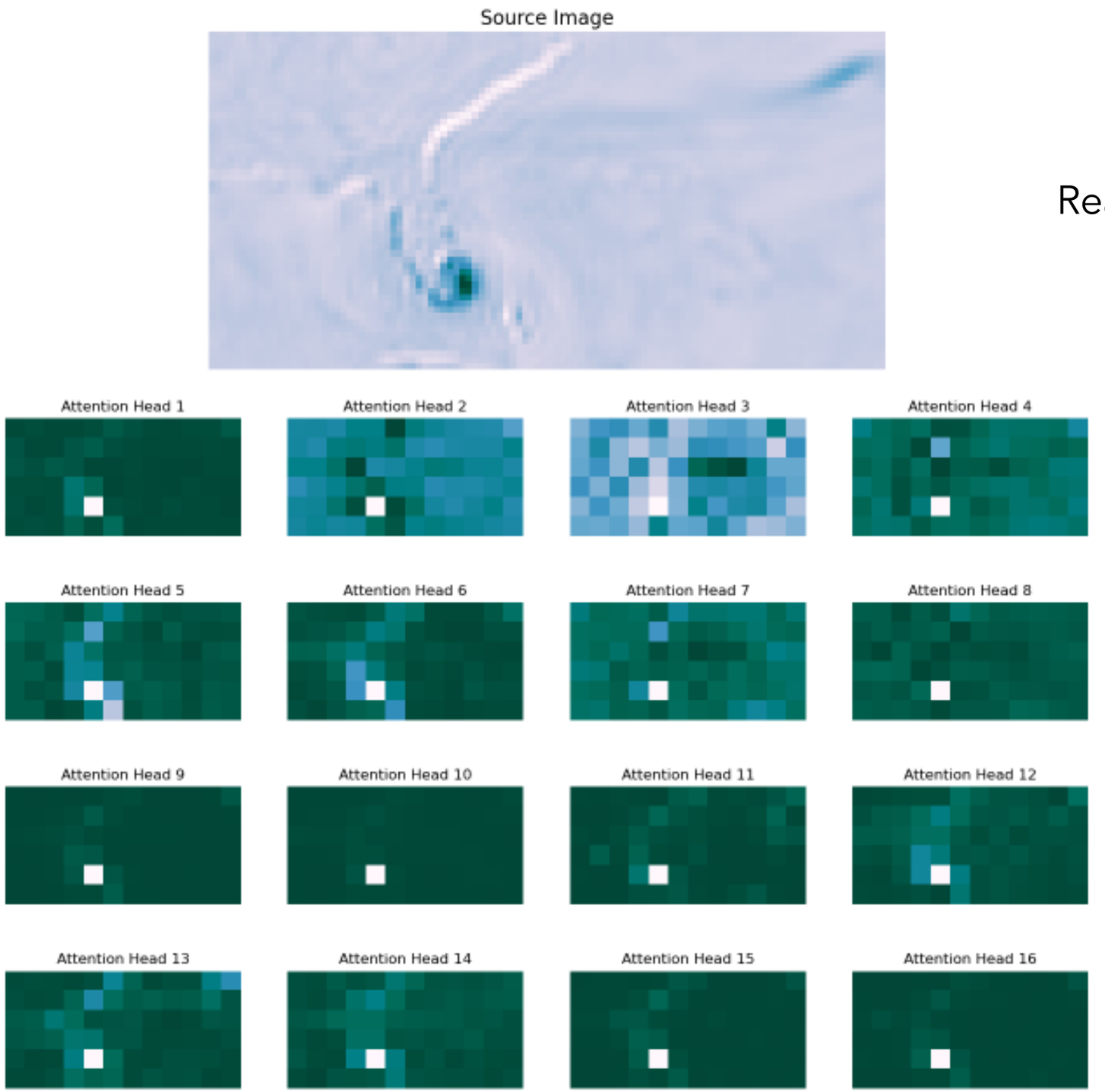
AtmoRep 0.1: 3.5B parameters

AtmoRep 1.0: 600M parameters, x6 faster & more accurate



- All the ablation studies were done at CERN
- Help of BSC support team thanks to the “Epicure” project to make the code faster (x6!)

