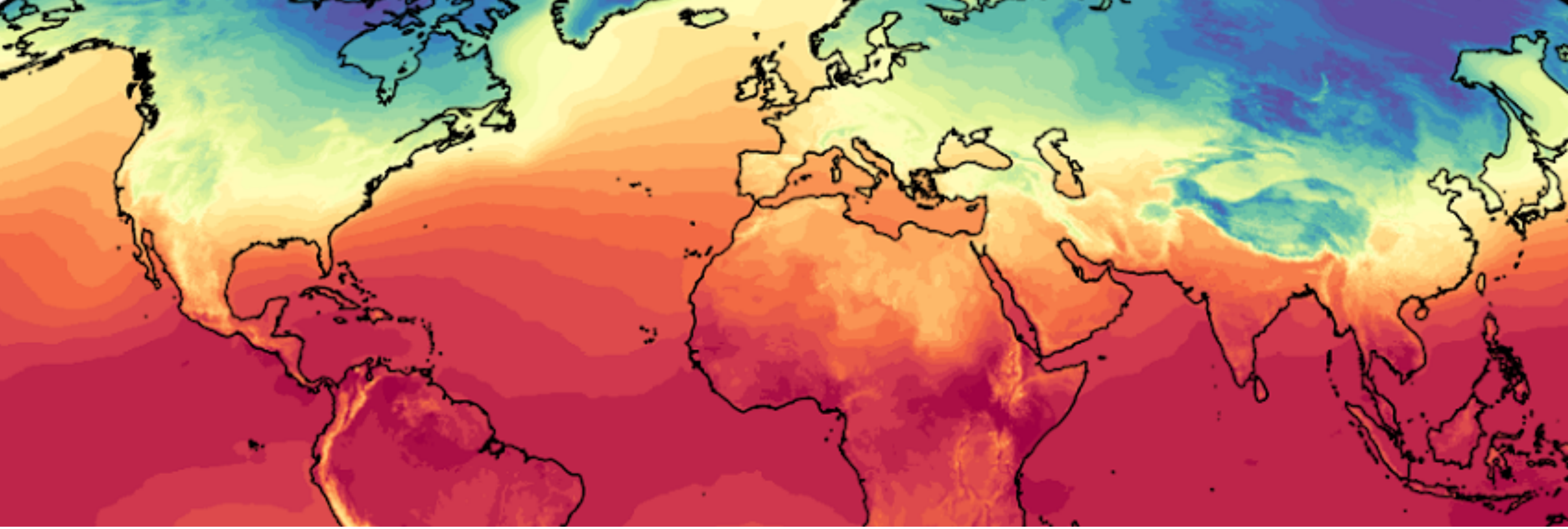


EMP^2:
Environmental Modelling and Prediction Platform

Christian Lessig, Ilaria Luise, Martin Schultz,
Alberto Di Meglio, Anna Ferrari, Sofia Vallecorsa et al.

CERN OpenLab Technical Workshop
16th March 2023

CERN | CIPEA CERN Innovation Programme on Environmental Applications

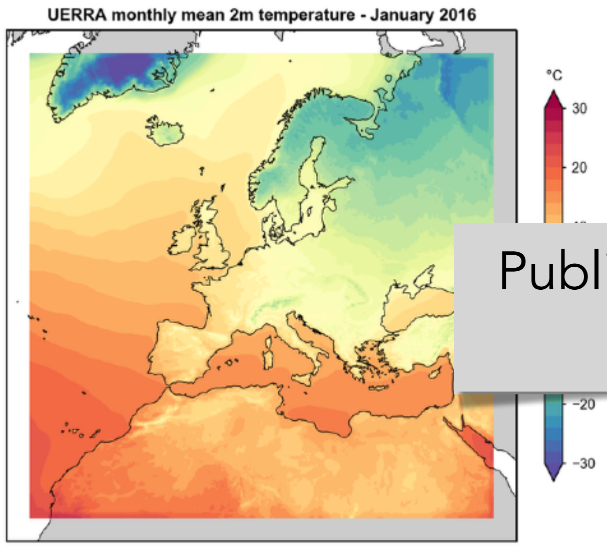
OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

JÜLICH Forschungszentrum

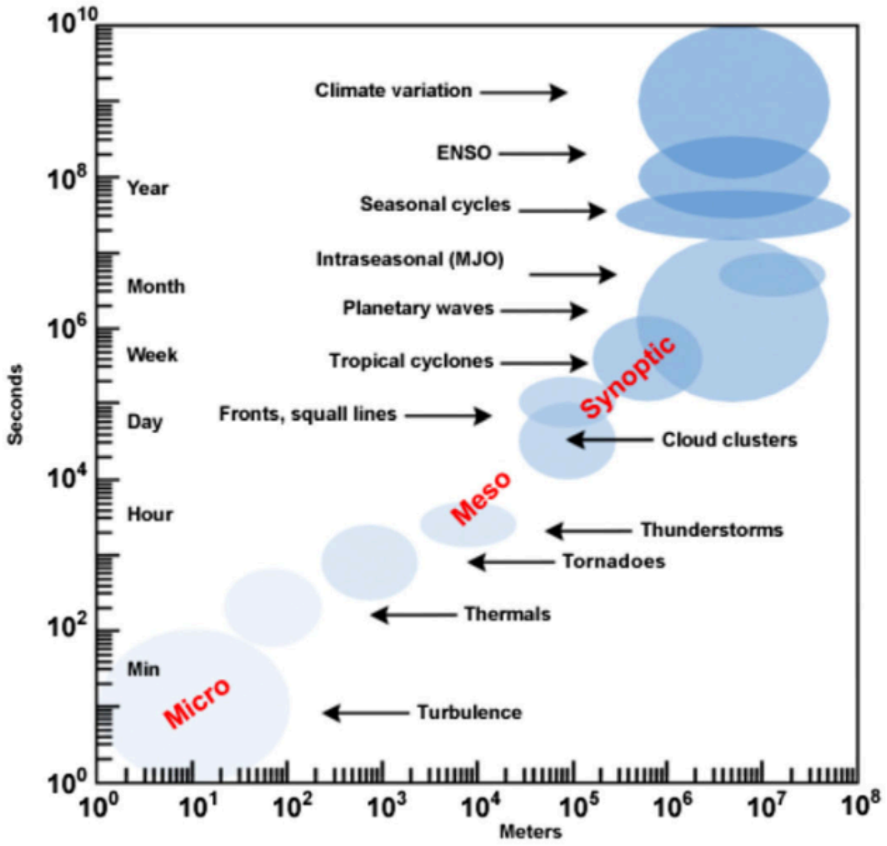
Motivation and scientific challenge

Atmosphere:

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



Publicly available datasets
eg. ERA5



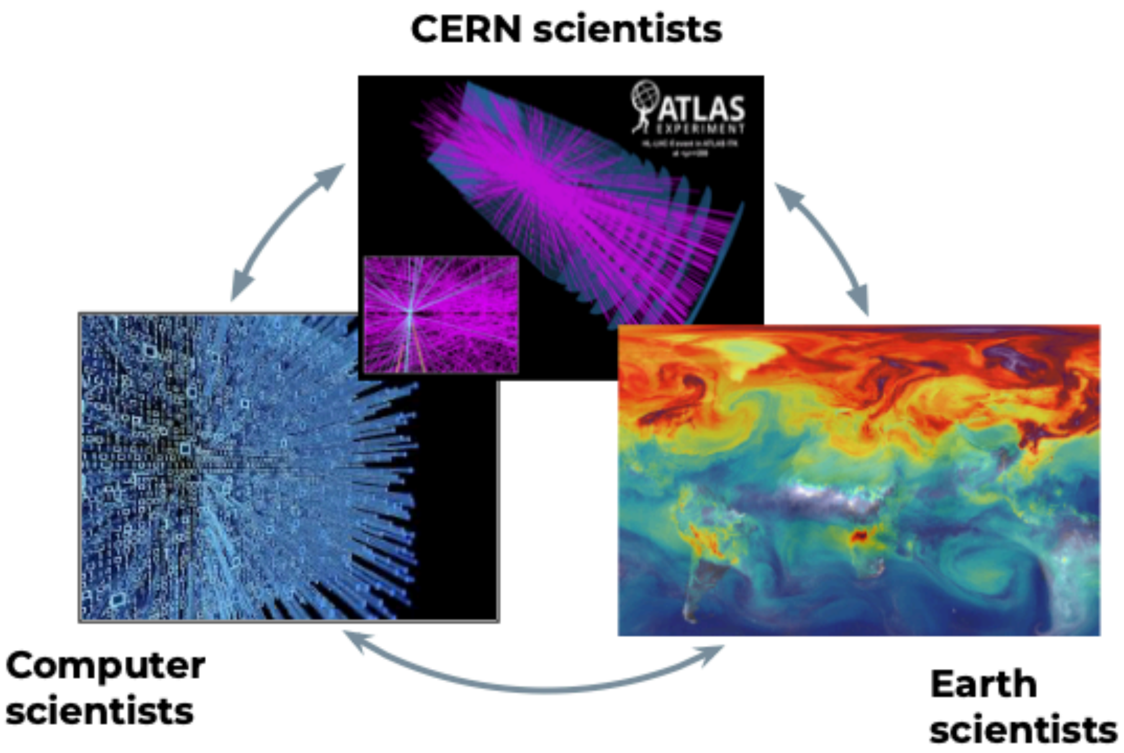
From V. M. Galfi, V. Lucarini, F. Ragone, and J. Wouters. Applications of large deviation theory in geophysical fluid dynamics and climate science. La Rivista del Nuovo Cimento, 44(6):291–363, 2021.

