

EMP²:

Environmental Modelling and Prediction Platform

Christian Lessig, [Ilaria Luise](#), Martin Schultz,
Michael Langguth, Alberto di Meglio, Maria Girone, Sofia Vallecorsa et al.

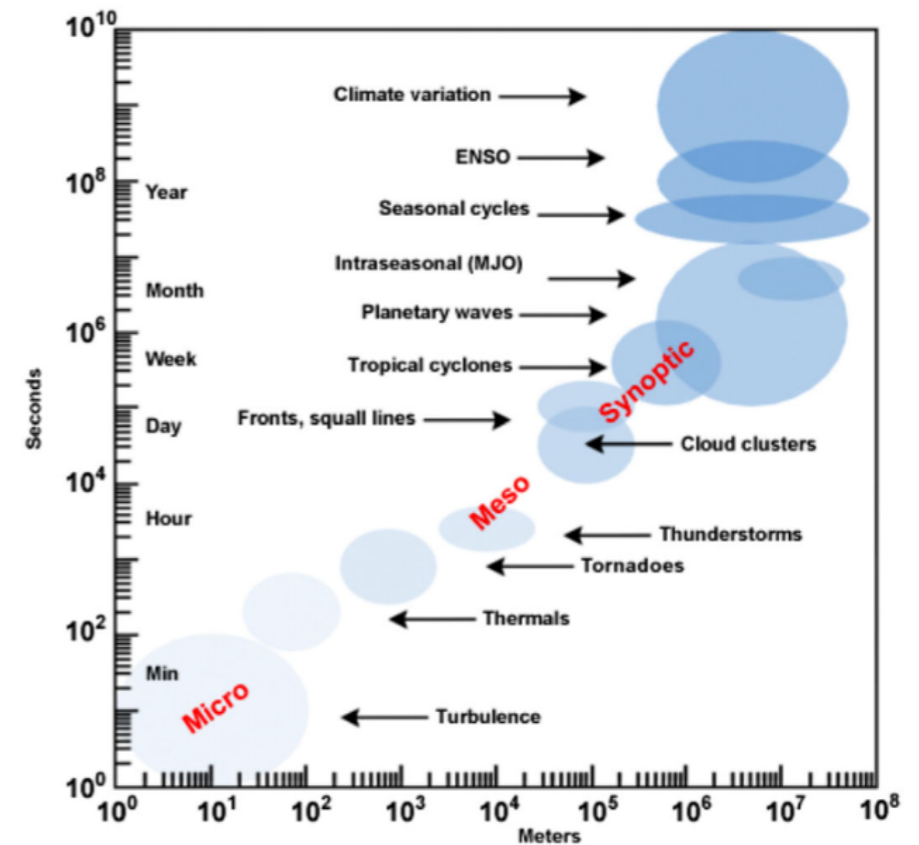
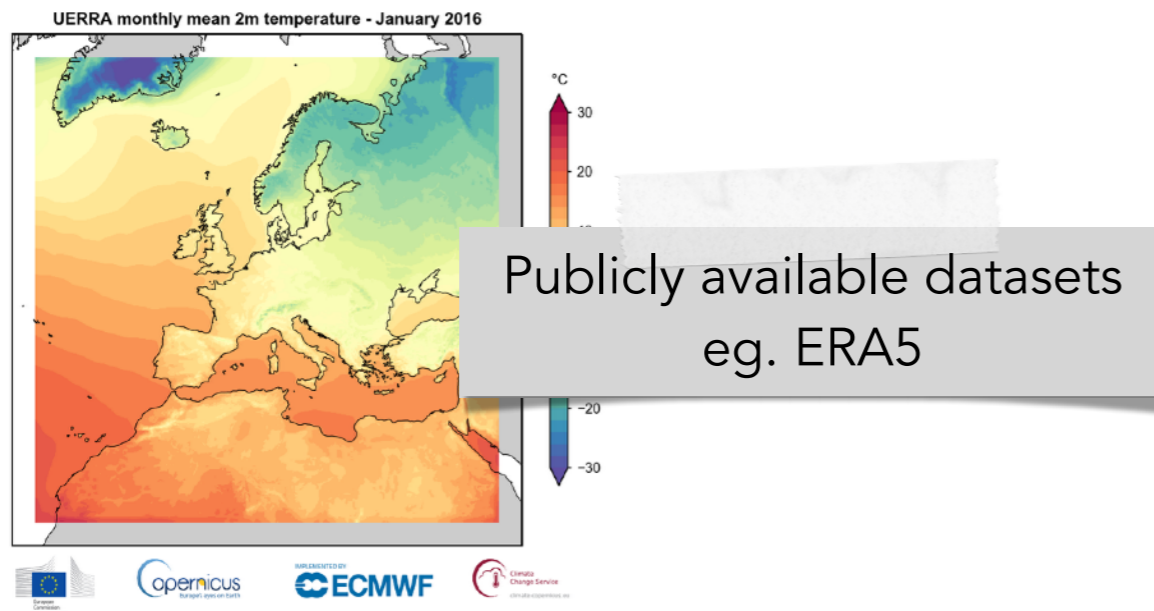
OpenLab Technical Workshop 2024 | March 2024 - CERN



Motivation and scientific challenge

Atmosphere:

- Complex phenomena involving multiple scales
- No complete classical model to simulate the dynamics
- Very large amounts of **observational** data available

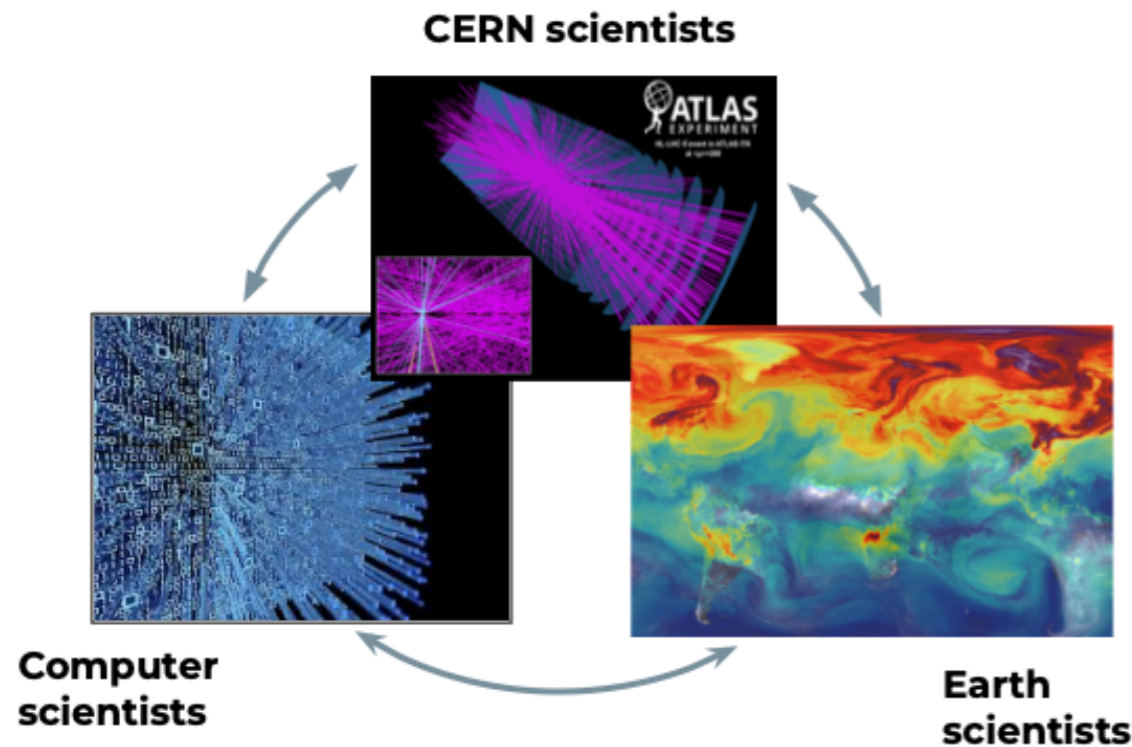


We have hundreds of TB of available atmospheric observations.

Can we use the information in these datasets for the next generation of improved weather and climate models?



AtmoRep: 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

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

Condense dataset information in a compact representation

better handle the information in downstream applications.

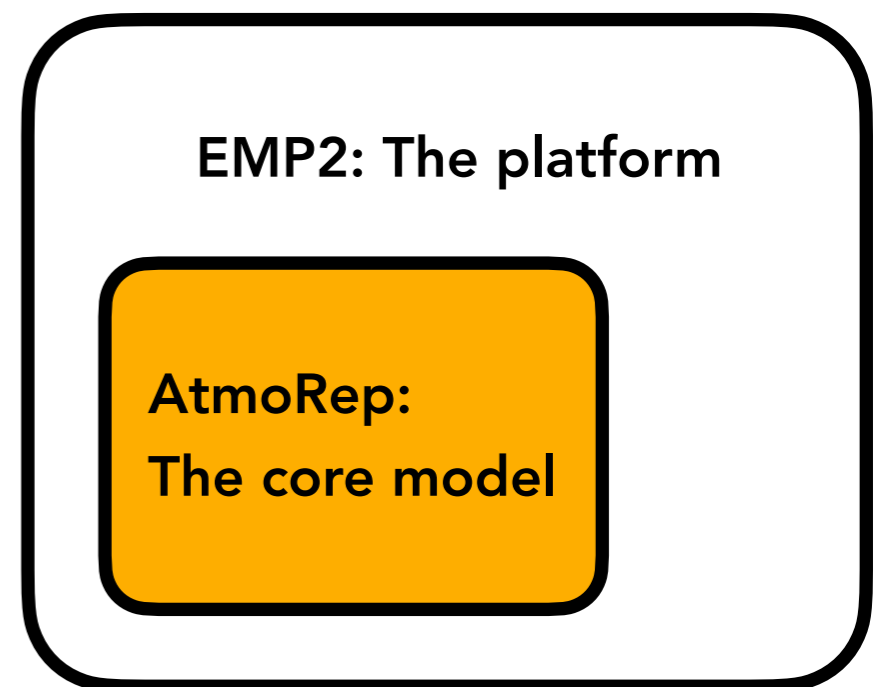
eg. condense the info in a few GB rather than TB

Common Goal:

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

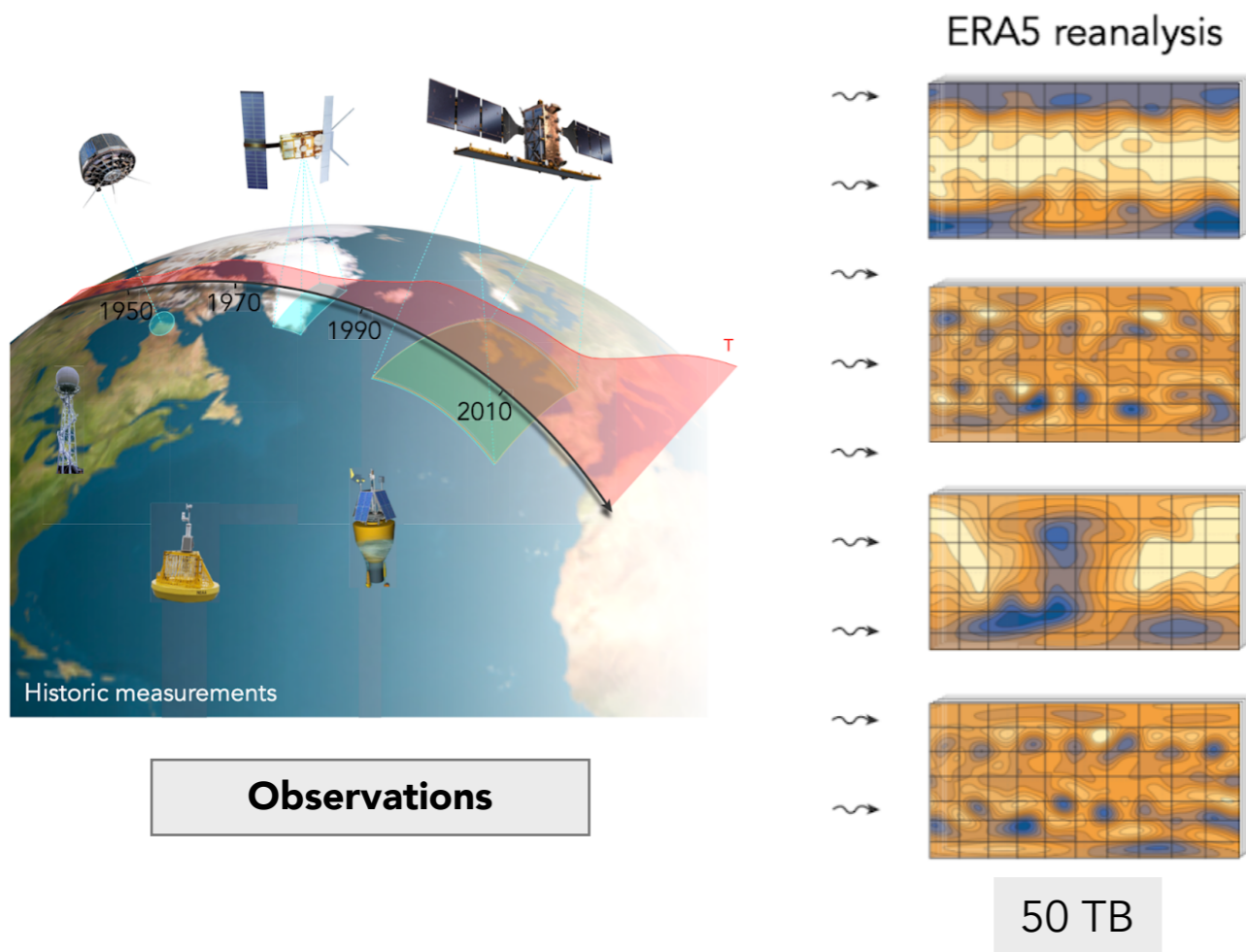
AtmoRep: A foundation model for the atmosphere