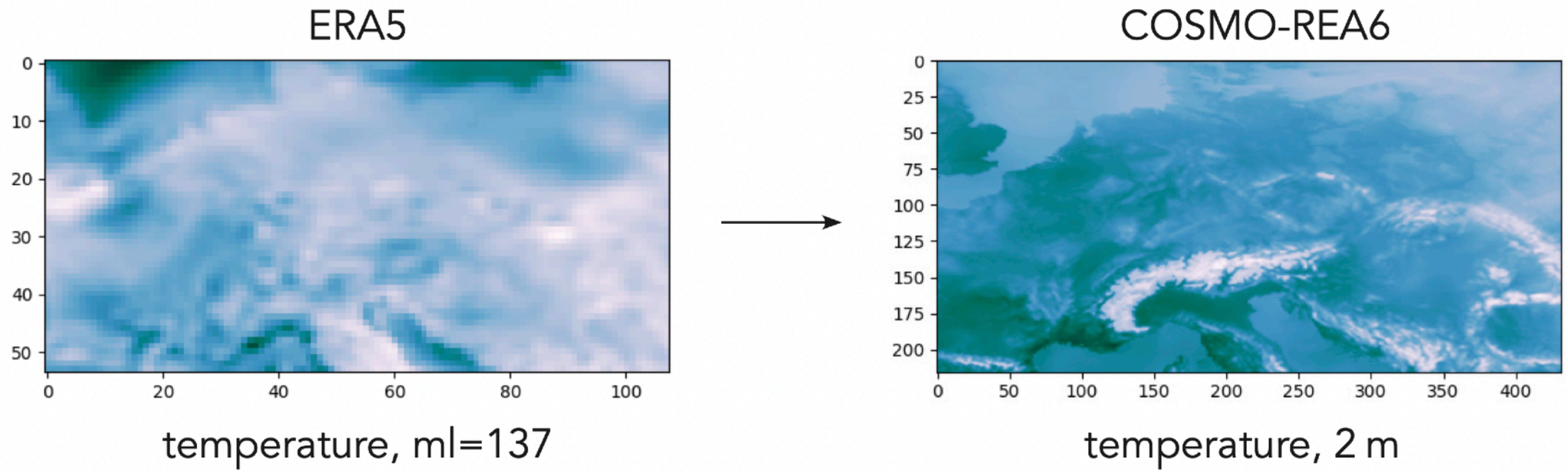
Results: Hurricane tracking



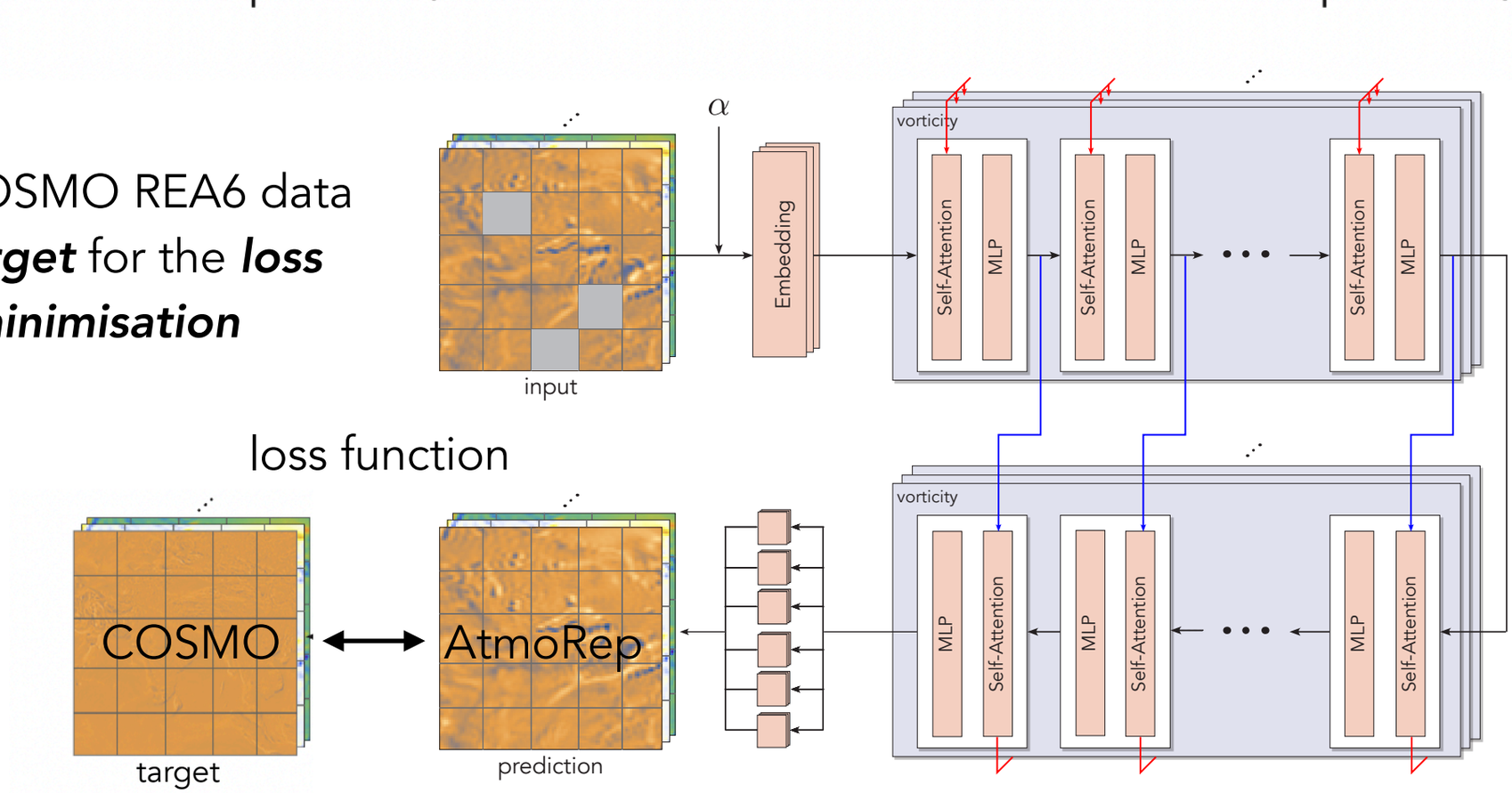
Read full report here

Ilaria Luise, CERN - ilaria.luise@cern.ch

Downscaling