Developments in hardware
(GPU clusters)
and software (exa-scale ML)

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?

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

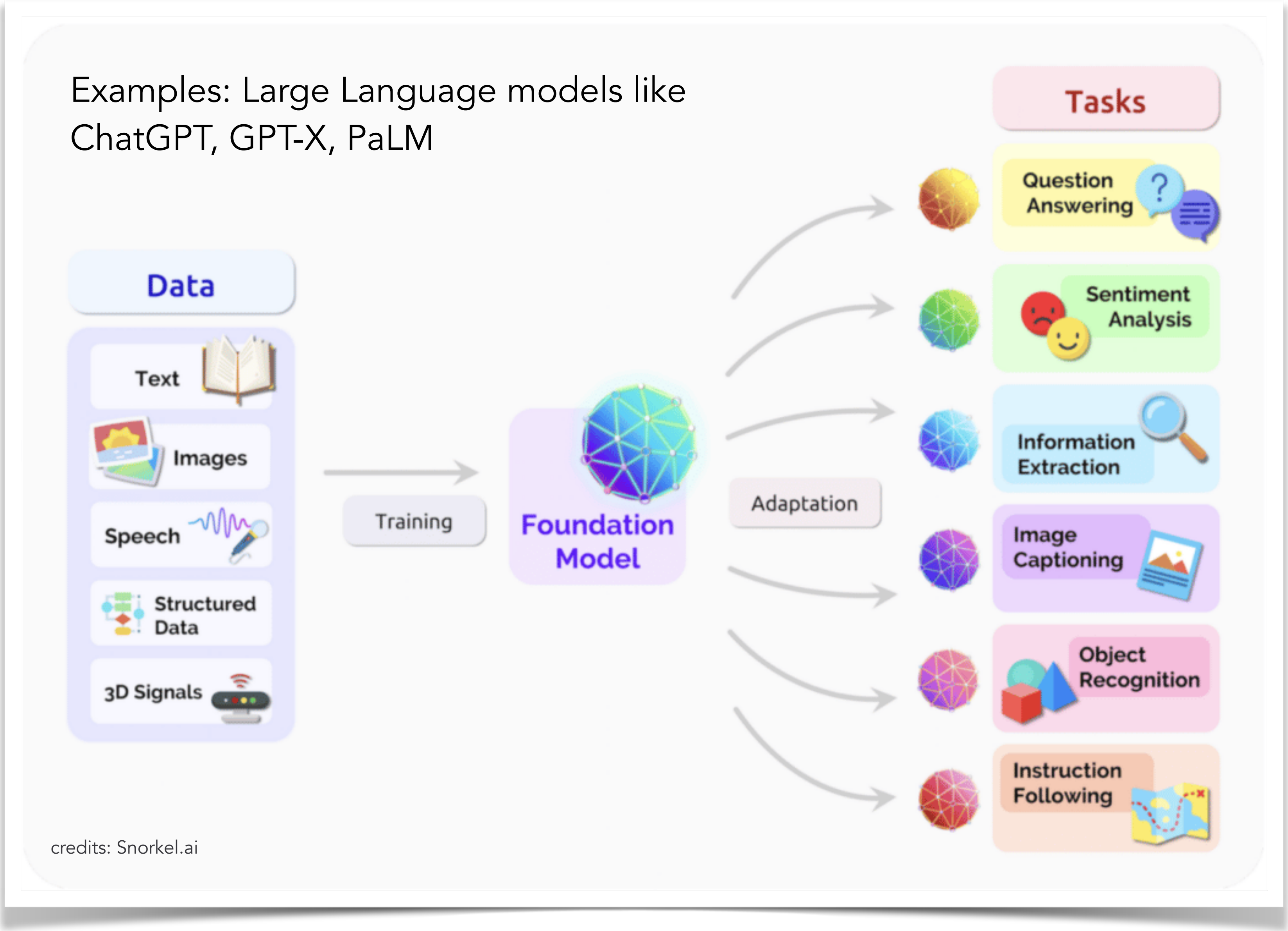
Common Goal:

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

Representation Learning

The EMP^2 model architecture

The beauty of foundation models

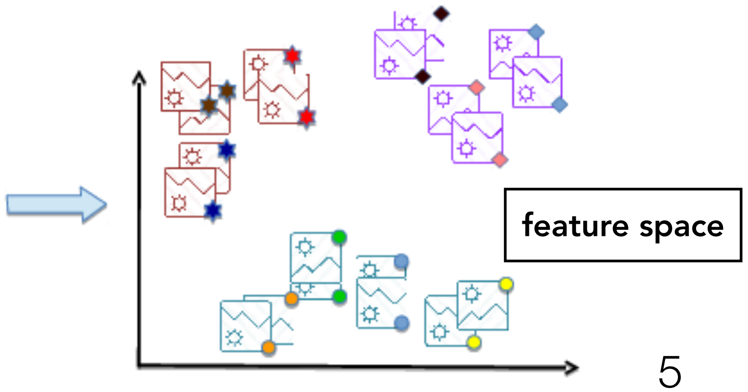


Examples: Large Language models like ChatGPT, GPT-X, PaLM

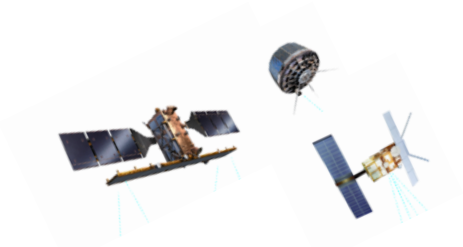
credits: Snorkel.ai

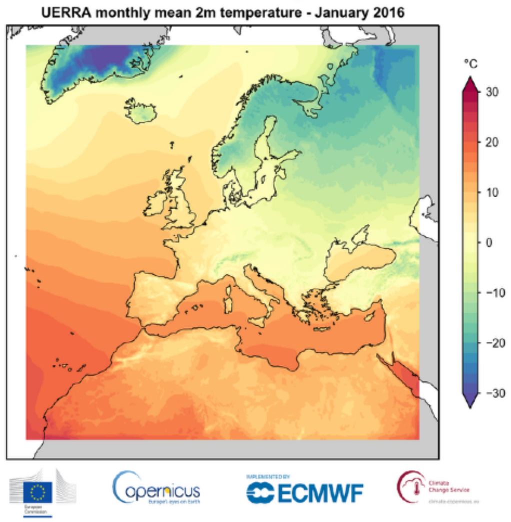
.. see Renato Cardoso's Talk on foundation models tomorrow at 9h30