The core model



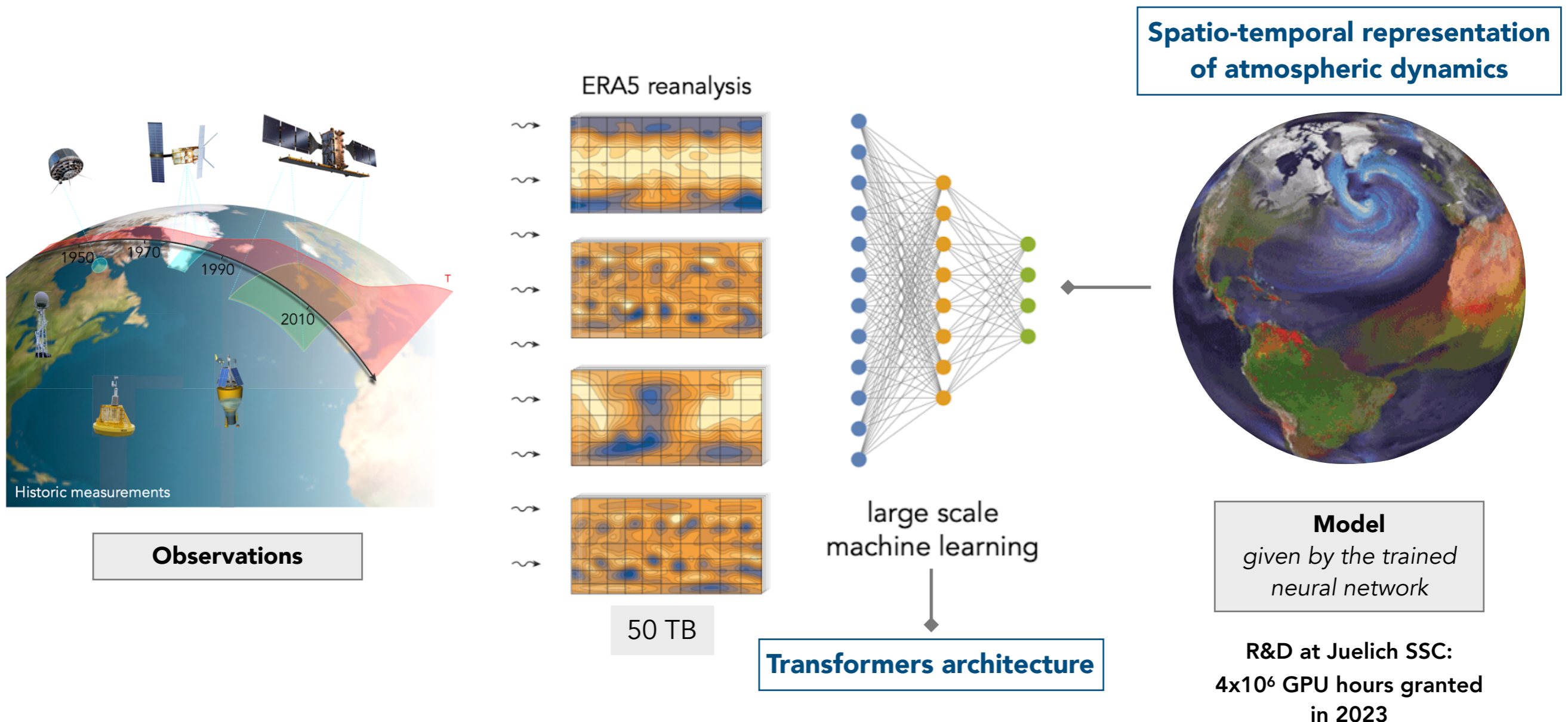
The project in a nutshell

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



The project in a nutshell

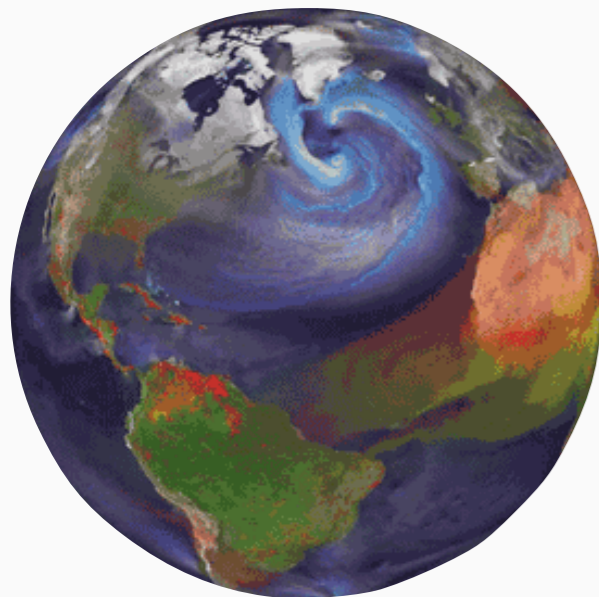
A machine-learning based global environmental model trained on terabytes of observational data



Applications: one model for multiple purposes

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

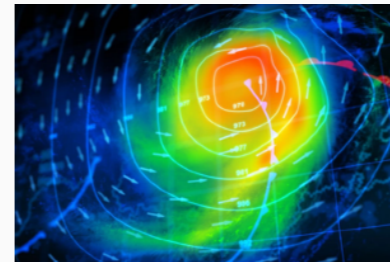
Spatio-temporal representation of atmospheric dynamics



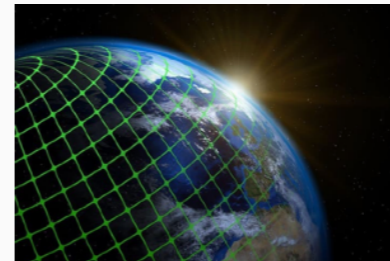
Model
given by the trained neural network

Task dependent
Adaptable smaller
networks

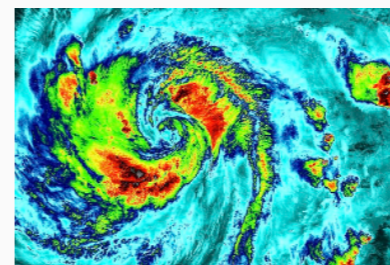
Adaptation



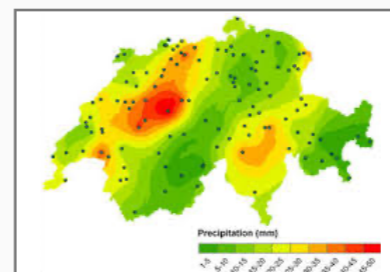
Weather predictions



Downscaling



Bias corrections

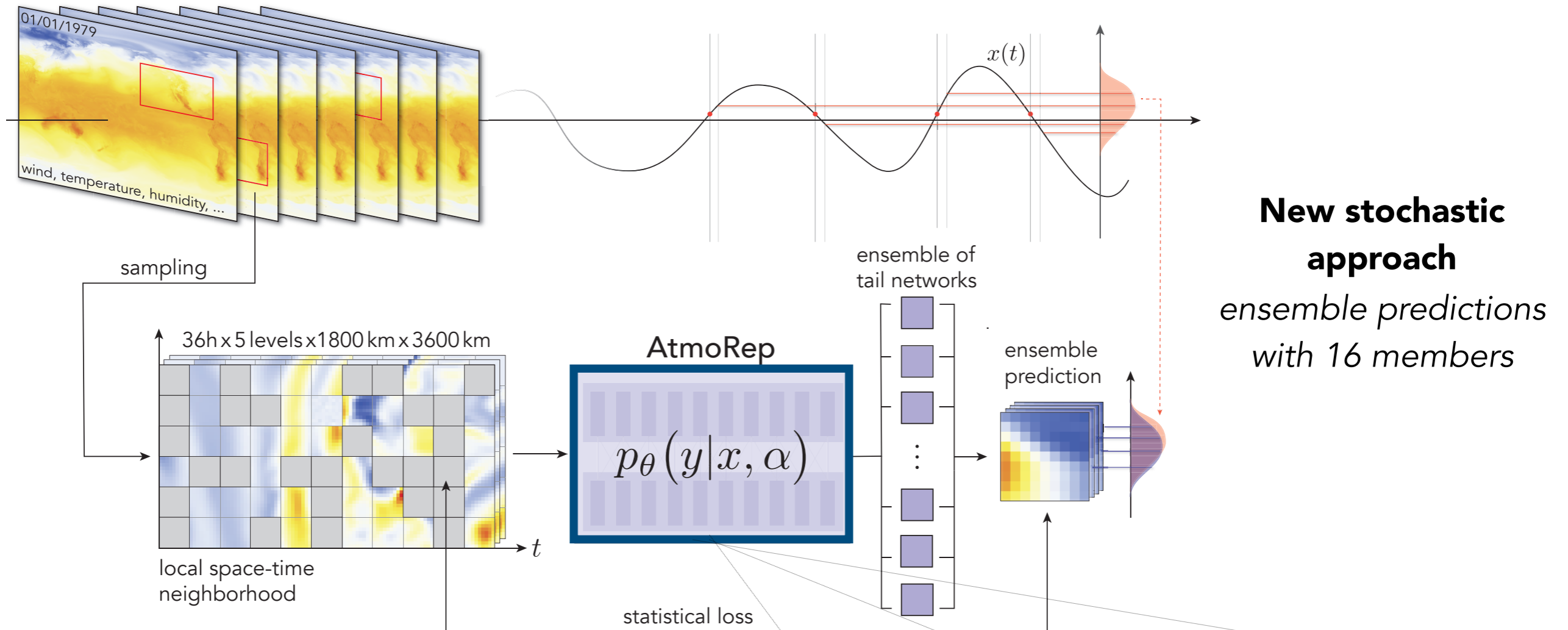


Spatio-temporal Interpolations (WIP)