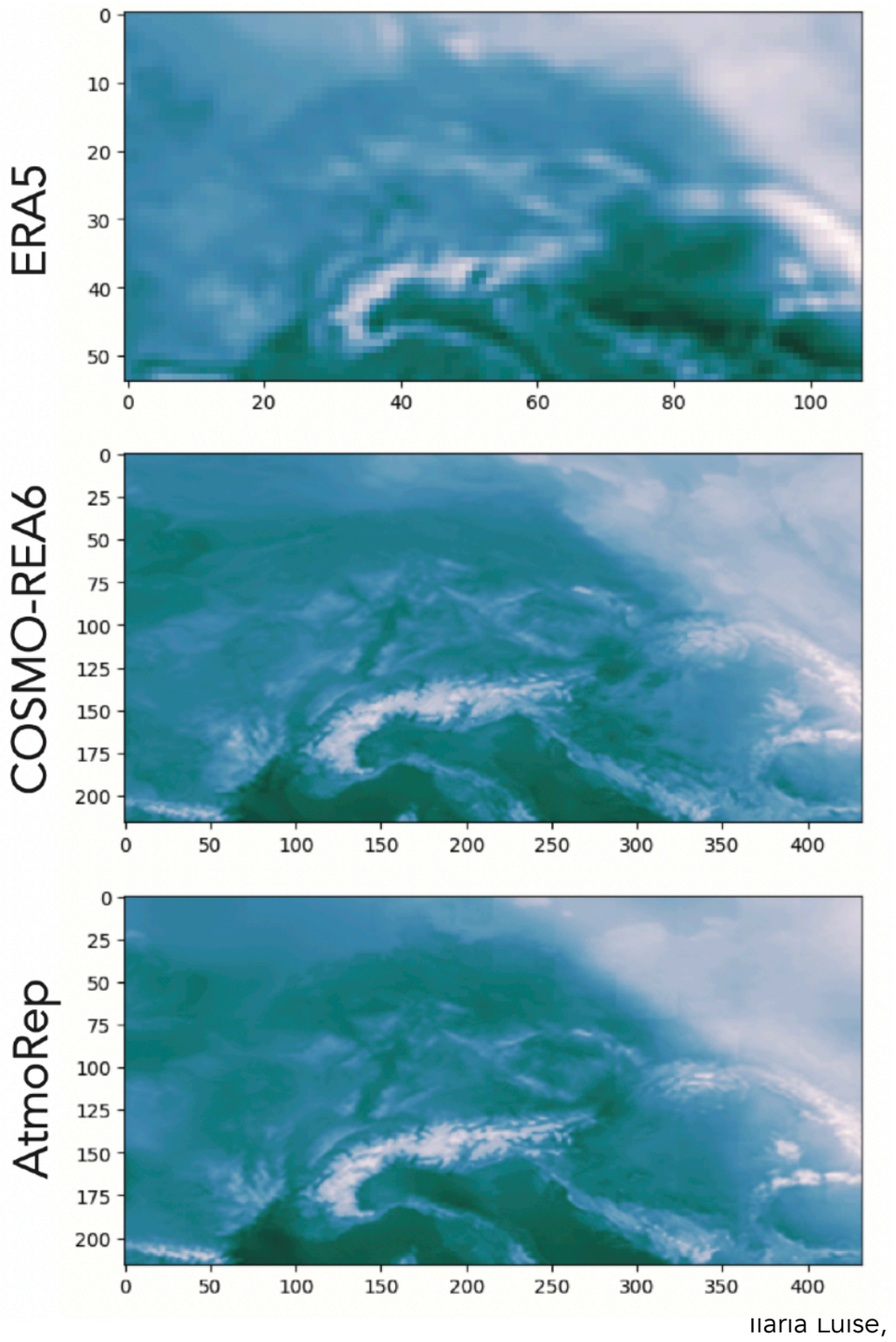


Use COSMO REA6 data as ***target*** for the ***loss minimisation***

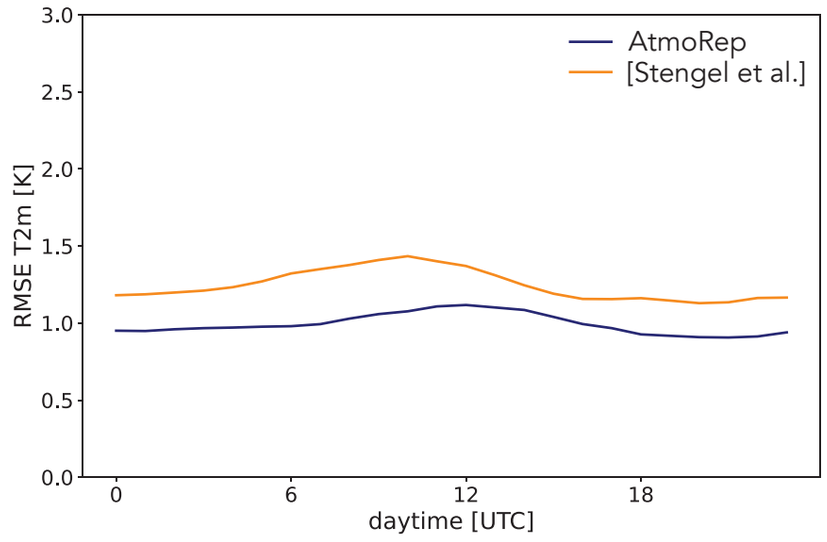


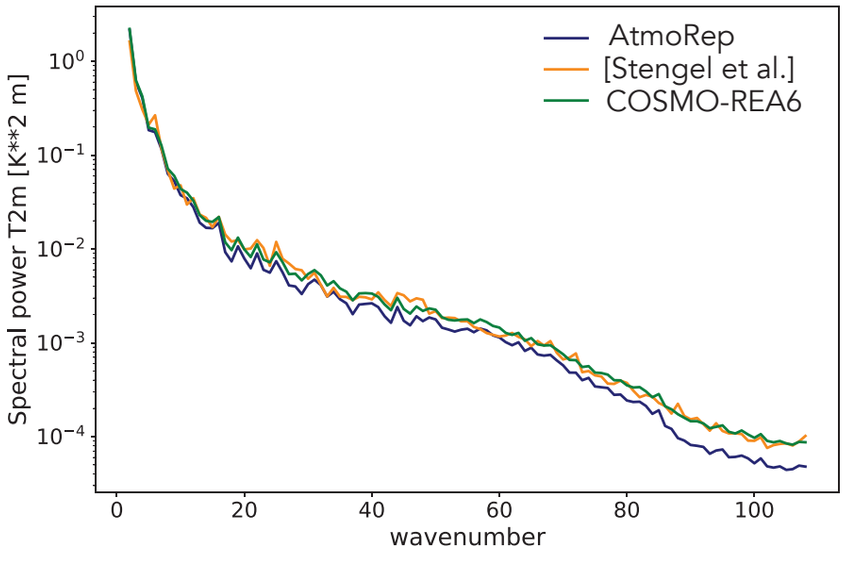
Downscaling

Use the COSMO REA6 dataset (6 km resolution vs ~32 km in ERA5) to create a downscaled version of AtmoRep



Comparison with a competing AI-based model for downscaling:

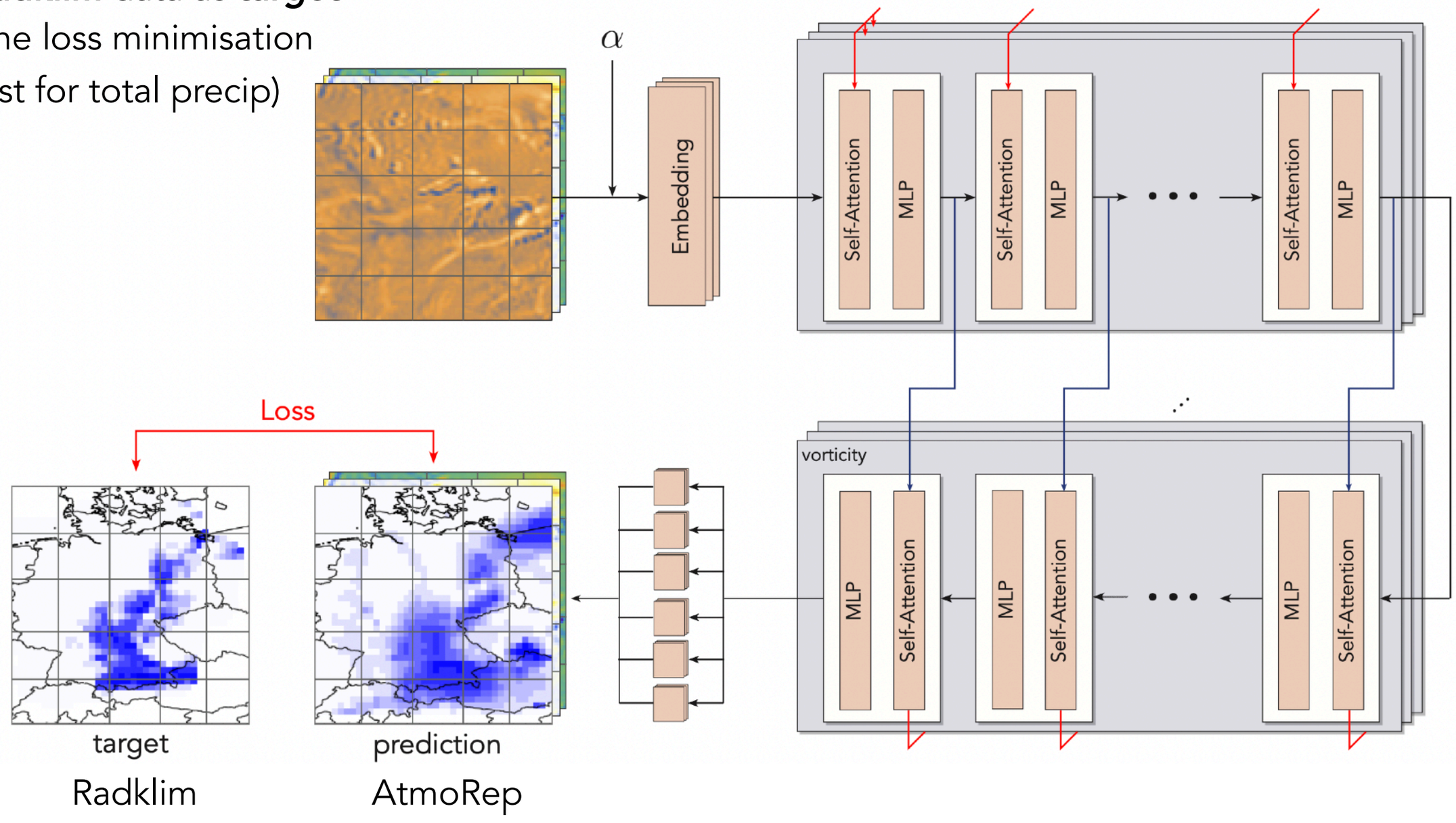




Bias corrections

Precipitation rates are known to be suboptimal in ERA5
Use RADKLIM radar data to fine-tune the precipitation rates in AtmoRep

Use ***Radklim*** data as ***target***
for the loss minimisation
(just for total precip)