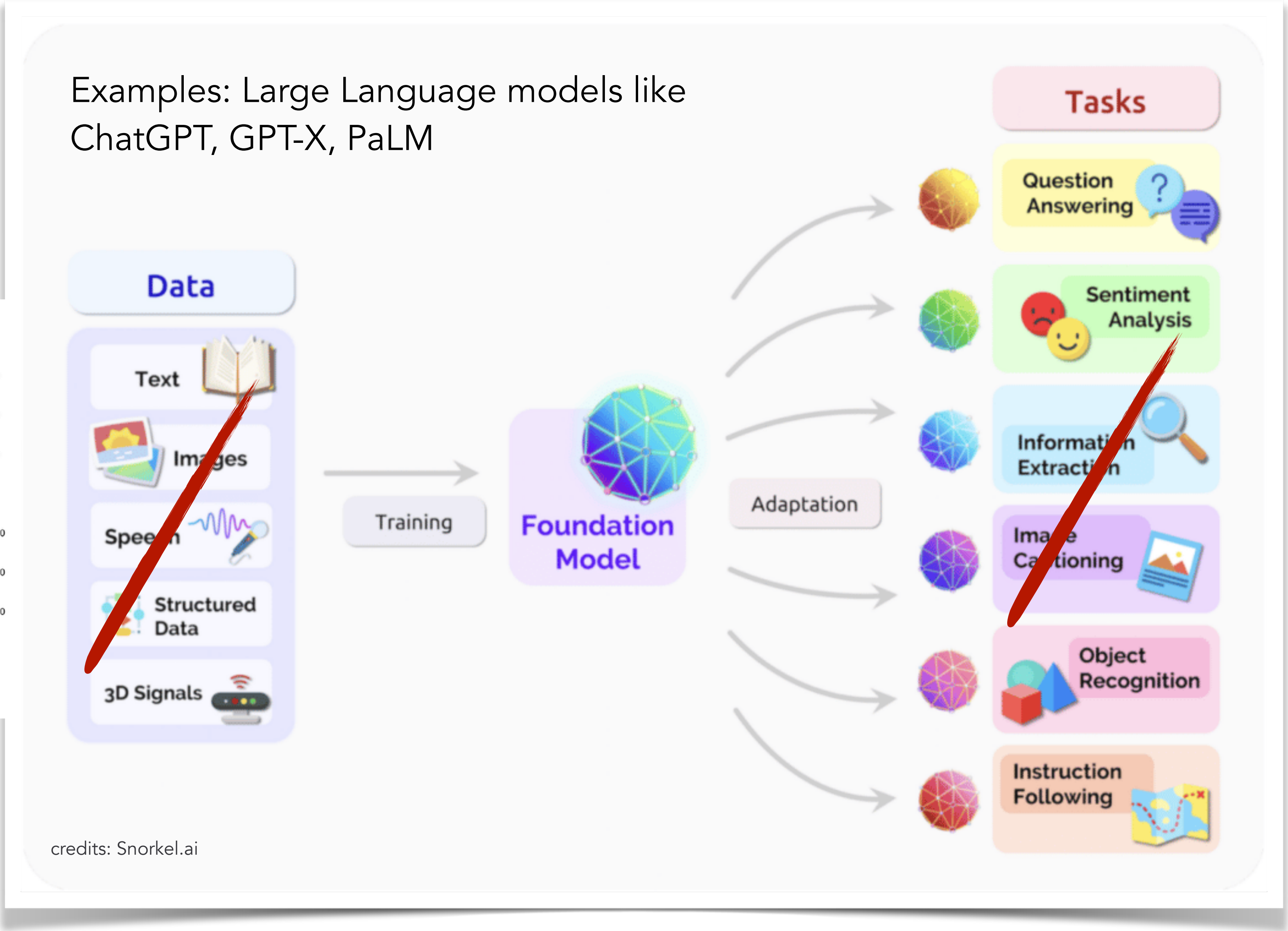
New concept: representation learning



The beauty of foundation models

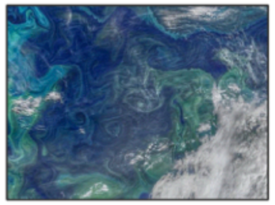




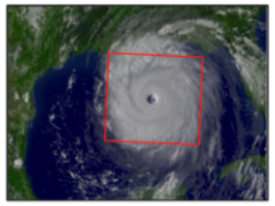


Examples: Large Language models like ChatGPT, GPT-X, PaLM

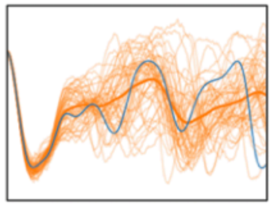
credits: Snorkel.ai



Forecasting



Impact analysis



Climate projections

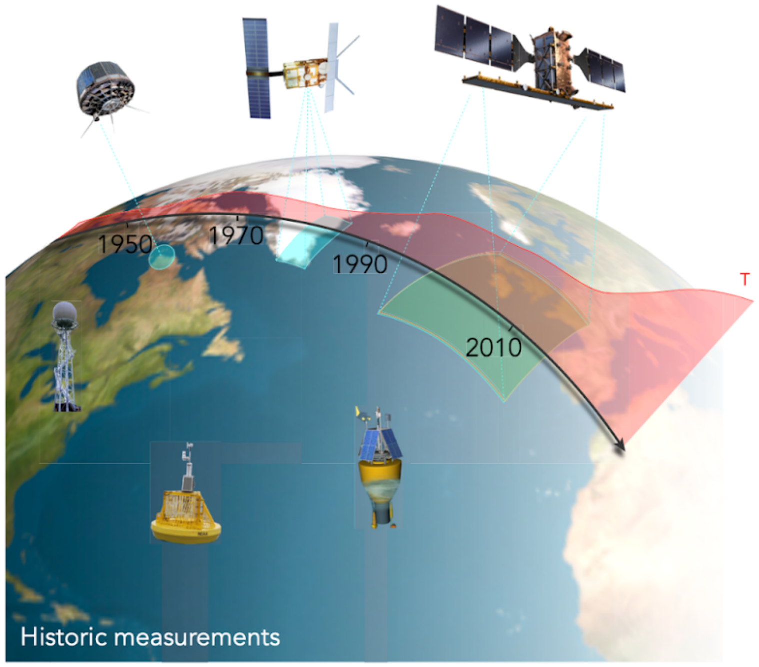
.. see Renato Cardoso's Talk on foundation models tomorrow at 9h30

Can we transfer the concept of representation learning from language models to fundamental science?

Learn a domain-specific but task-independent representation that is useful for a large range of scientific applications. Challenge: Need to deal with much more complex processes and datasets than in NLP

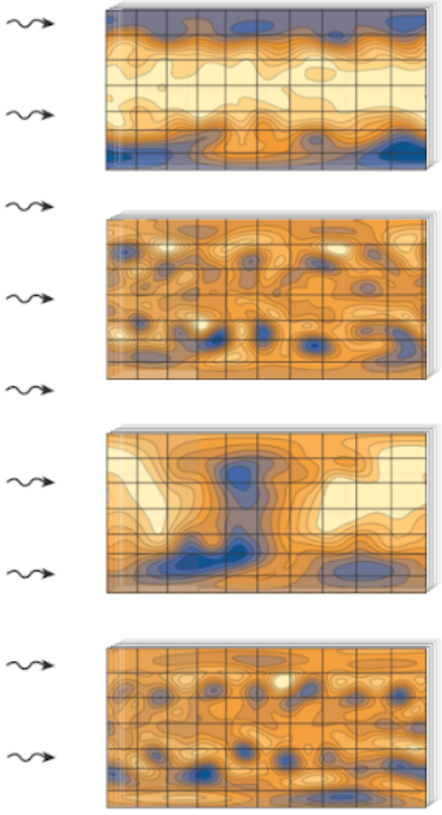
The project in a nutshell

First proof-of-concept of a machine-learning based global environmental model trained on terabytes of observational data