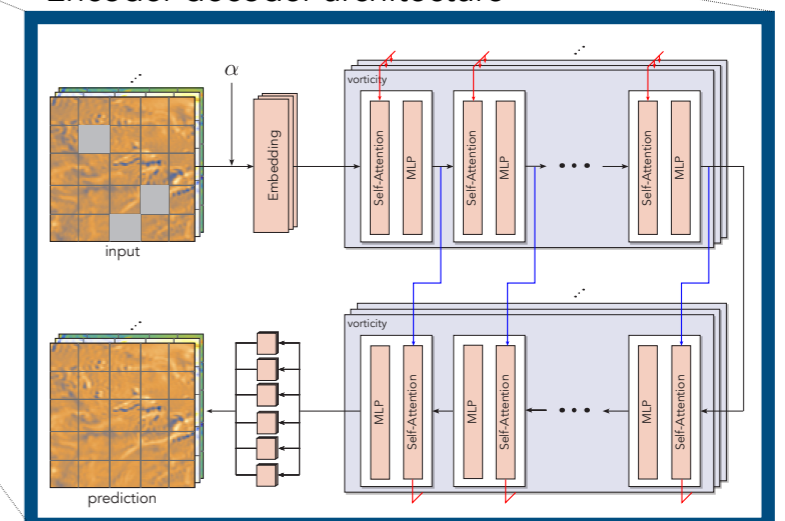
The network architecture

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



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

Encoder decoder architecture

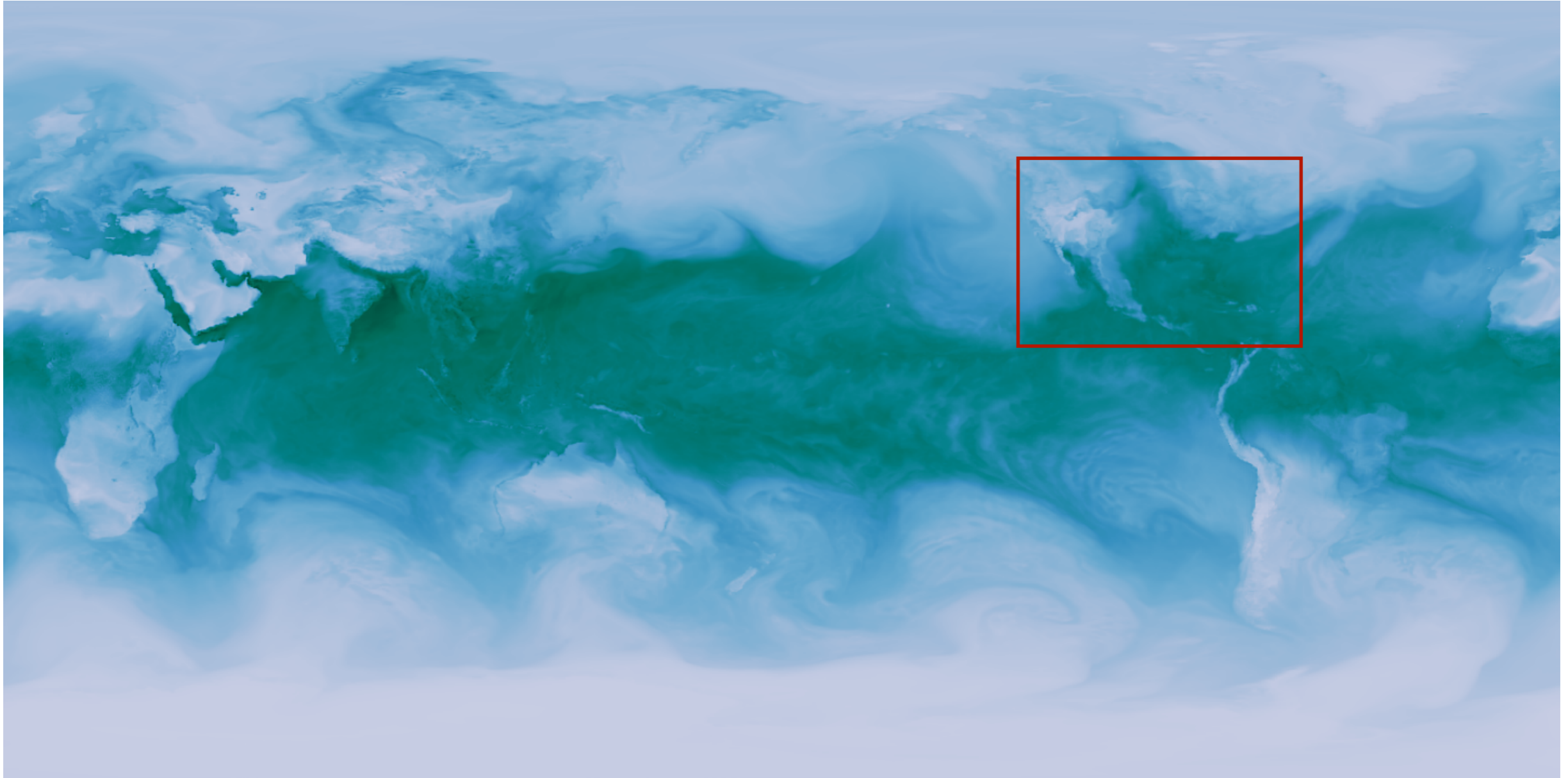


Short term weather forecasting

Zero-shot applications

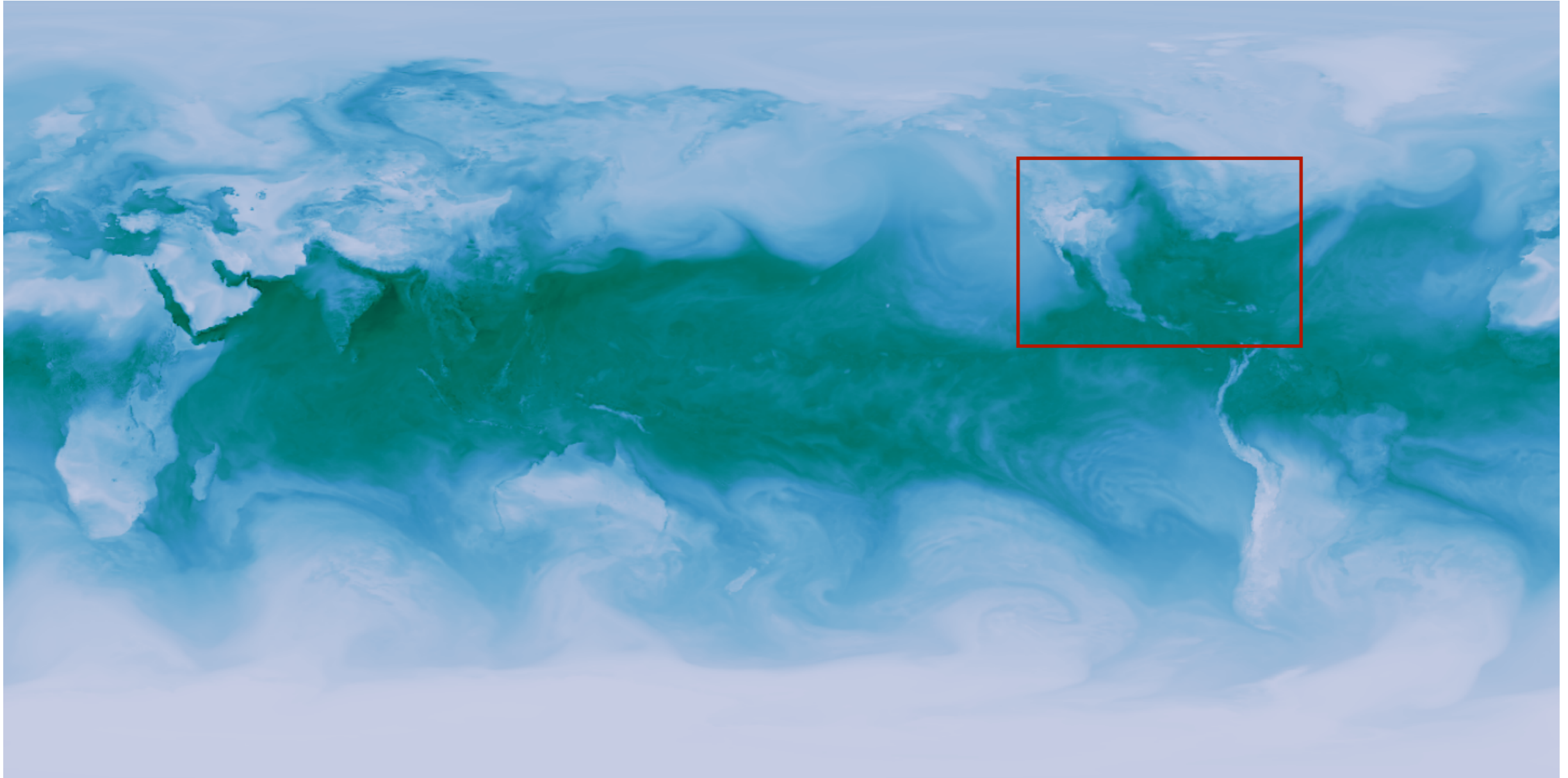
Results: Target - ERA5

specific humidity, June 15th 2018 13:00 UTC



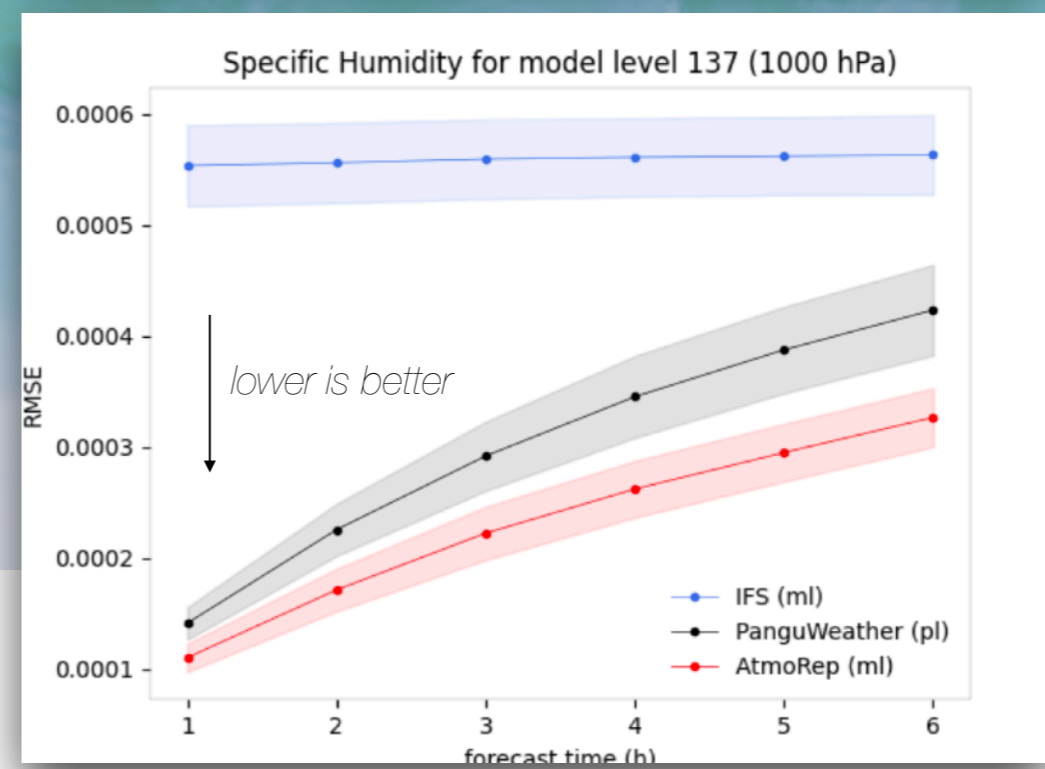
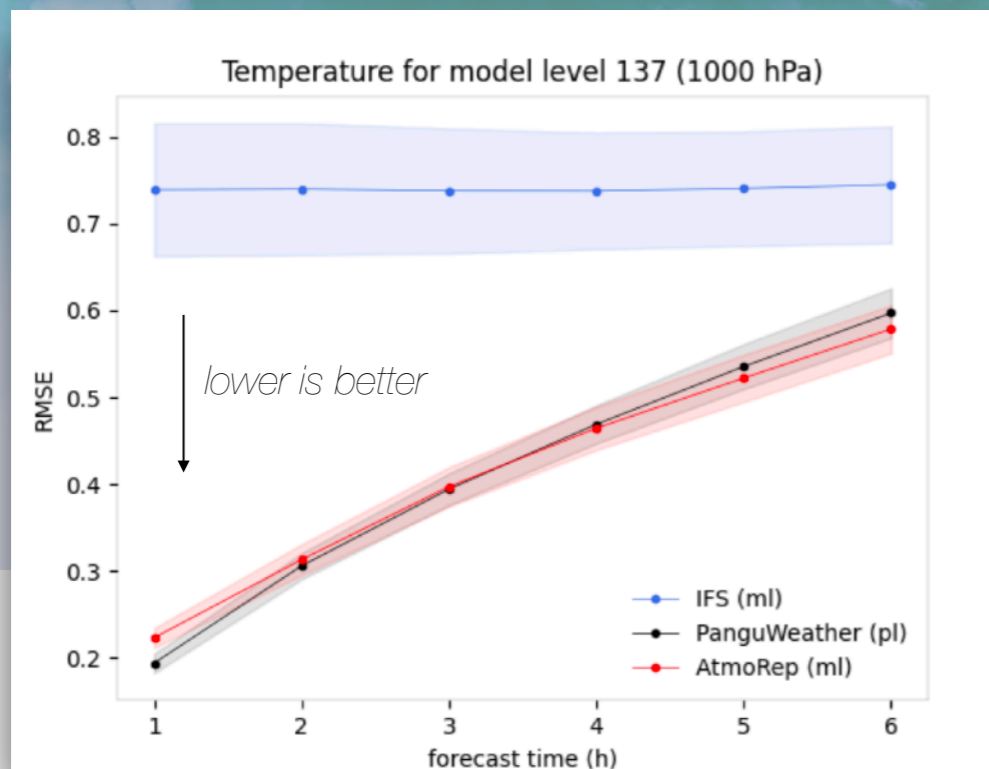
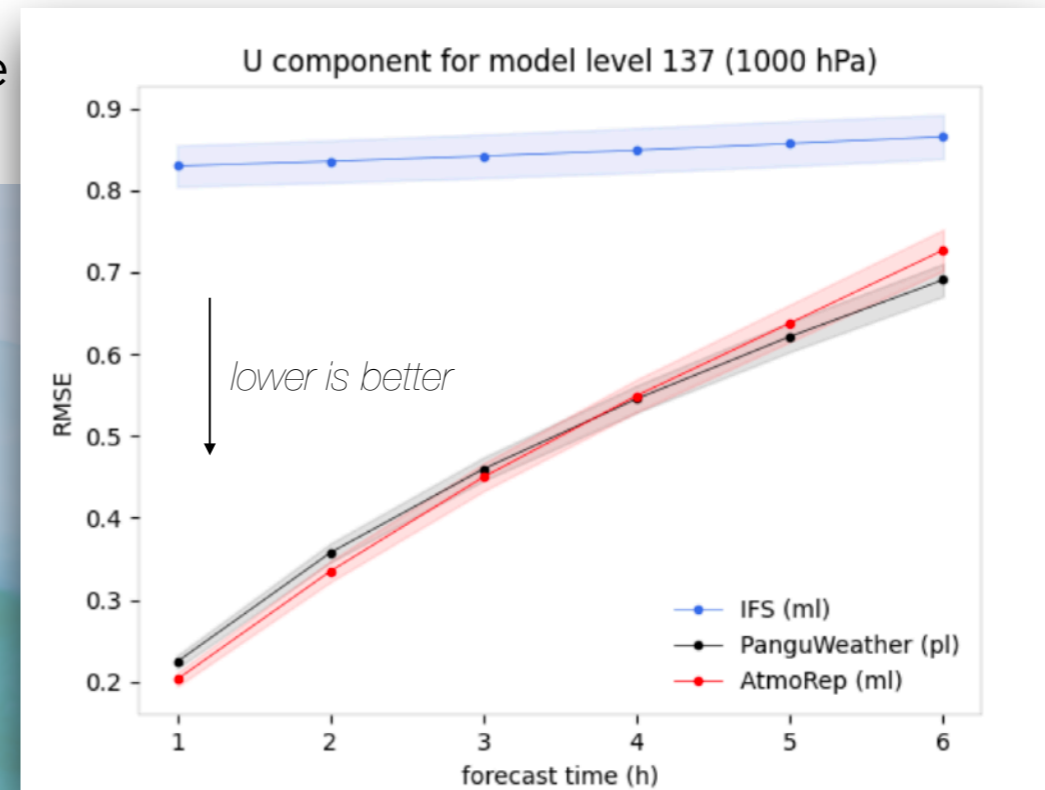
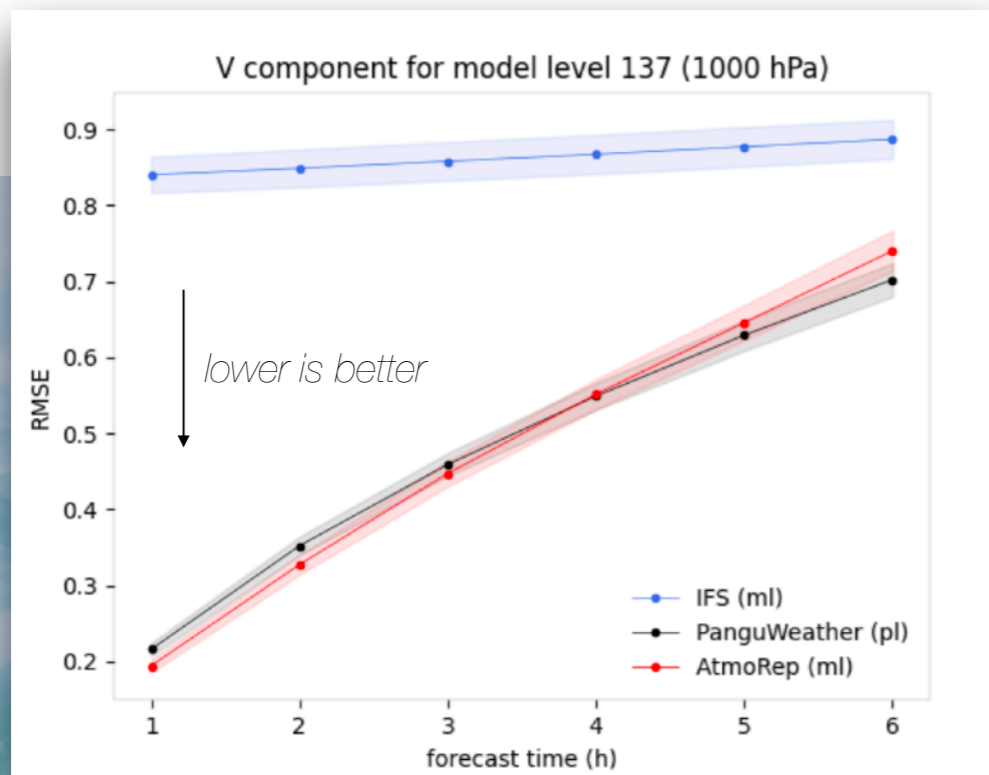
Results: Prediction - AtmoRep