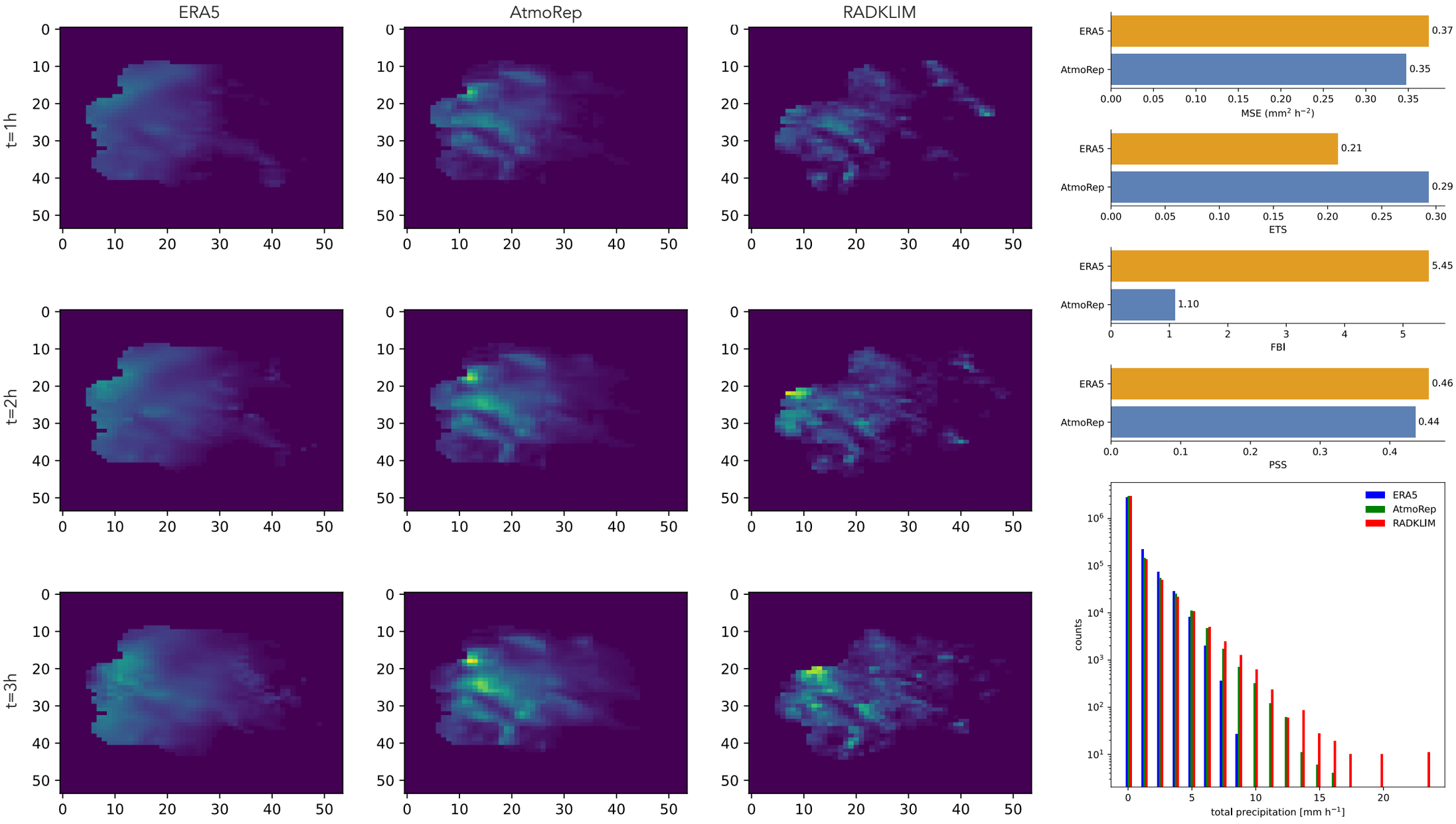


Ilaria Luise, CERN - ilaria.luise@cern.ch

Bias corrections: Results

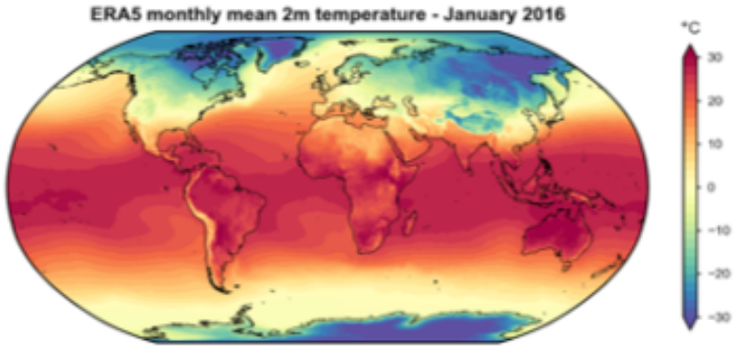
Precipitation rates are known to be suboptimal in ERA5