Observations

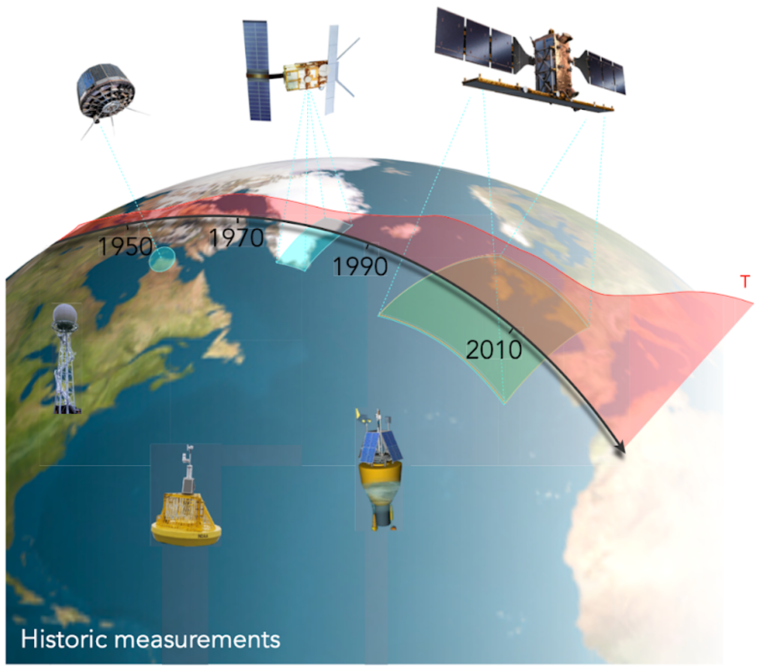
ERA5 reanalysis



500 TB

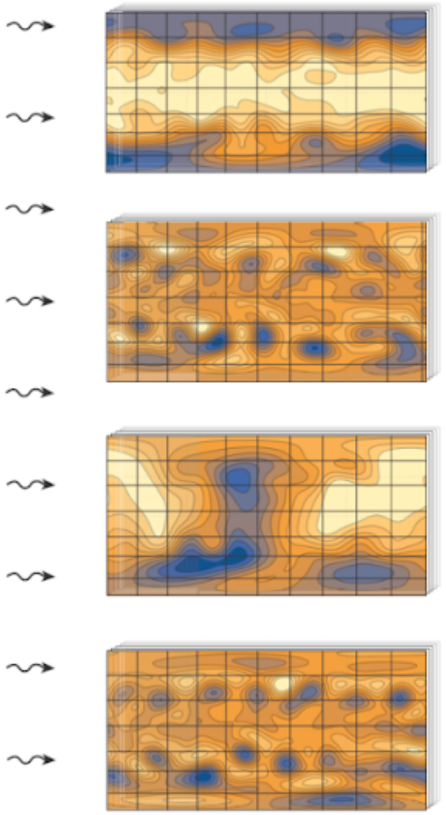
The project in a nutshell

First proof-of-concept of a machine-learning based global environmental model trained on terabytes of observational data

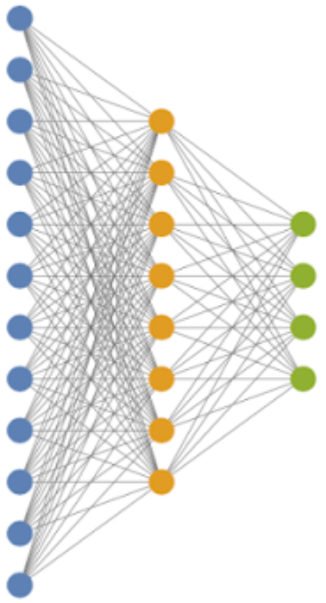


Observations

ERA5 reanalysis



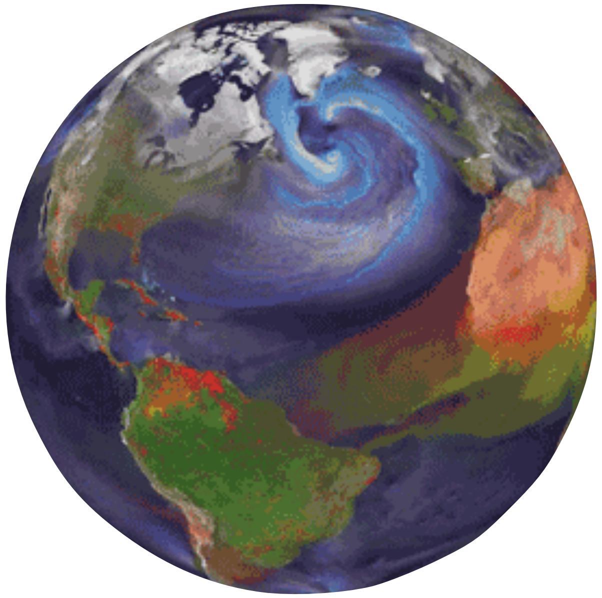
500 TB



large scale machine learning

Transformers architecture

Spatio-temporal representation of atmospheric dynamics



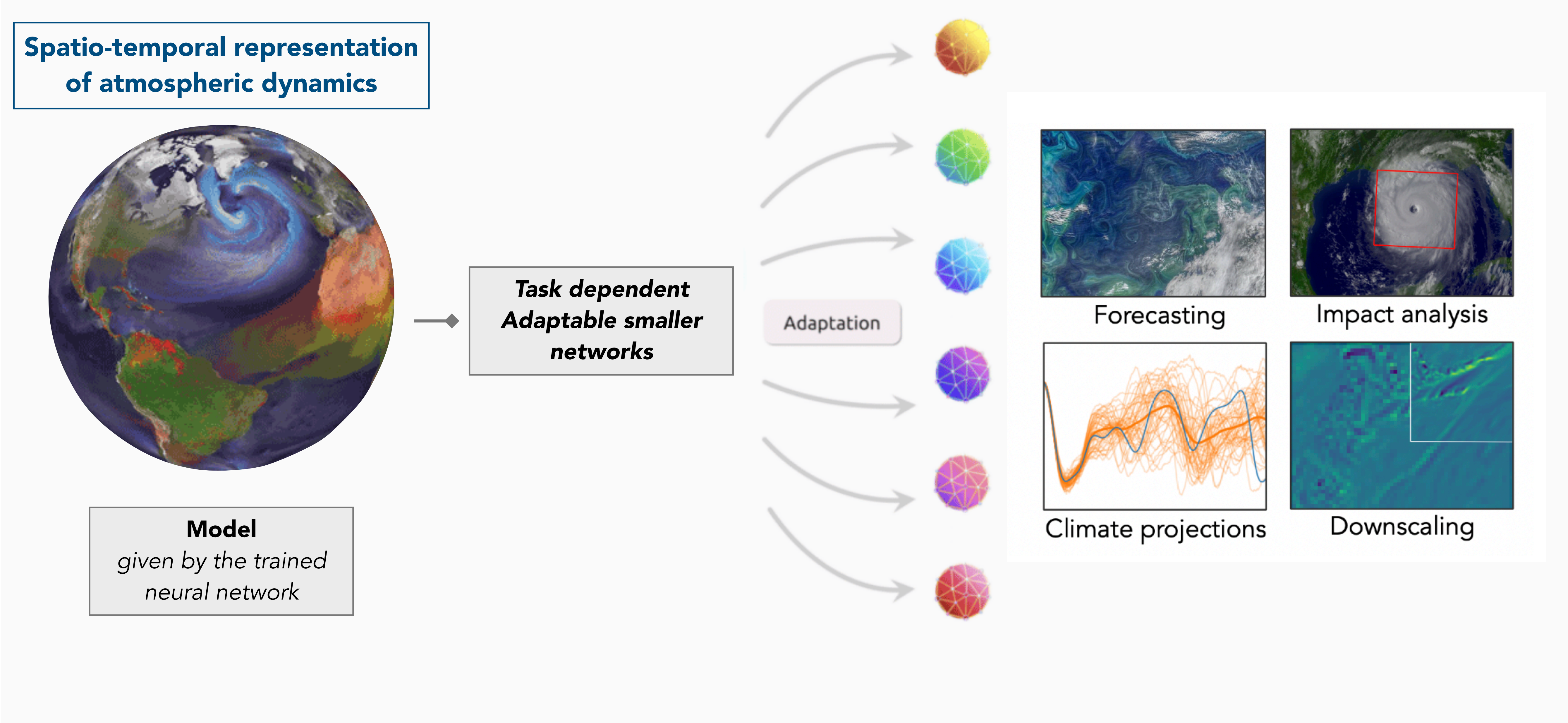
Model
given by the trained neural network