specific humidity, June 15th 2018 13:00 UTC



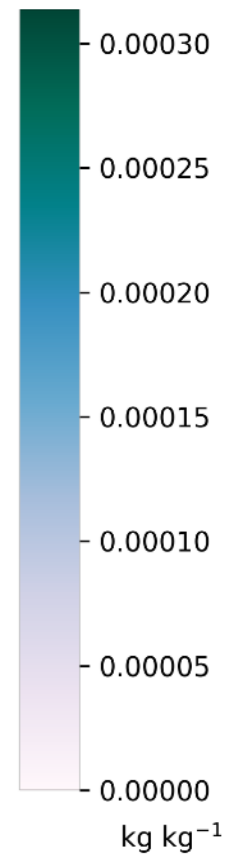
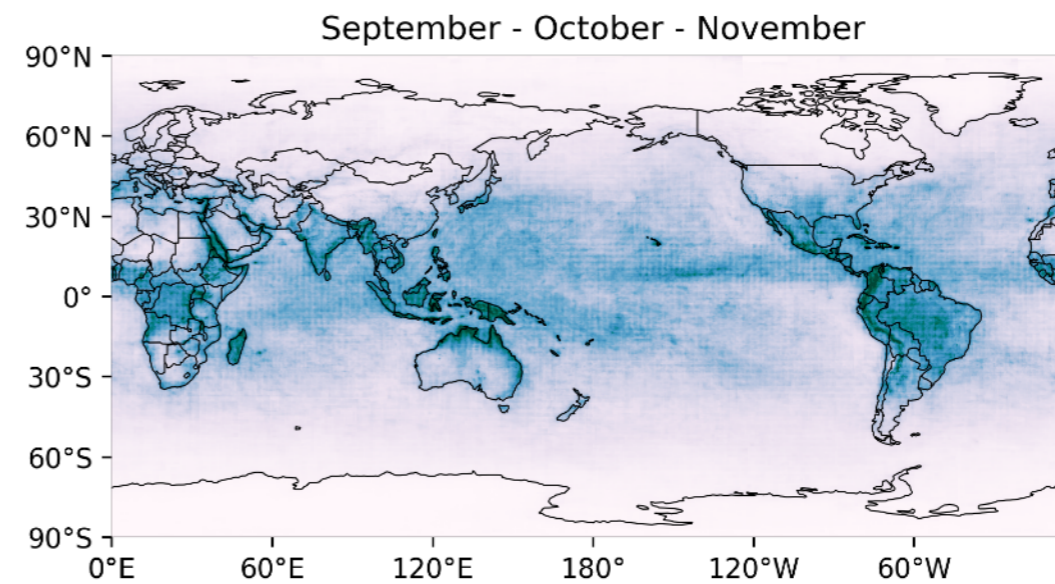
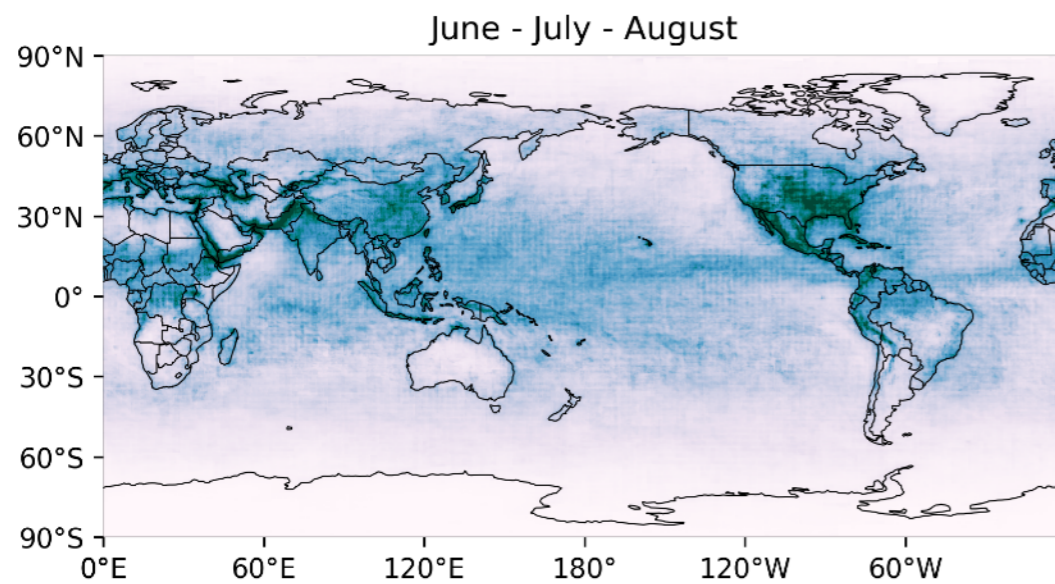
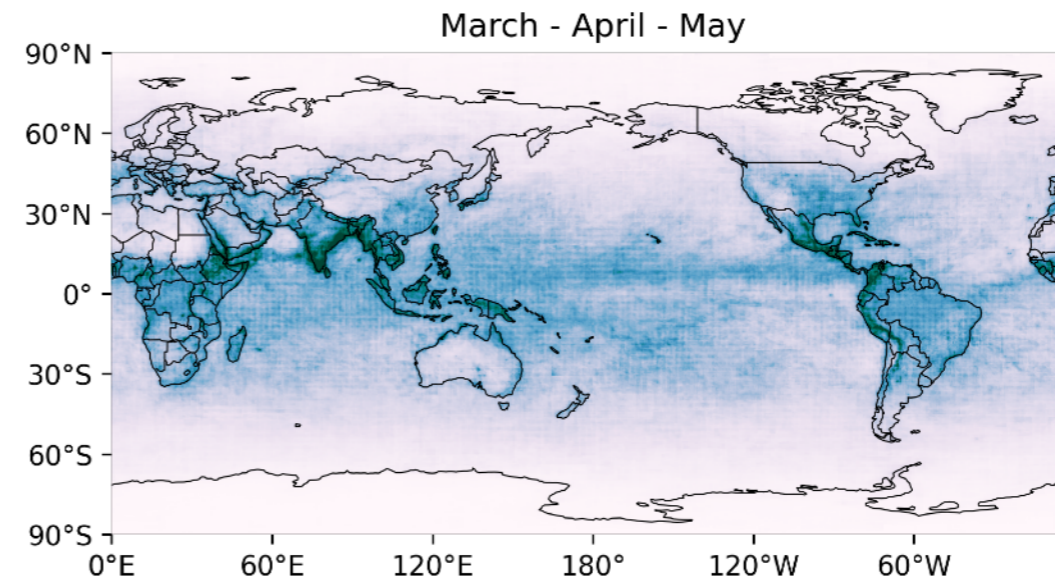
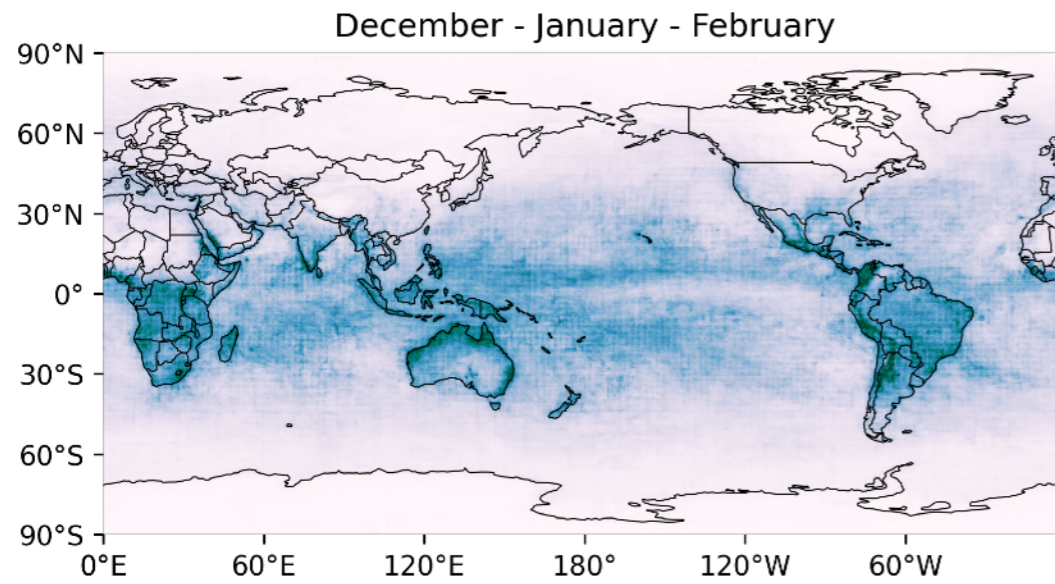
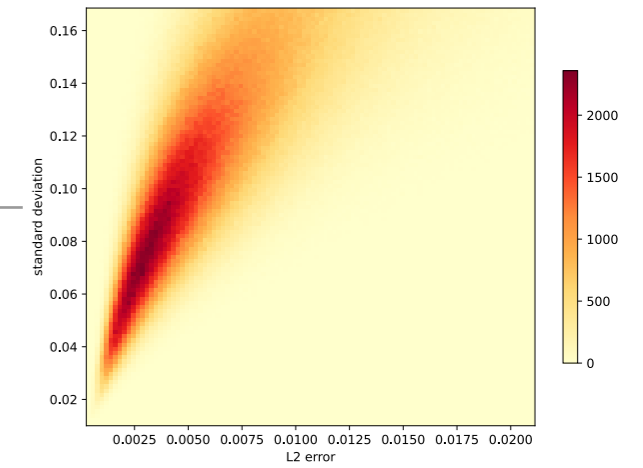
Results: Prediction - AtmoRep

RMSE = Root Mean Square Error



Results: ensemble variability

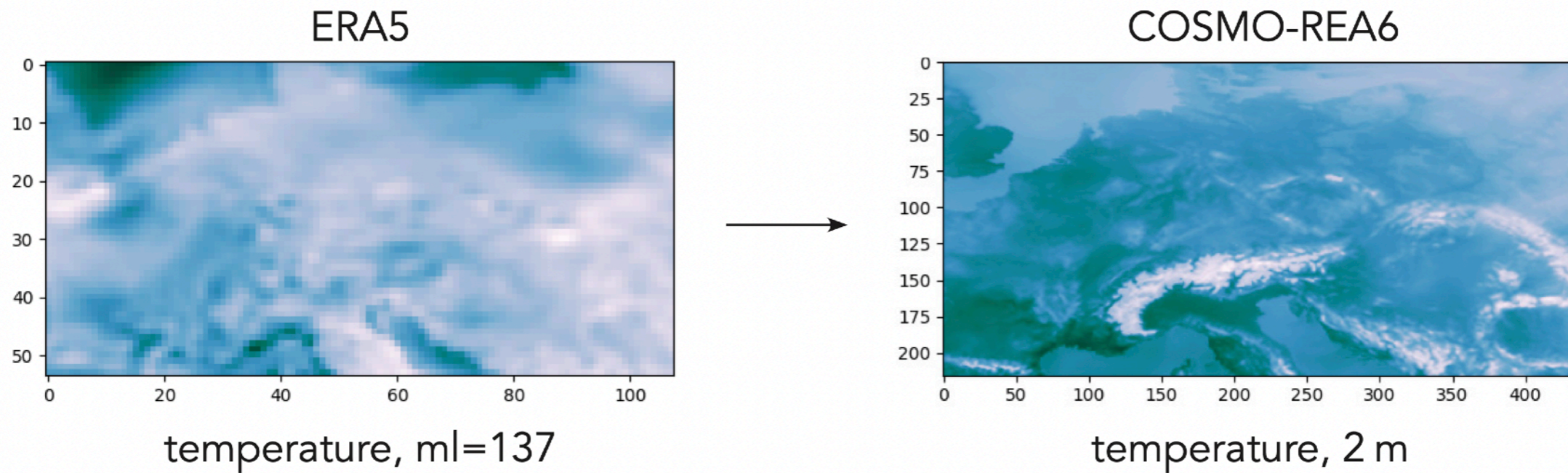
specific humidity: standard deviation of the ensemble