Use RADKLIM radar data to fine-tune the precipitation rates in AtmoRep



Ilaria Luise, CERN - ilaria.luise@cern.ch

Backup

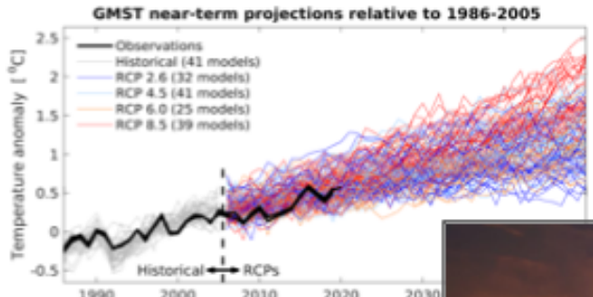


Operations

Technology

Collect

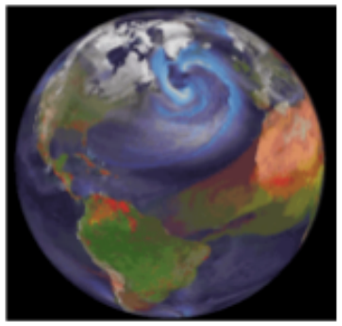
Aware



Artificial Intelligence

Big Data Analytics

Compute

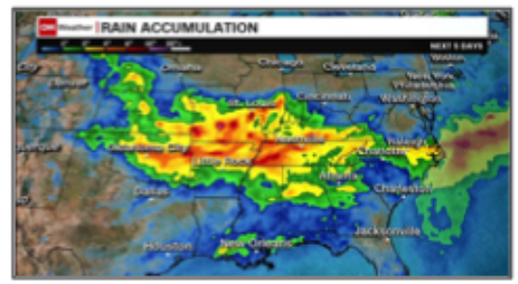


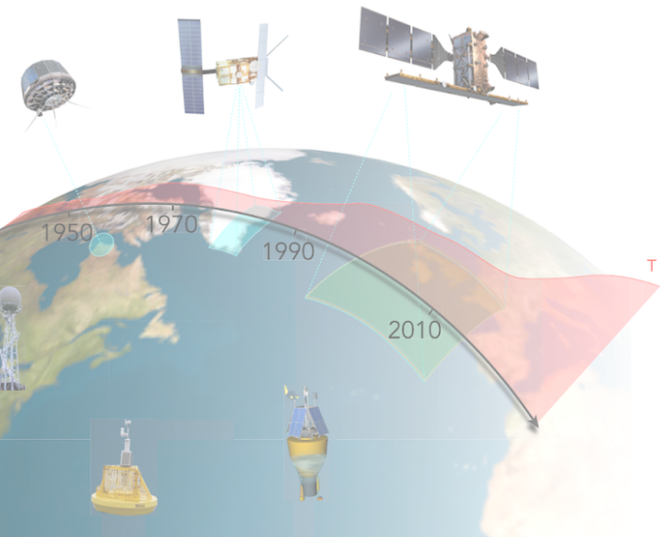
Respond

Visualization Tools

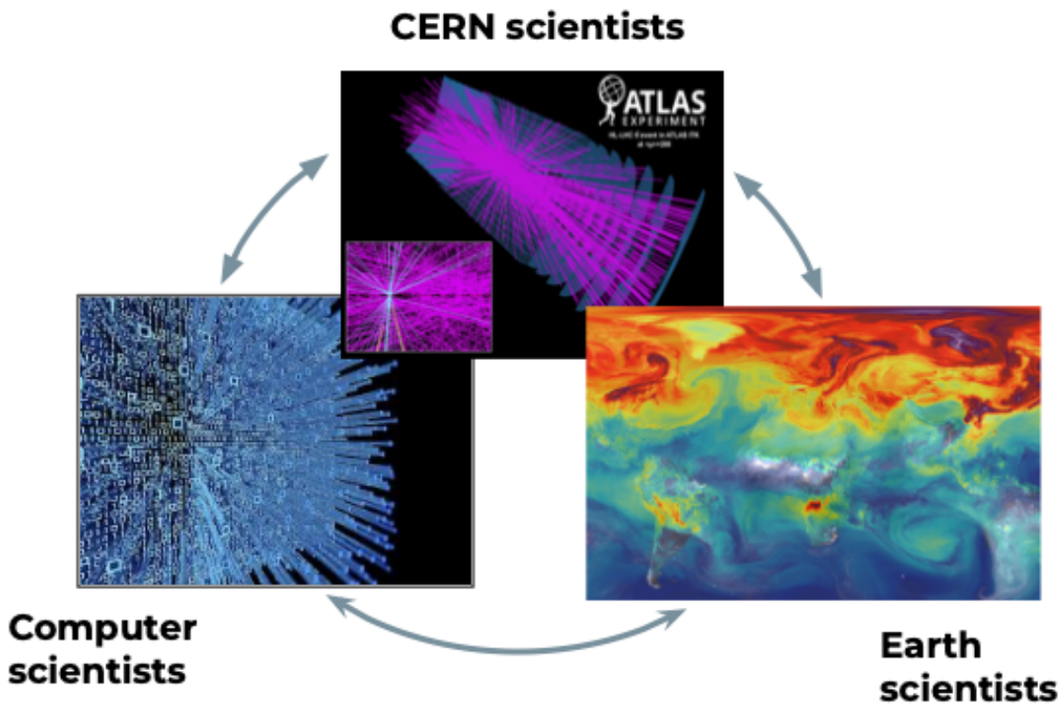
Visualize

Predict





Introduction



Why CERN?

Solve common scientific challenge(s) in high-energy physics and weather/climate science using AI/ML

Model complex, nonlinear phenomena and improve current simulations

Access multi-scale dependencies of a given process

Earth science: eg. better understand convection phenomena

CERN: eg. particle-jet showers reconstruction

Condense dataset information in a compact representation

better handle the information in downstream applications.

eg. condense the info in a few GB rather than TB

Explore potential of unsupervised learning for scientific applications

Extract new information directly from data

eg. learn unknown correlation patterns

Earth science: eg. early detection of extreme events

CERN: eg. anomaly detection

Common Goal:

Develop a proof of concept of representation learning for scientific applications based on observations