R&D at Juelich SSC:
6x10^5 GPU hours granted in 2022

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



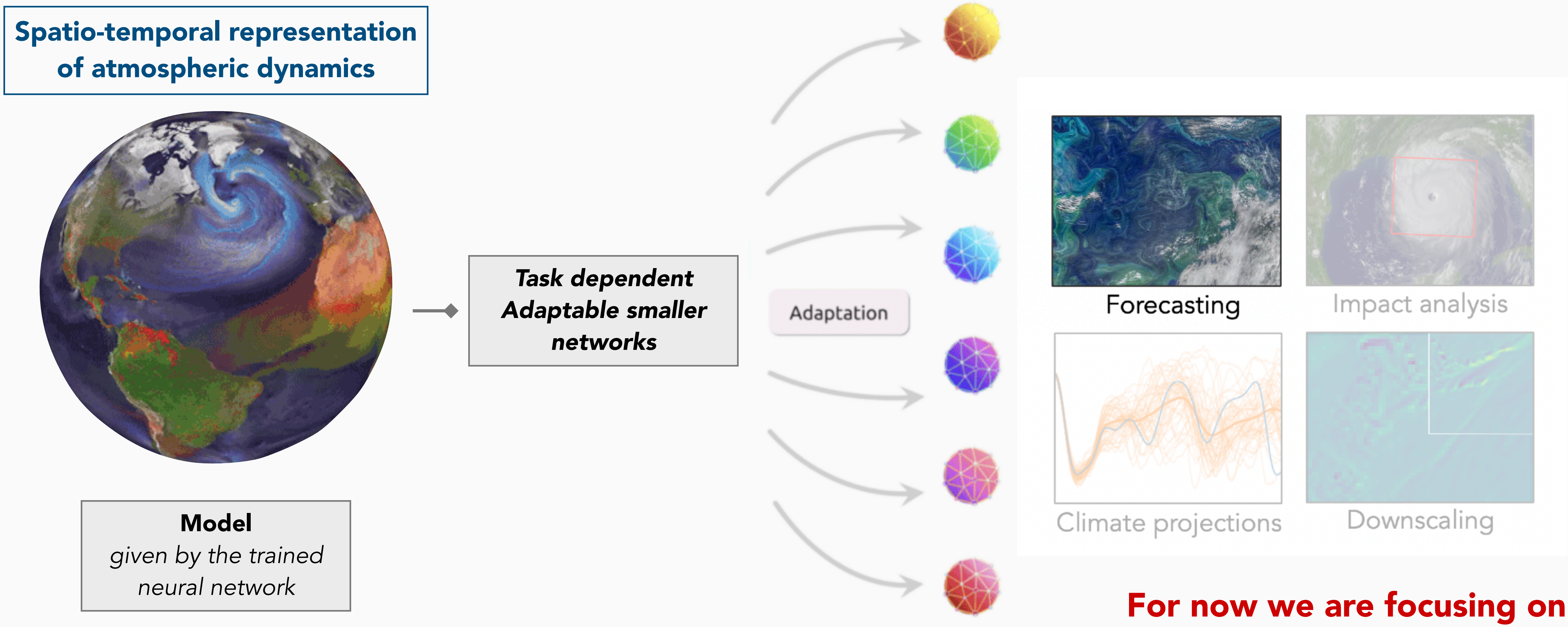
Spatio-temporal representation of atmospheric dynamics

Model
given by the trained neural network

Task dependent Adaptable smaller networks

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

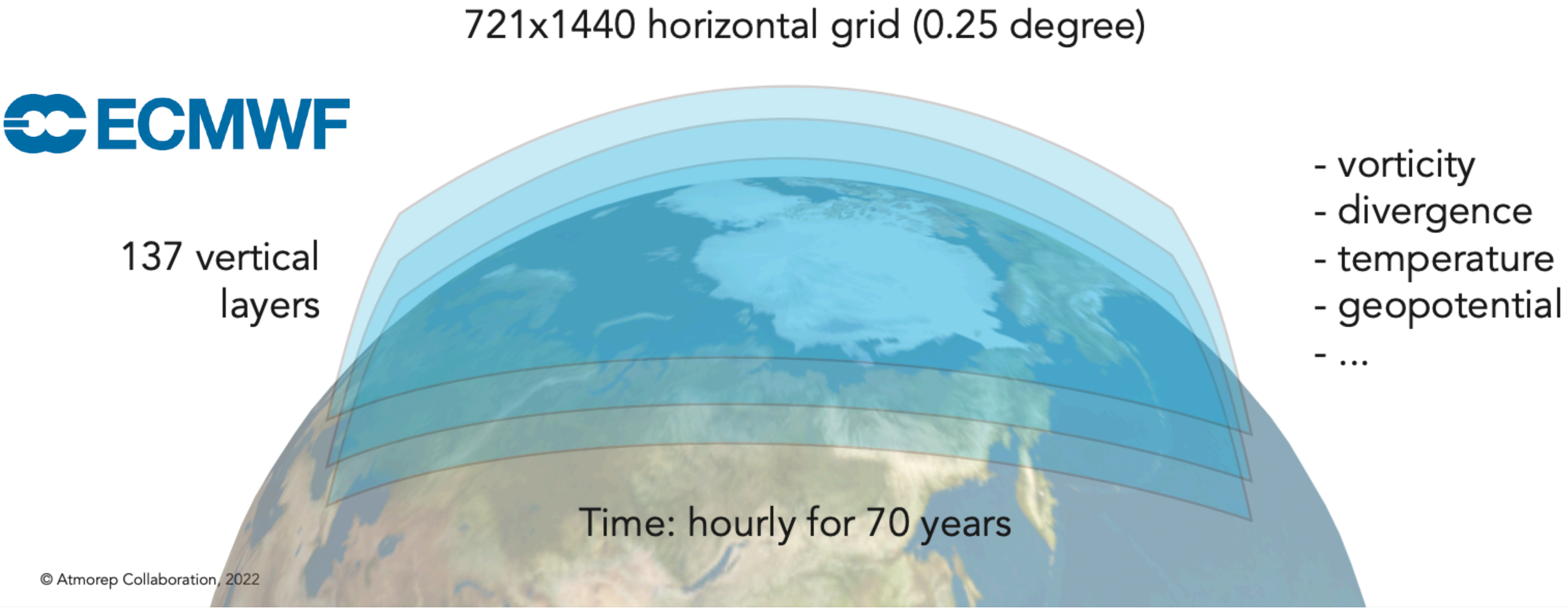
For now we are focusing on short term forecasting

The dataset

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, temperature, geopotential height, specific humidity, orography
- Space: 721 x 1440 x 6 vertical layers
- Time: **randomly sample** over 24 time steps per day for 365 days for 70 years