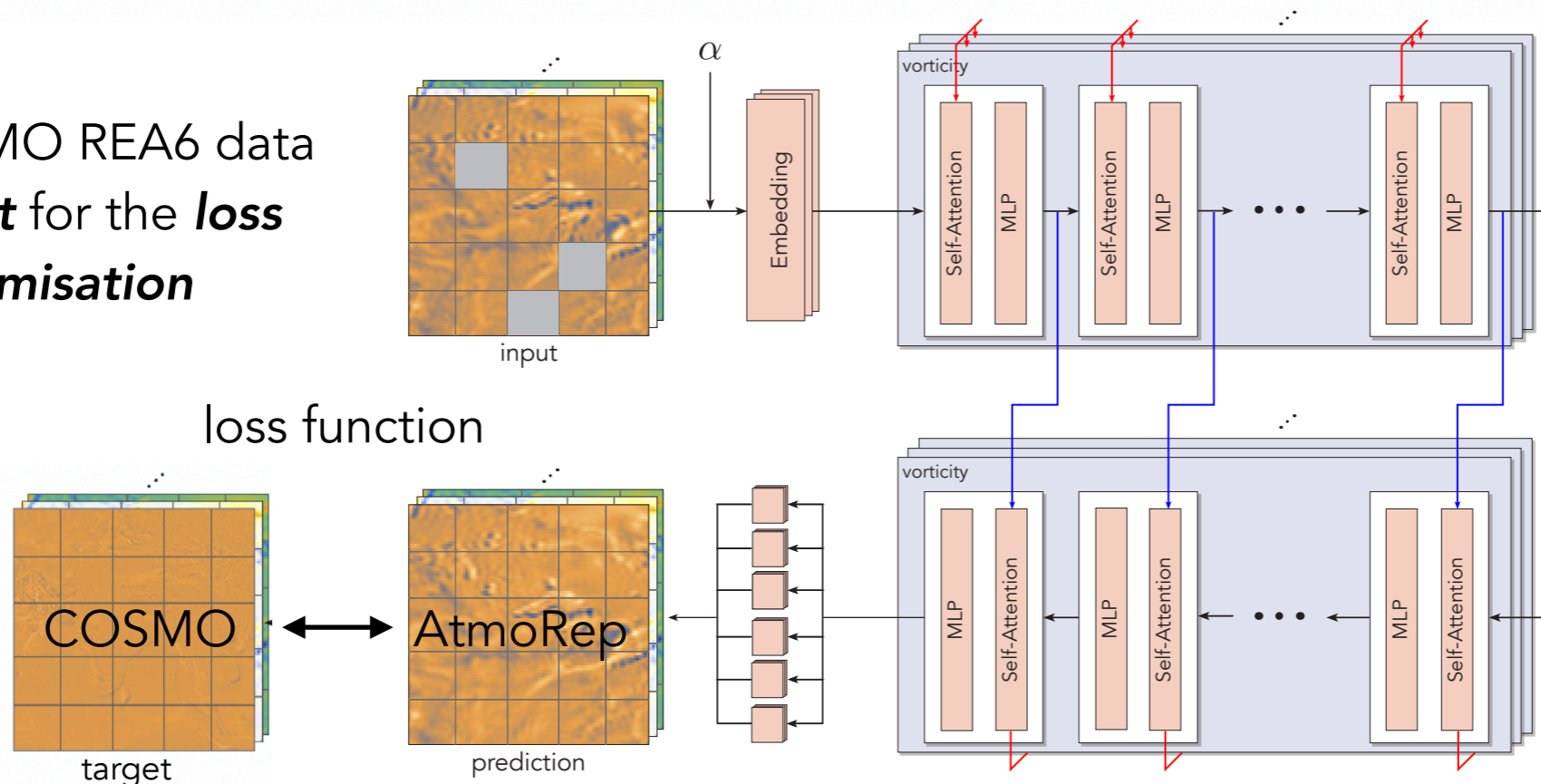
Fine tuning on real data

Data driven precipitation corrections & downscaling

Downscaling

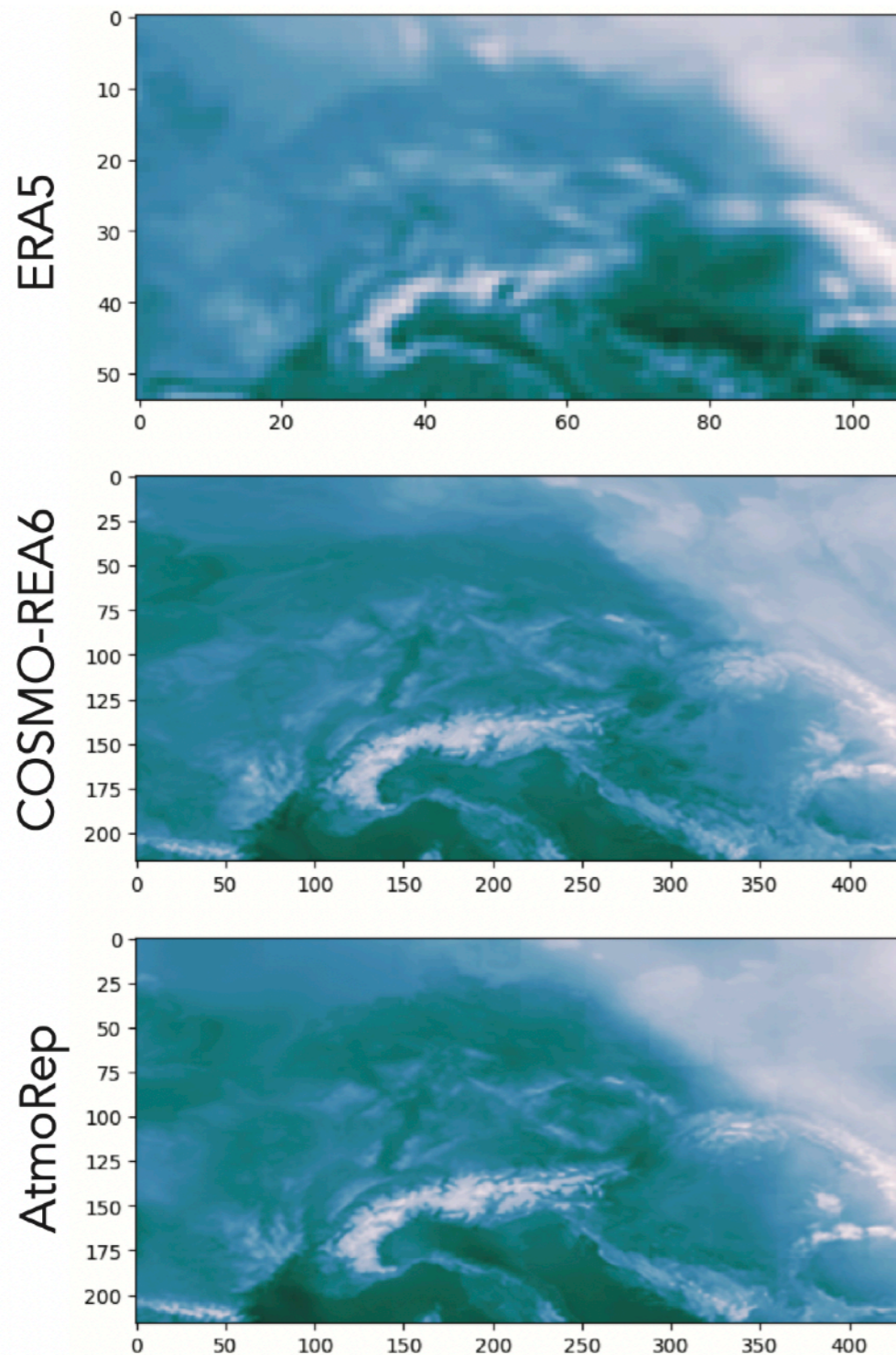


Use COSMO REA6 data as **target** for the **loss minimisation**

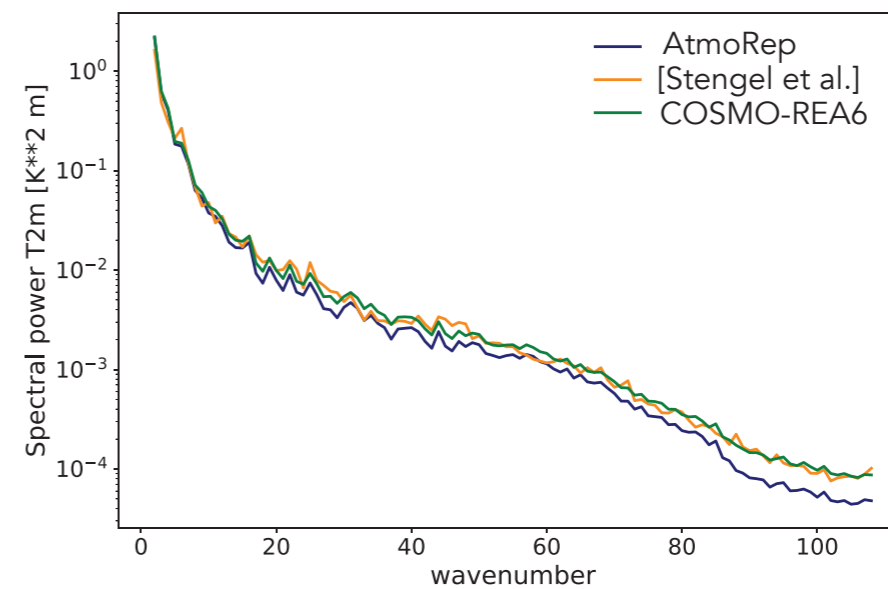
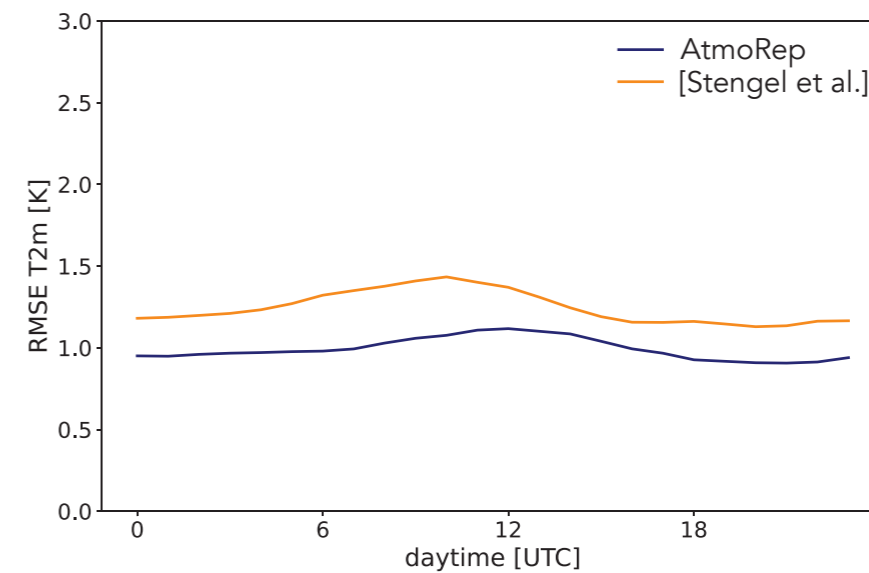


Downscaling

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:

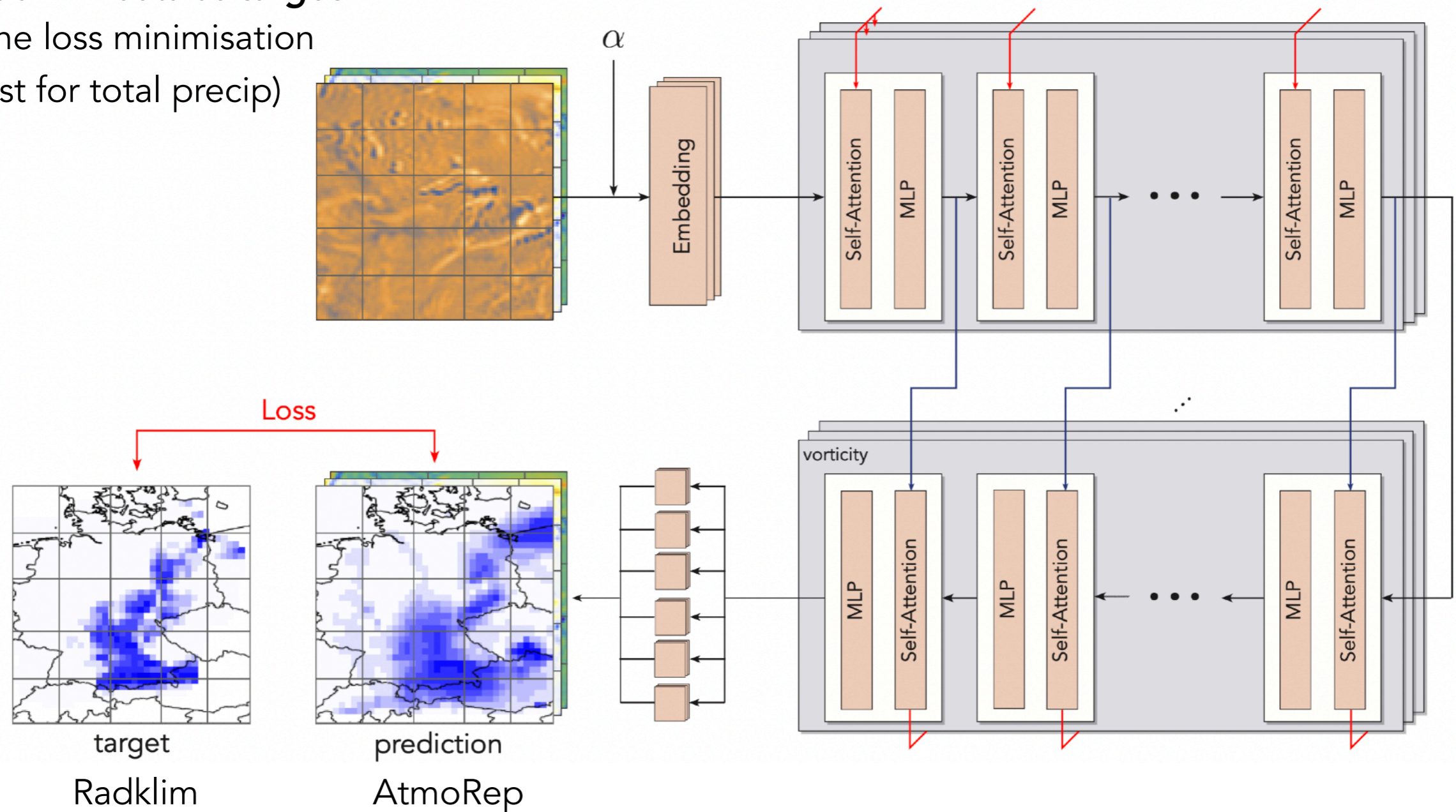


Bias corrections

Precipitation rates are known to be suboptimal in ERA5

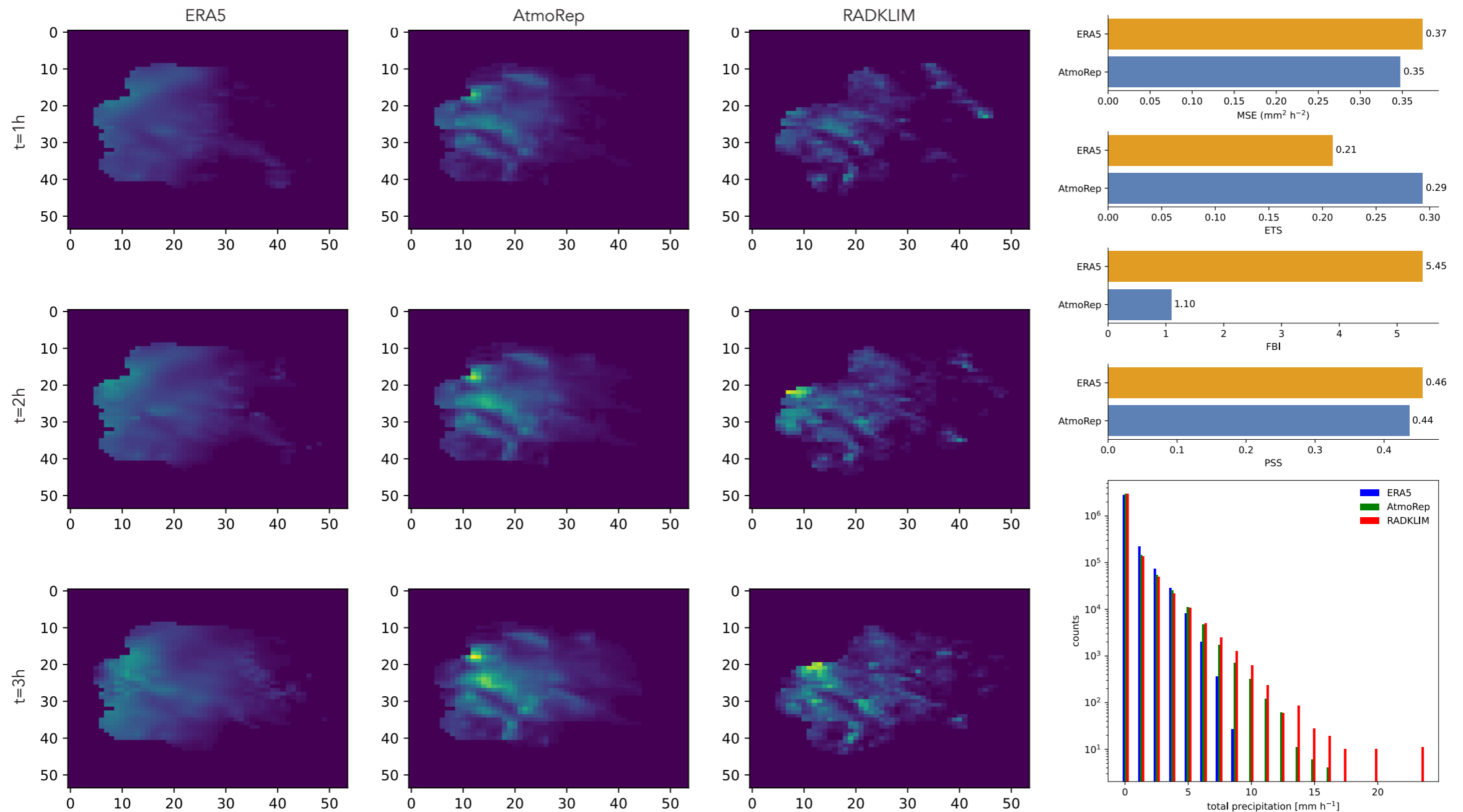
Use RADKLIM radar data to fine-tune the precipitation rates in AtmoRep

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



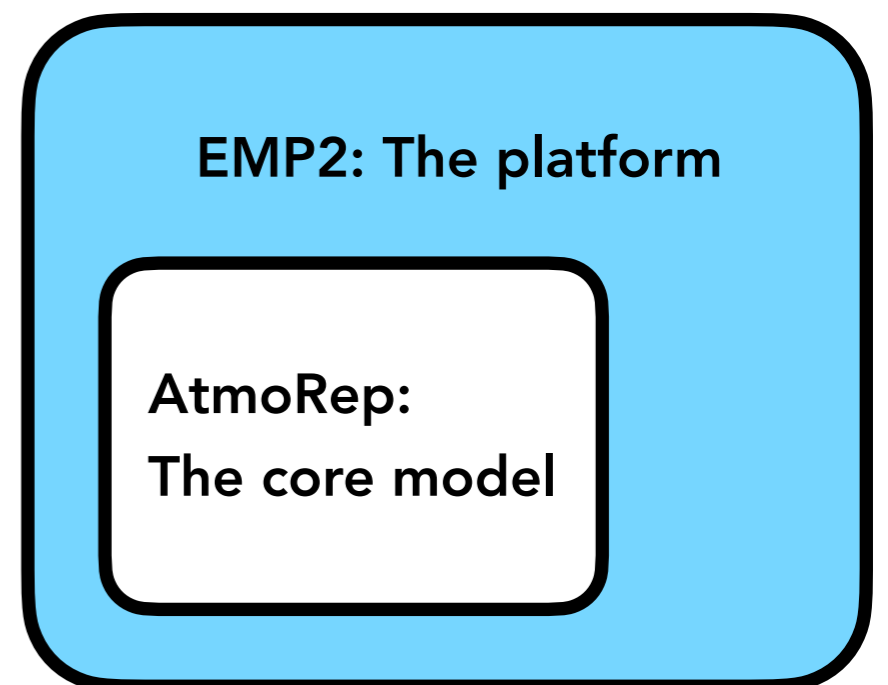
Bias corrections: Results

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



Environmental modelling and prediction platform

The user-friendly platform



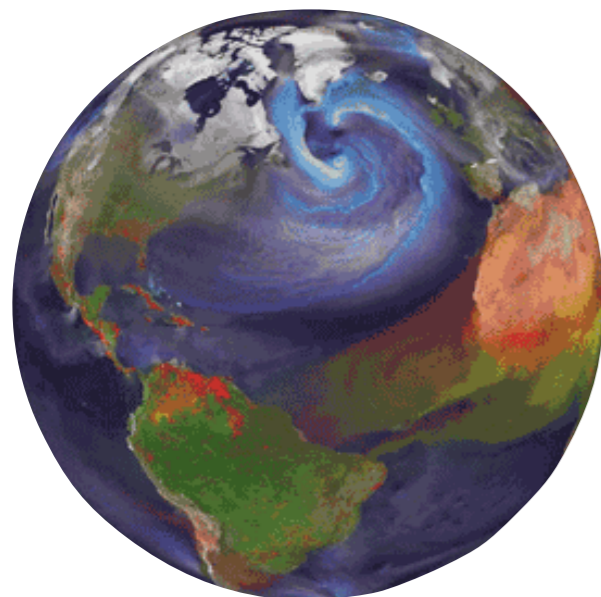
Generalisation from HPC centres to clouds

Future: develop the API & the user interface

Challenges: How foundation models interface with Digital Twin architectures?

Can we use foundation models as backbone for several digital twins
(e.g. core model = common atmospheric representation)?

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

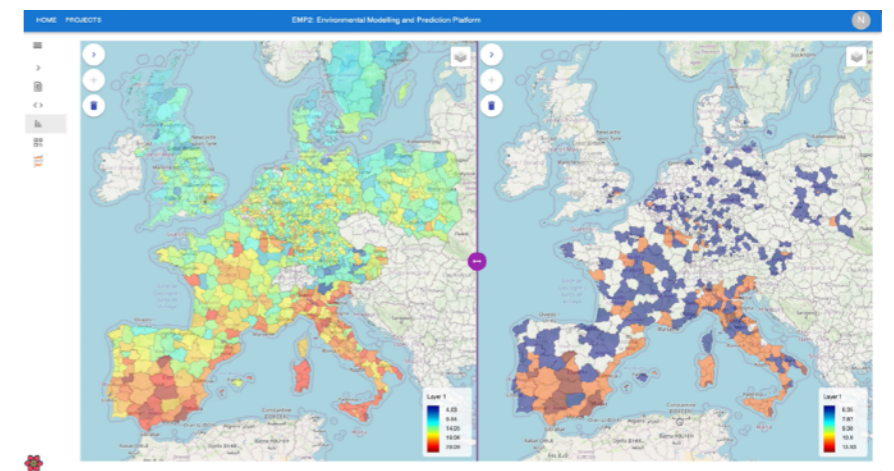


(New CS4OD)



interTwin

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



Goal: test EMP² within a digital twin existing architecture.

EMP² will be implemented as one of the use cases to test the Digital Twin architecture developed through the InterTwin Project

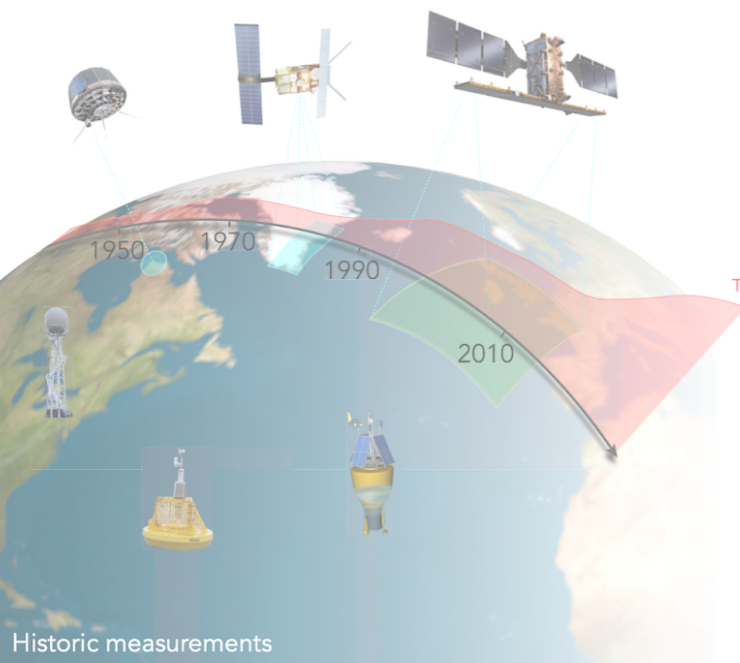
Conclusions

AtmoRep: First prototype of a multi-purpose model for Earth system applications

The **model is available and testable** on the current applications: nowcasting, downscaling, temporal interpolation and precipitation corrections.

More infos:

- Code is available on GitHub: [link](#)
- More infos on the website: www.atmorep.org
- **Pre-print on ArXiv: [link](#)**

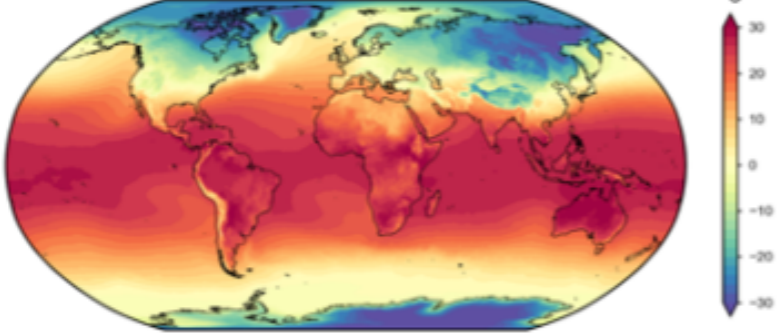


.. and some long term plans:

- How to integrate "raw" observations?
- Coupled atmosphere+ocean system?

Backup

ERA5 monthly mean 2m temperature - January 2016

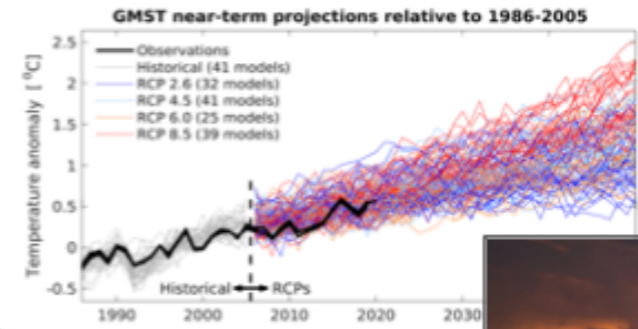


Operations

Technology

Collect

Aware



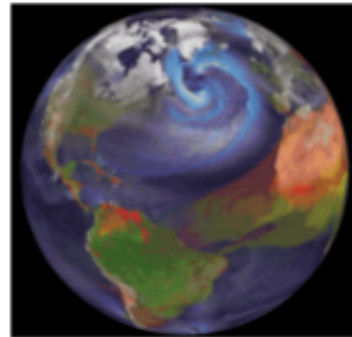
Artificial Intelligence



Big Data Analytics



Compute



Respond



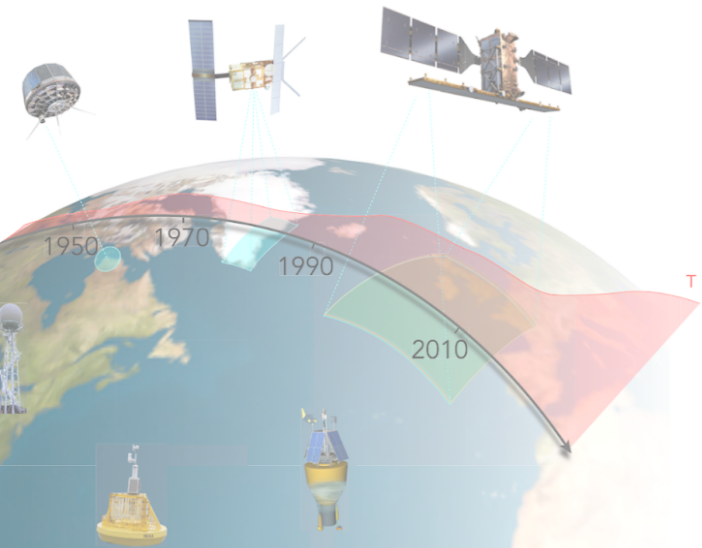
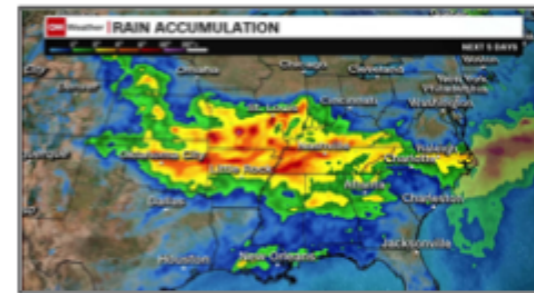
Visualization Tools



Visualize



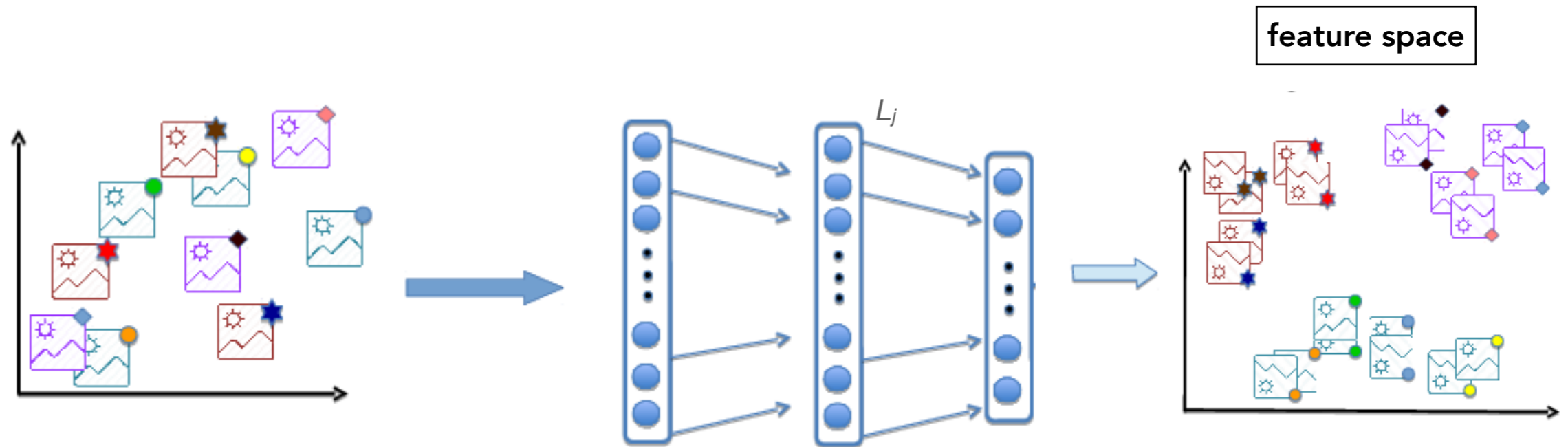
Predict



Key ingredient: Representation learning

Representation learning:

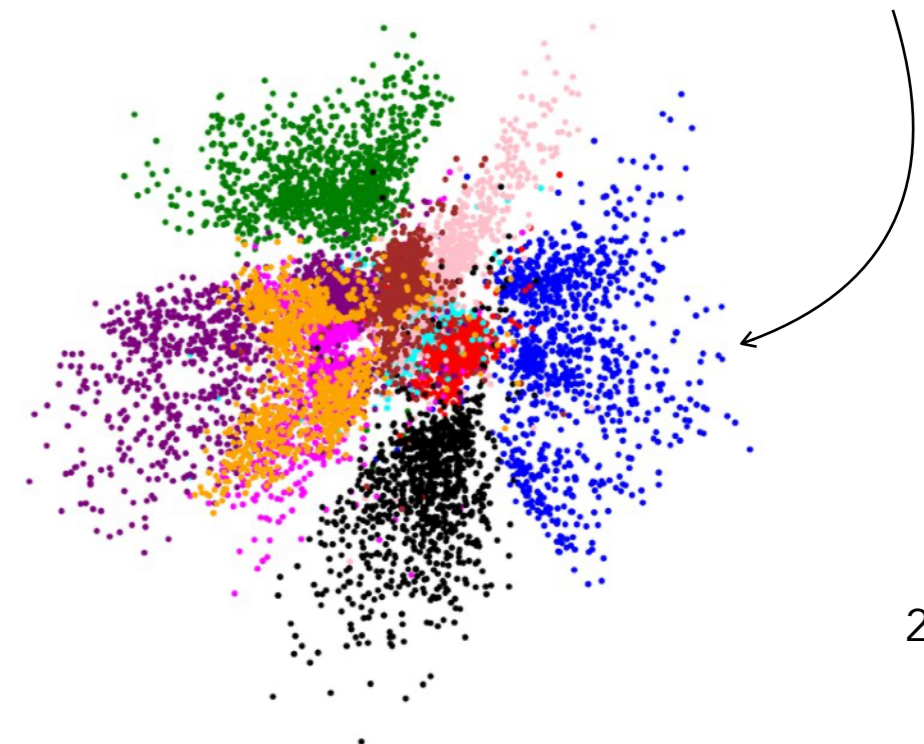
- Learn a **task-independent representation** of the data in the **feature space** of the neural network