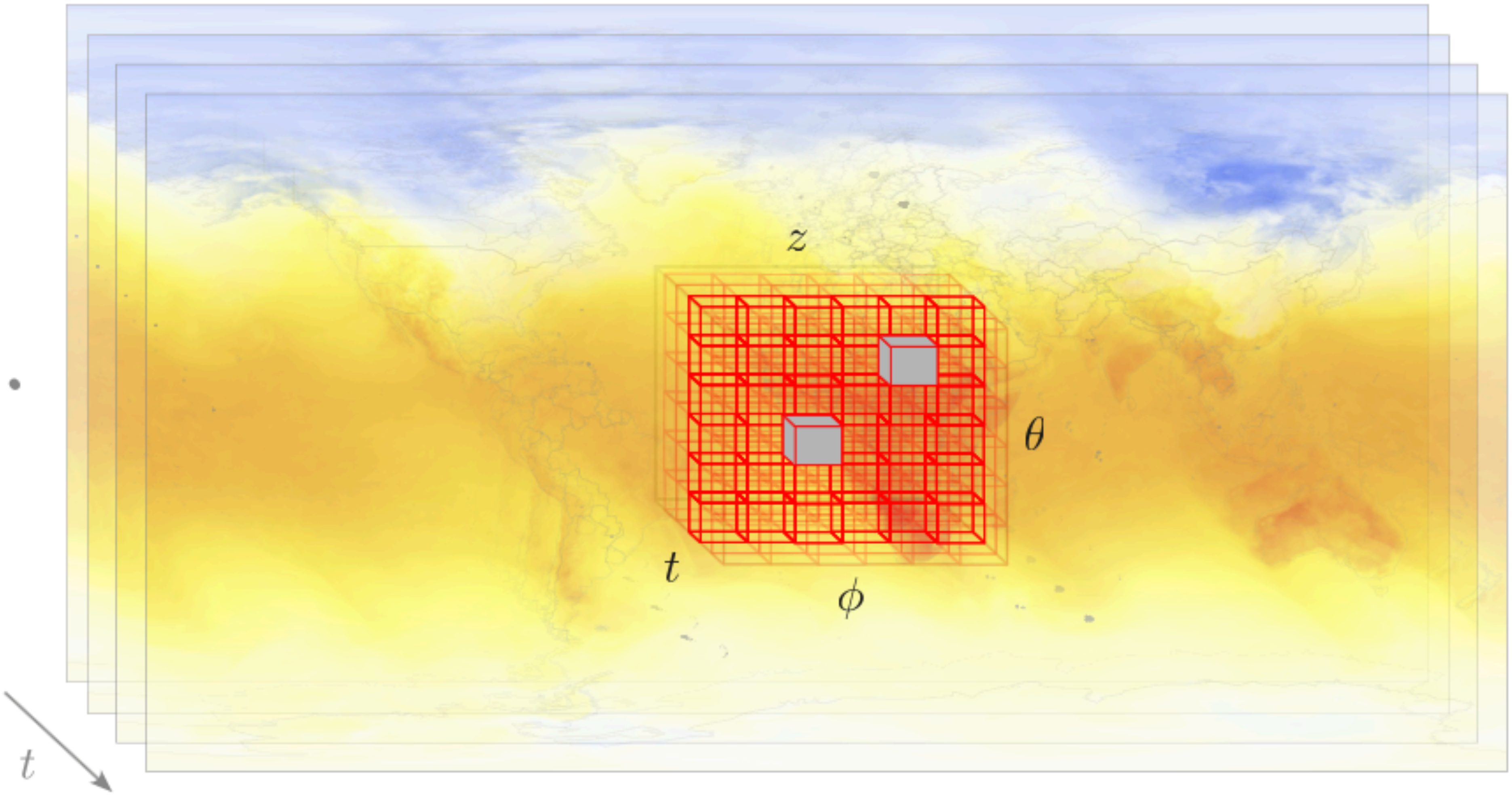


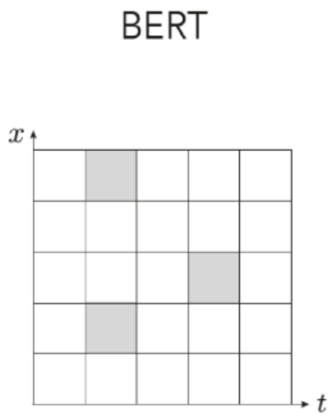
© Atmorep Collaboration, 2022

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

Preliminary results

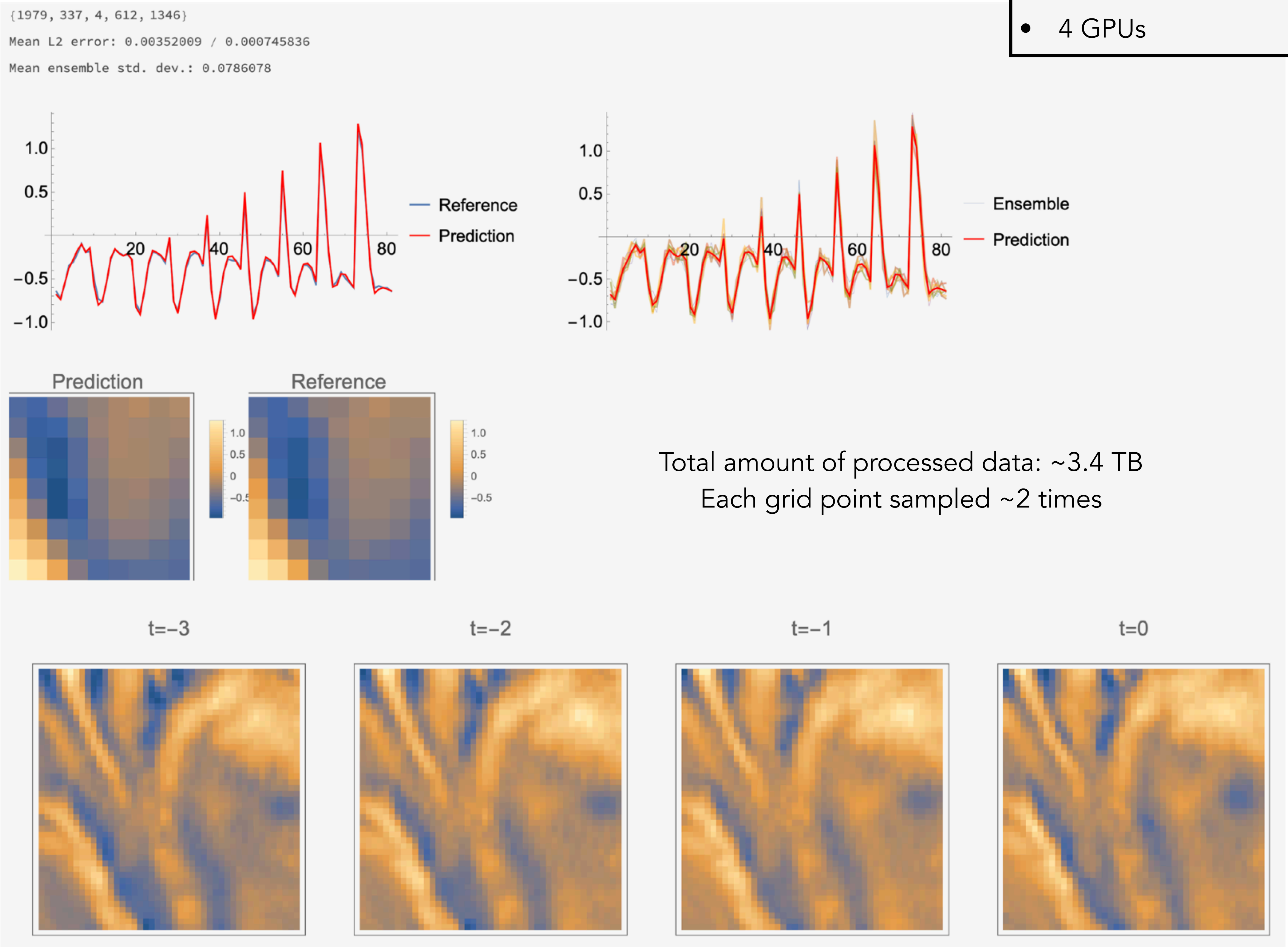
Vorticity 975 hPa (+ orography)

- 14 epochs
- 256 batch size
- 4 nodes
- 4 GPUs

1D comparison

2D comparison

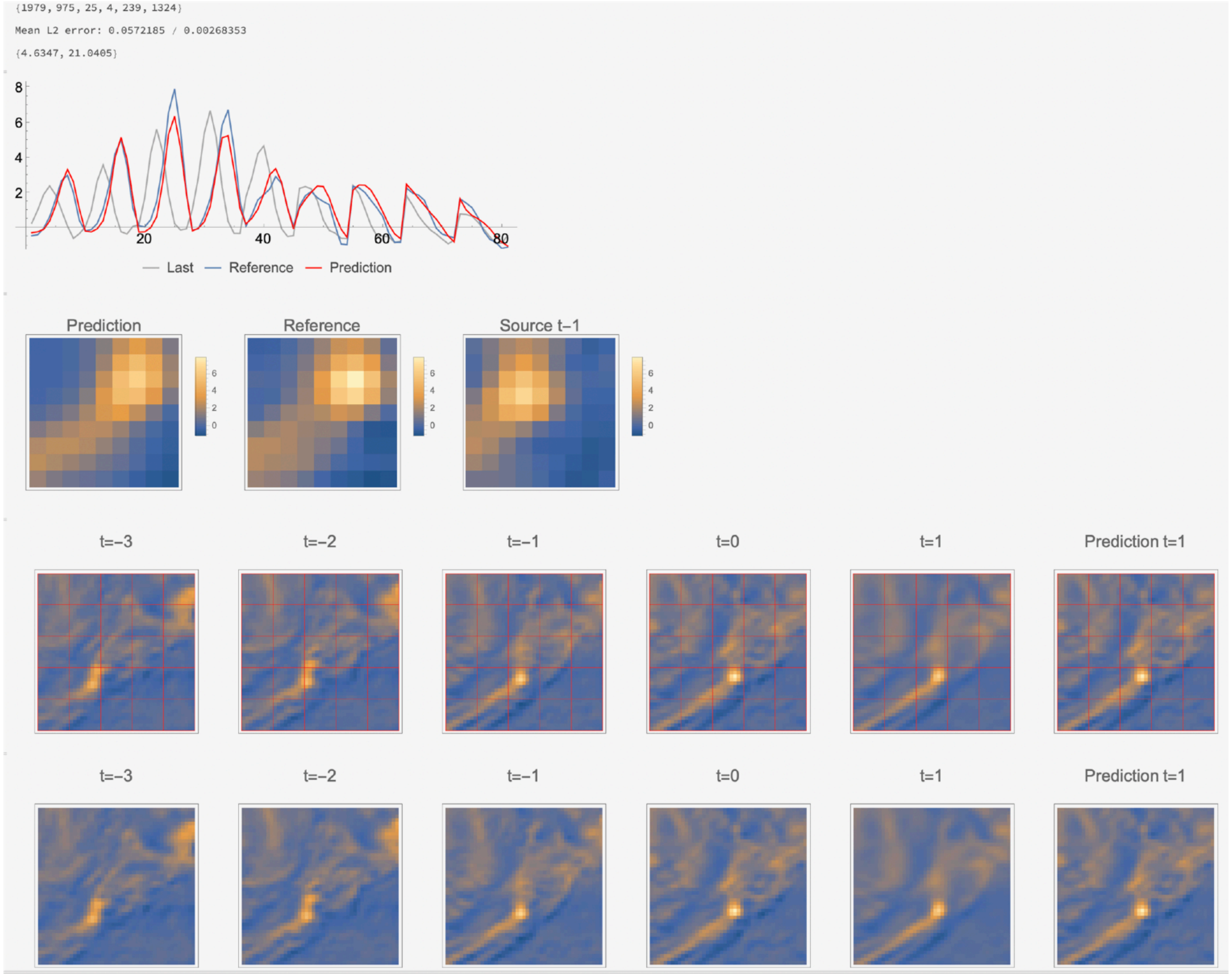
sampled cube



```
{1979, 337, 4, 612, 1346}
Mean L2 error: 0.00352009 / 0.000745836
Mean ensemble std. dev.: 0.0786078
```