Effective "encoding" of the data useful for many applications

$$L_j : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

Data are clustered according to common properties in the "feature space"



Publicly available pre-processed dataset of hourly spaced interpolated Earth observations: The ERA5 reanalysis from ECMWF

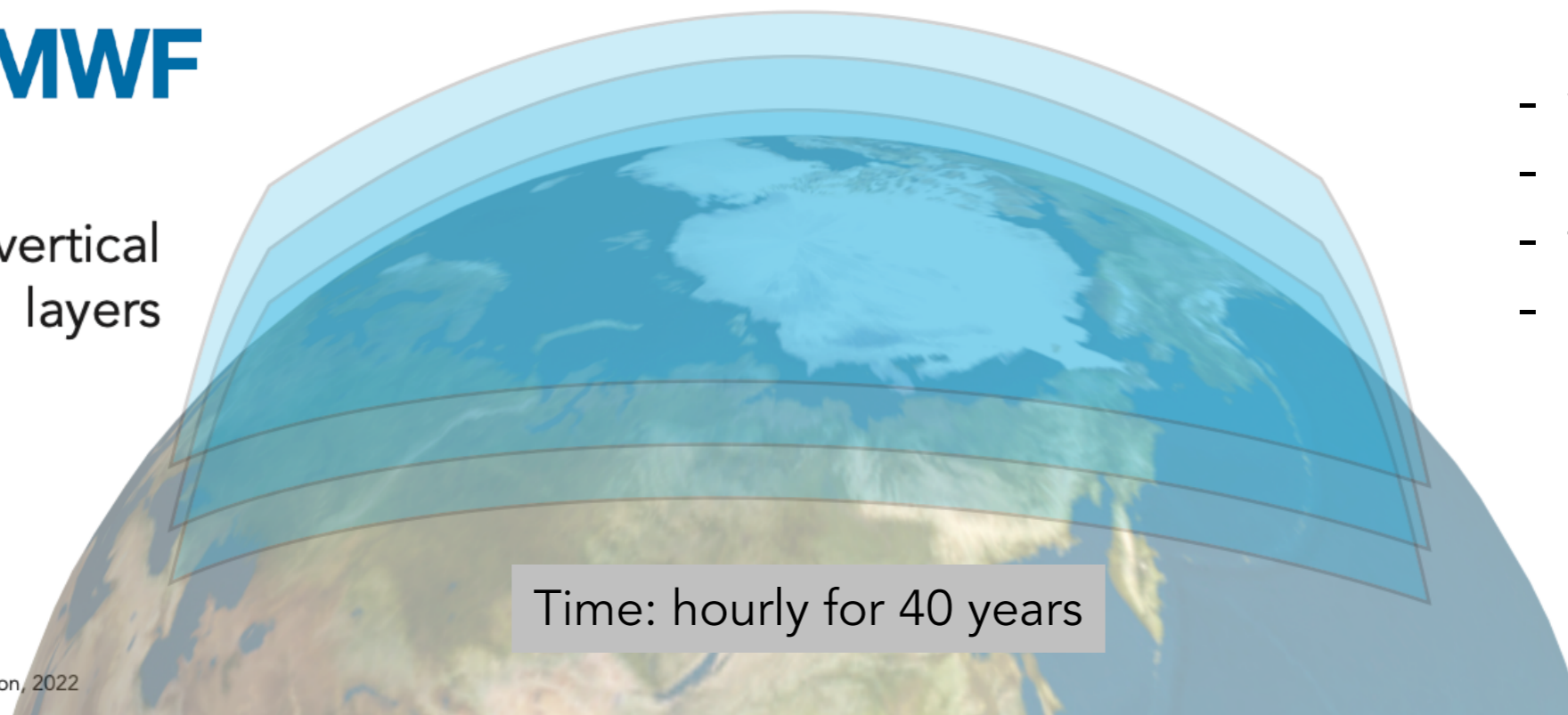
Subset of ERA5 reanalysis used at the moment for training:

- Physical fields: vorticity, divergence (or wind velocity), vertical velocity, temperature, specific humidity, total precipitation
- Space: 721 x 1440 x 5 vertical layers
- Time: **randomly sample** over 24 time steps per day for 365 days for 40 years

721x1440 horizontal grid (0.25 degree)



137 vertical layers

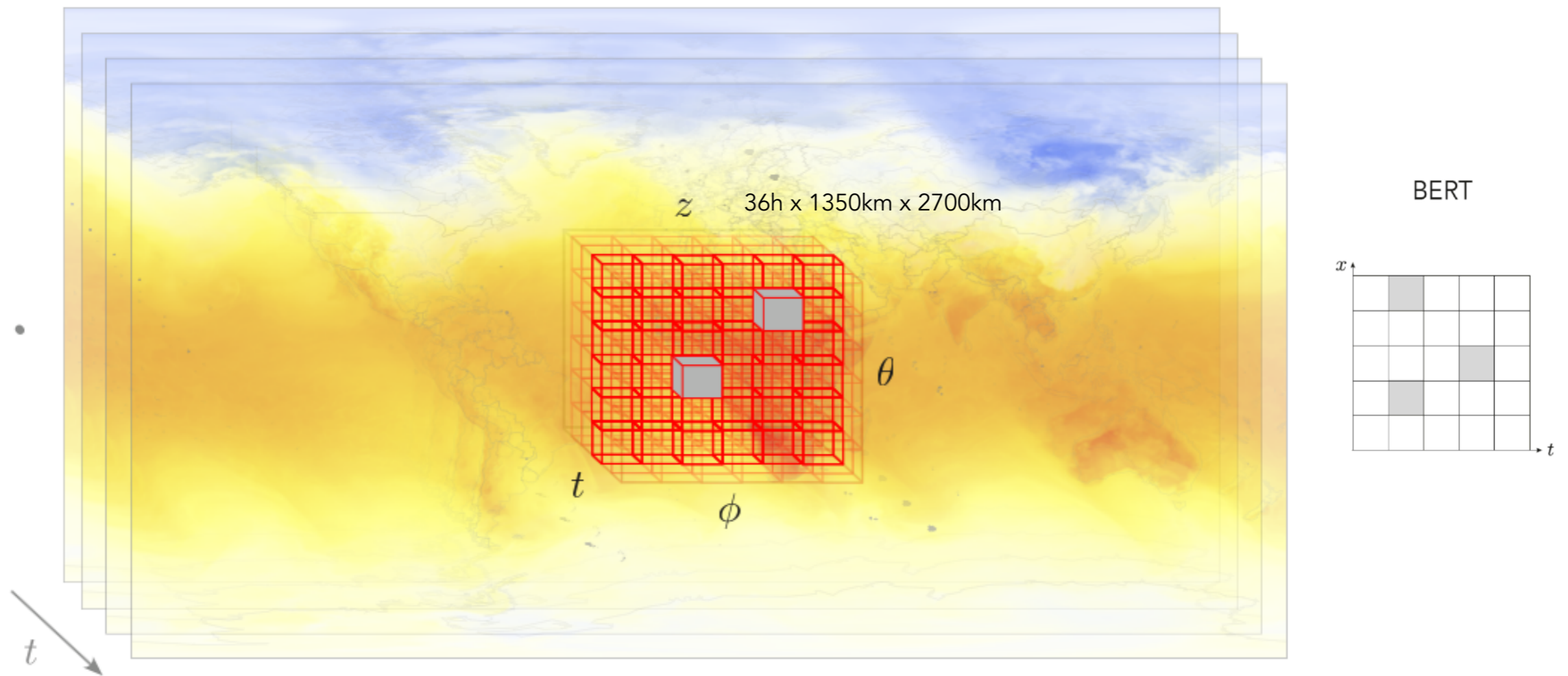


- vorticity
- divergence
- temperature
- ...

Intermezzo: The training protocol

Use a variation of BERT masked language model from self supervised trainings in NLP

Random sampling of neighbourhoods for training → stochastic gradient descent



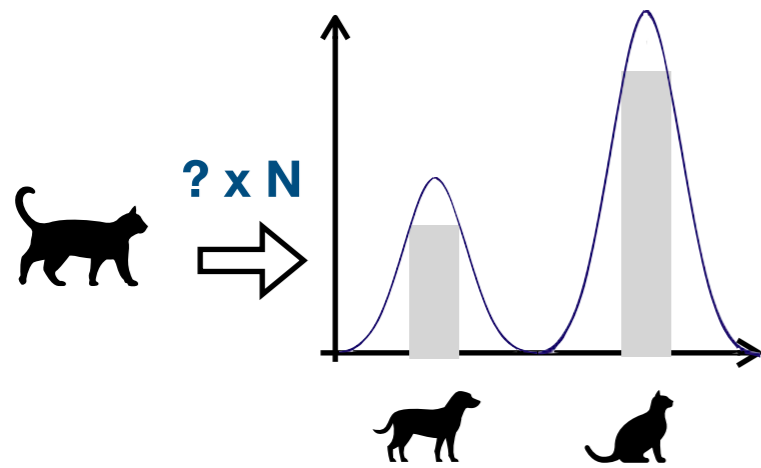
Split cube in small space-time regions (3D cubes) → tokens

Mask random tokens within the hyper-cube and try to predict them back

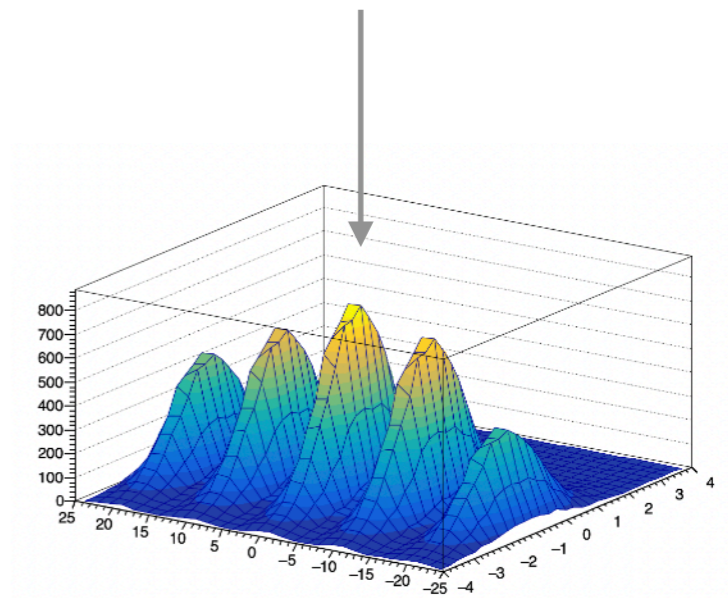
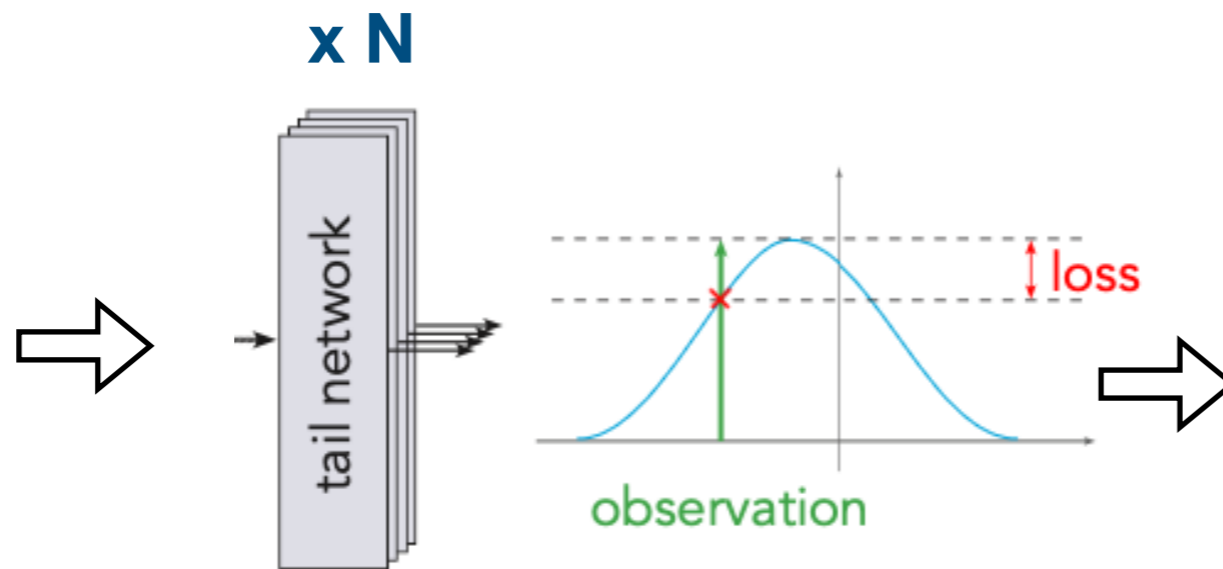
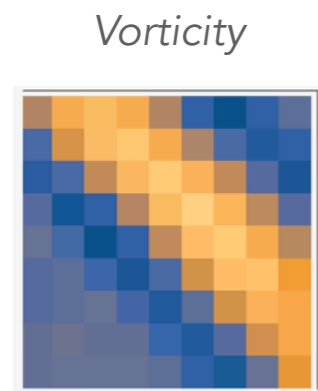
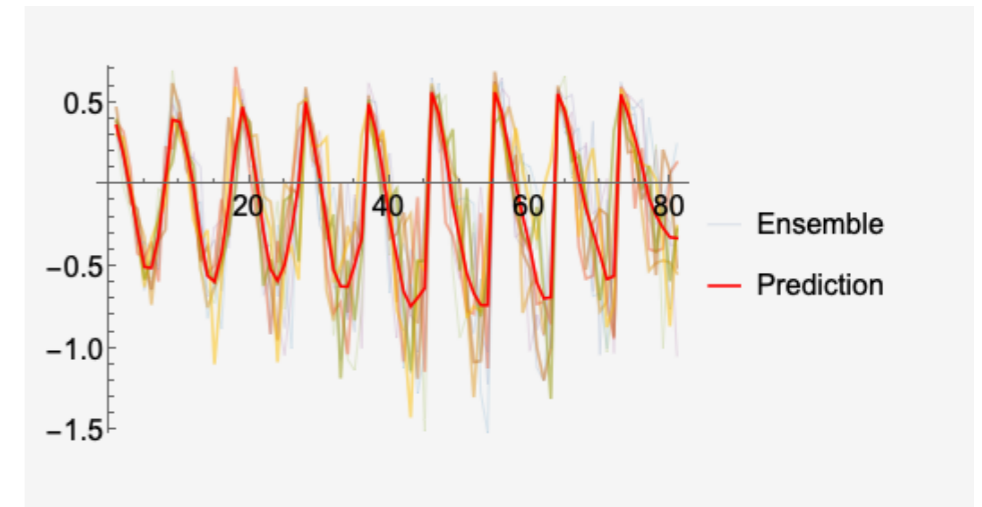
visually: learn representation dynamics through interpolation

Default: 12 x 6 x 12 tokens with 3 x 9 x 9 grid points

Key ingredients: statistical loss



Inspired by
**cross entropy loss in
classification problems**



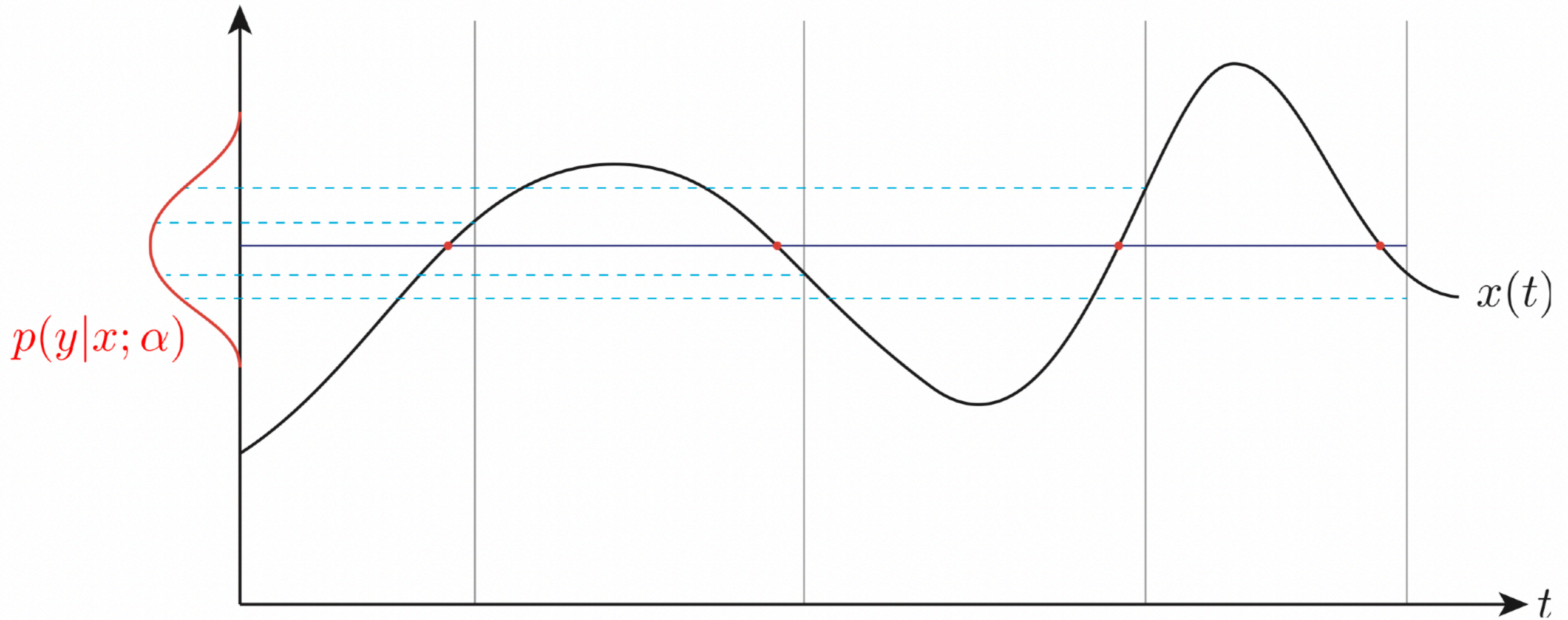
**Interpret as probability
distribution for each pixel
(assumed Gaussian)**

*statistical interpretation:
measure the difference between the
pdf of the ML classification model
and the predicted distribution*

**Ensemble of tail networks
generates N predictions for
each pixel**

Loss: Minimize the difference between the mean of the distribution and the true value

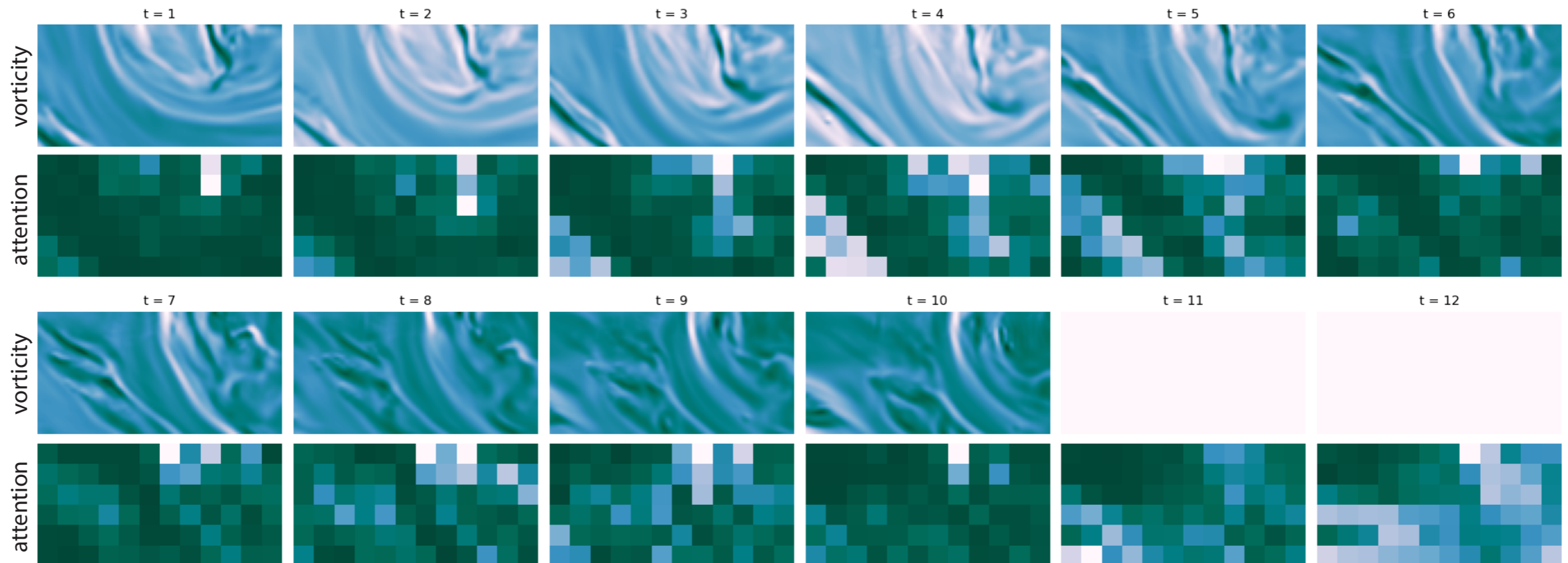
Ensemble variability



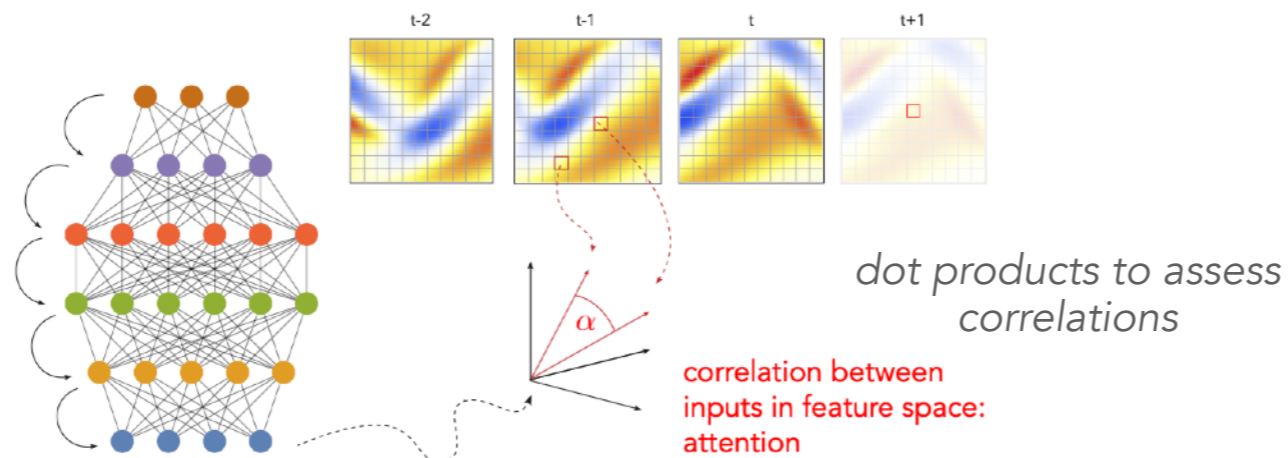
Attention maps and interpretability

Inspect the self-attention mechanism:

can we identify physics phenomena (e.g. hurricane formation) before they are even created?



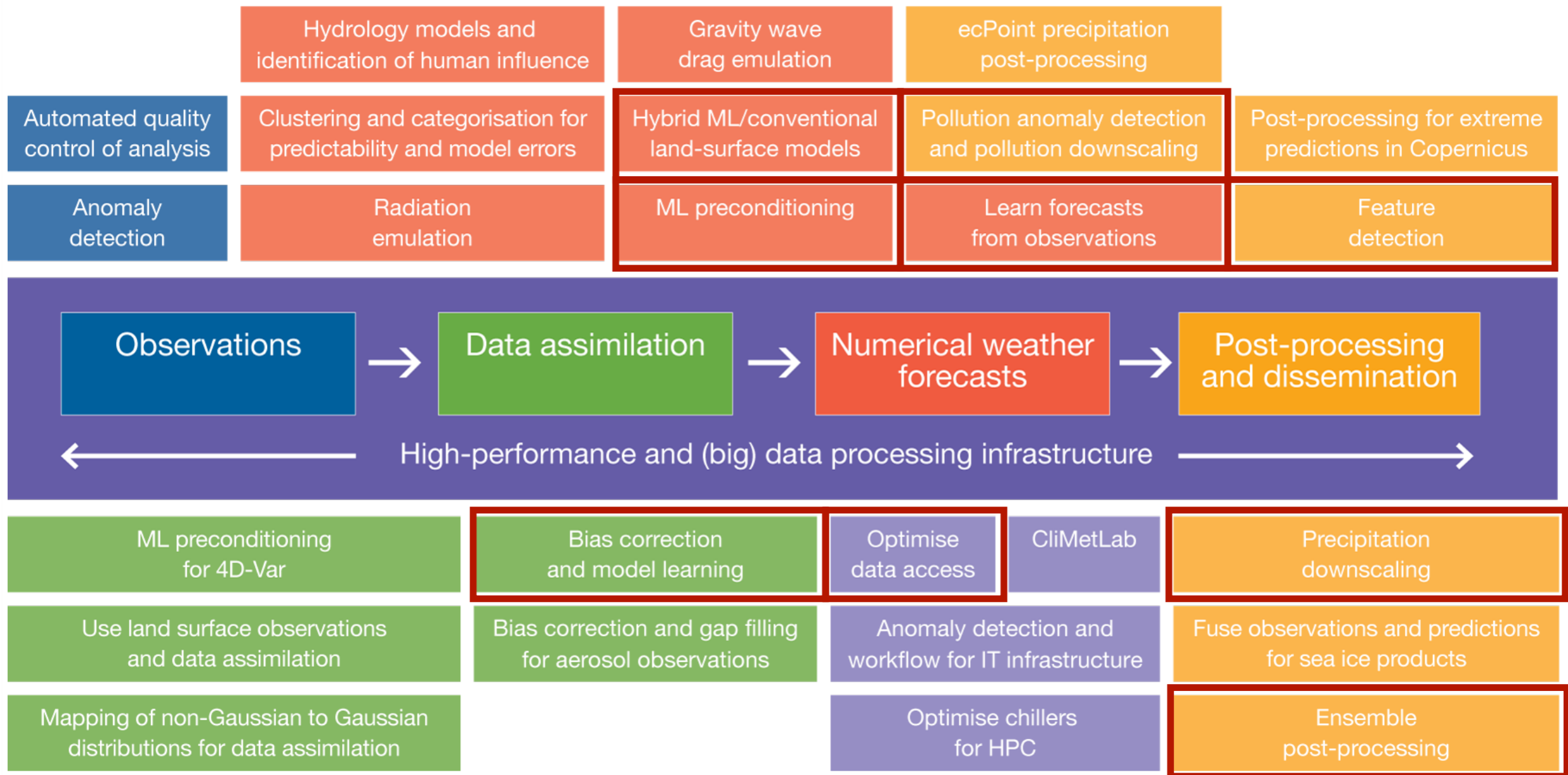
Attention:



A full set of possible applications

From P. Dueben (ECMWF): https://events.ecmwf.int/event/232/attachments/963/1688/Presentation_slides.pdf

Many application areas for machine learning across ECMWF



EMP² vs ClimaX: differences & similarities



6 February 2023!

ClimaX: A foundation model for weather and climate

Tung Nguyen^{*1,3}, Johannes Brandstetter², Ashish Kapoor¹,
Jayesh K. Gupta^{†1}, and Aditya Grover^{†1,3}

¹Microsoft Autonomous Systems and Robotics Research, ²Microsoft Research AI4Science, ³UCLA

ClimaX

EMP²

Both are foundation models based on visual transformers!

Investigating similar downstream applications

Trained on a randomised forecasting objective

Goal: reconstruct states in the future

using ERA5 on pressure level variables

deterministic predictions

single transformer

Concatenation of fields in the variable aggregation step

private company

BERT-style training adapted to scientific data

reconstruct *masked tokens* within a random hypercube

using ERA5 on model level variables

stochastic predictions

Model uncertainty quantification through newly defined statistical loss

stack of transformers

one transformer for each field, coupled with cross attention.

public research

→ The model is less
"forecasting oriented"

→ This is what ECMWF uses:
an eye into the integration
within their systems.

→ Plug and play approach:
new fields can be easily
integrated.