Total amount of processed data: ~3.4 TB
Each grid point sampled ~2 times

Preliminary results: 1h forecasting



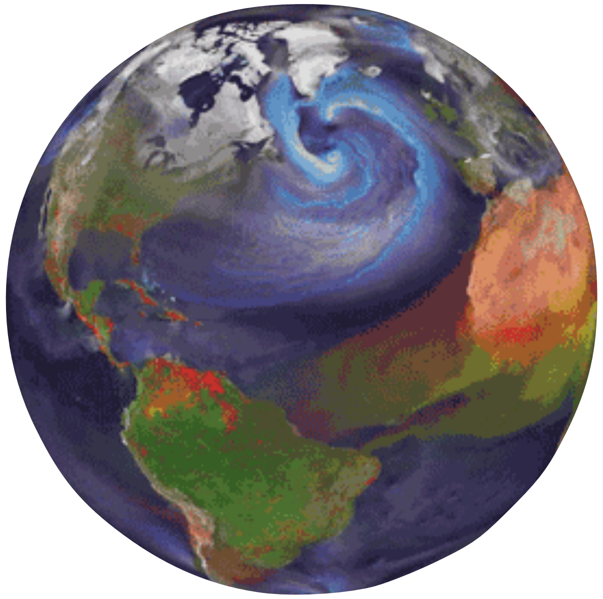
```
{1979, 975, 25, 4, 239, 1324}
Mean L2 error: 0.0572185 / 0.00268353
{4.6347, 21.0405}
```

Generalisation from HPC centres to clouds

Future: develop the API & the user interface

.. see last talk about the InterTwin project

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

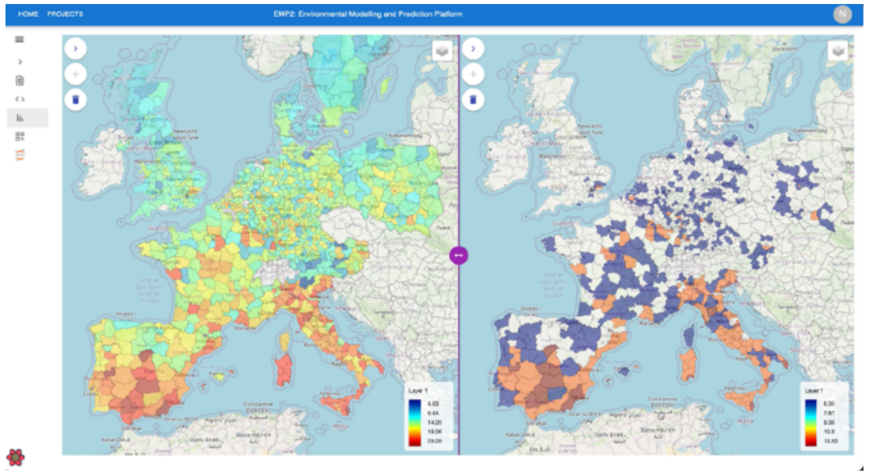


Onboard

(New CS4OD)

interTwin

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



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

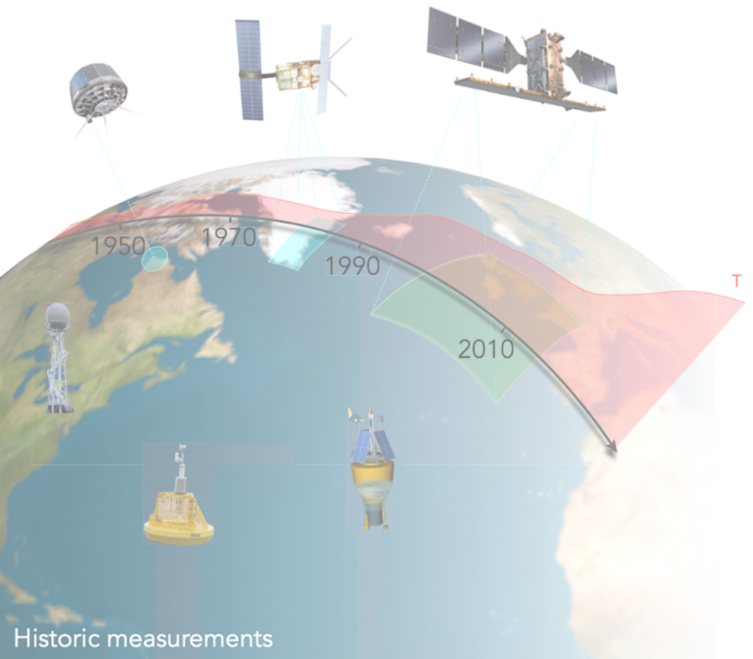
Conclusions

EMP^2: Environmental Modelling and Prediction Platform

- Exciting scientific challenges ahead on how to better exploit the large amounts of available unlabelled data using AI/ML.
- Transformers have been proven a powerful and scalable architecture. Can we use them for scientific applications to solve some of these challenges?

EMP^2 current status:

- Implementing the machine learning architecture. Now testing the multi-field and multi-level architecture. Long runs at JSC planned in the next weeks.
- Efforts to test the model on downscaling and bias correction applications are ramping up.



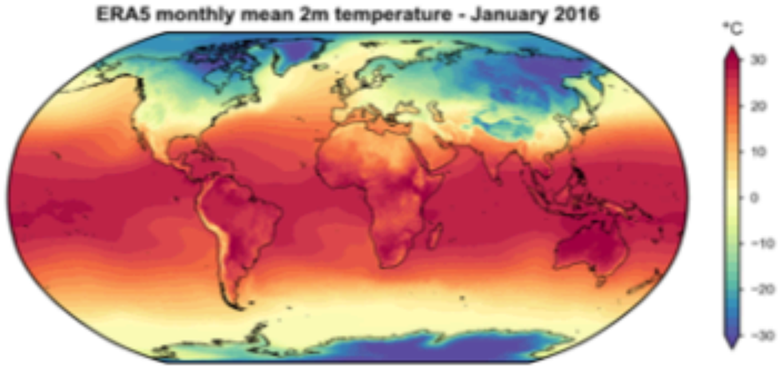
.. and some long term plans:

- Implementation in the Digital Twin engine as use case to test the InterTwin architecture

Questions for EMP^2 2.0:

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

AtmoRep

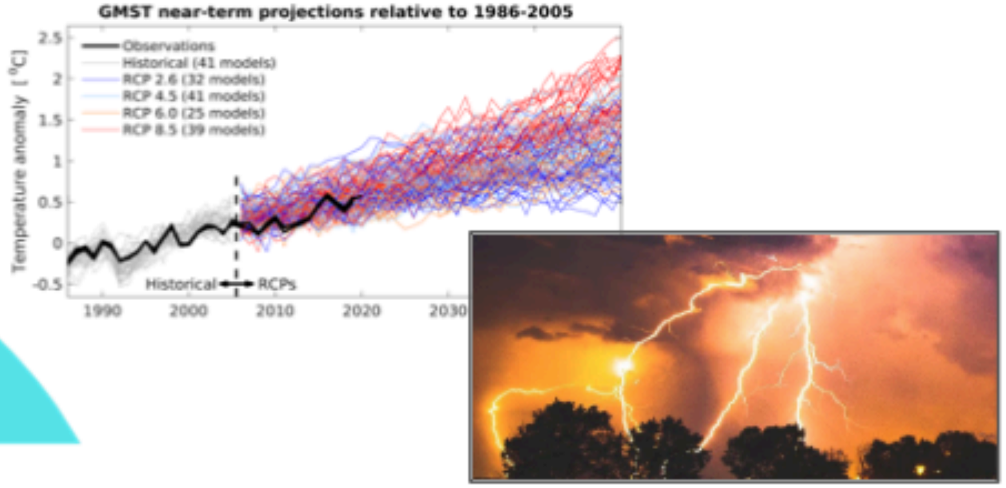


Technology

Operations

Collect

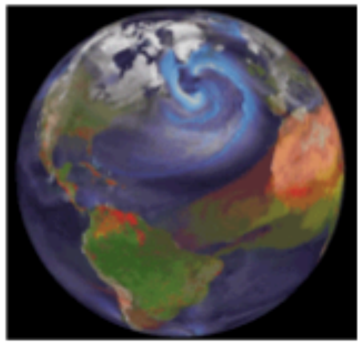
Aware



Artificial Intelligence

Big Data Analytics

Compute

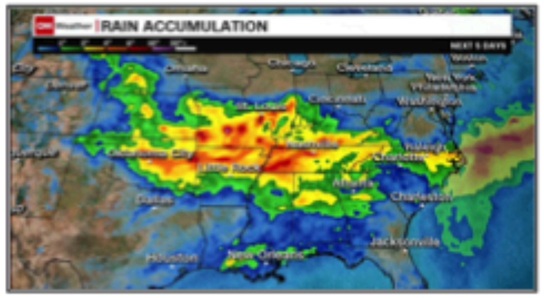


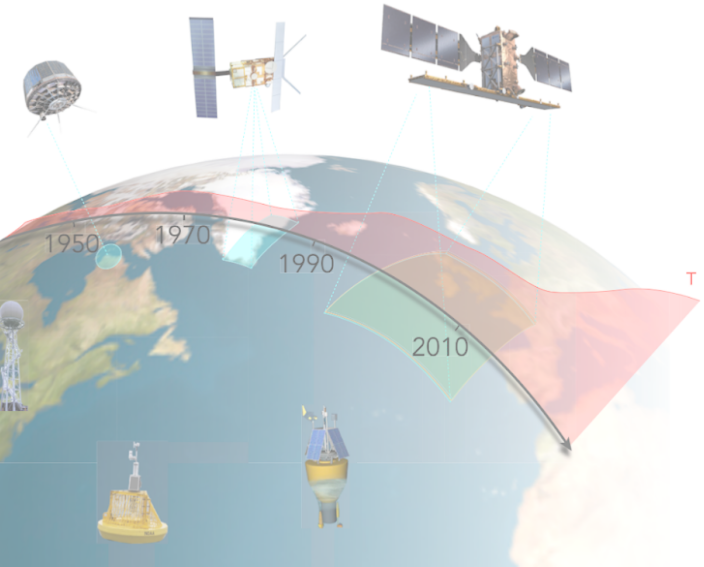
Respond

Visualization Tools

Visualize

Predict



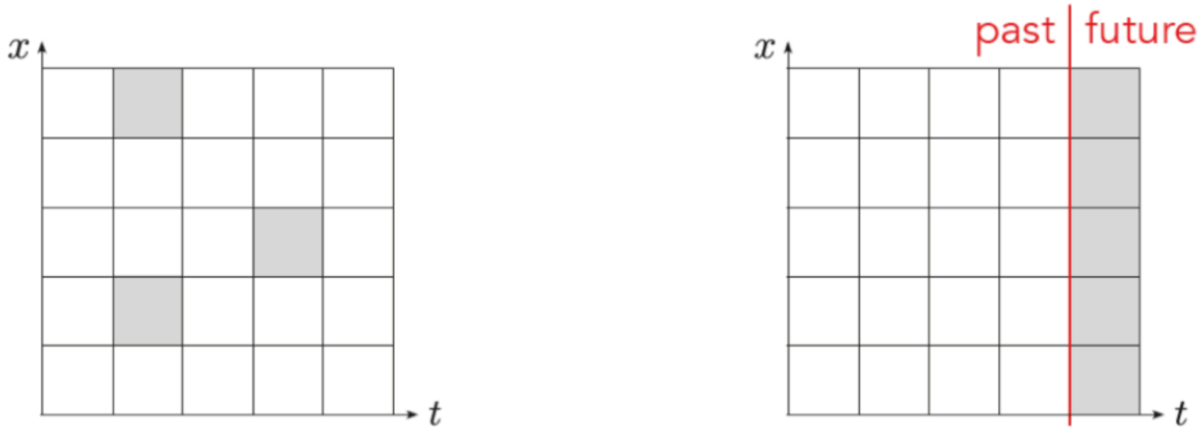


Zero shot forecast

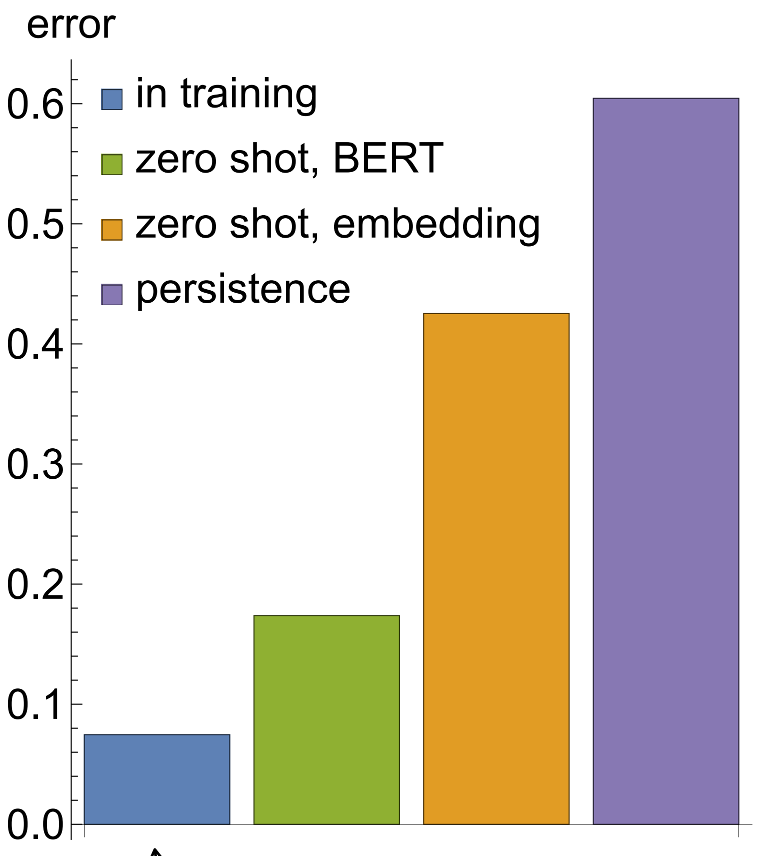
Vorticity (+ orography)
- 256 batch size
- 1 node, 4 GPUs

BERT

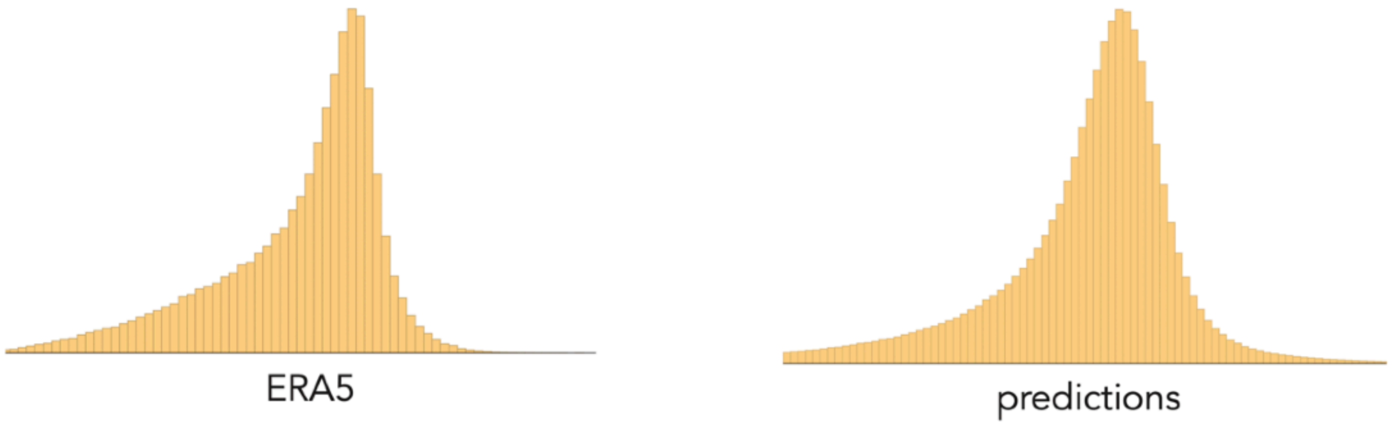
BERT-Forecast



Move from BERT to BERT "forecast"



Cape town: historgram of vorticity



The network learns the representation, clear improvement w.r.t. persistence