EMP²: Environmental Modeling and Prediction Platform

Climate and societal changes

Earth is facing a planetary emergency:
⇒ climate is heating up too quickly
⇒ loss of habitat for more than 1 million species
⇒ pollution continues to poison water, air and land

Science

Concrete actions start at local scale:
in small communities & cities

Policy making

Global scale  Local scale
Local communities & climate change:

Data should be collected at a potential site for 2–3 years

Analyse data on wind speed, flow constance.

Final decision

Concrete example:
A local administration needs to select a location for a wind farm

Classical models are too complex and expensive for local communities
⇒ data driven models are cheaper and faster

How to meet the best decision with the lowest impact on the planet?

link to blog article: here
Local communities & climate change: digital Twin

**Concrete example:**
A local administration needs to select a location for a wind farm

**Digital models:**
enable efficient, outcome focused decision making

**New solution:**
**Digital Twin models**

Environmental model for predicting winds (global model)

Geographic information of the chosen sites

Wind speed forecasting for each site

link to blog article: [here](#)
Digital Twin: main ingredients

Data availability
⇒ observational data for the process to be modelled

Accuracy
⇒ mathematical model of the observed process needs to be accurate

Efficient computational infrastructure
⇒ need supercomputing infrastructure for accurate prediction

Visualisation and Accessibility
⇒ targeted user friendly interface of the model
The EMP² project:

Goal: Proof-of-concept for a data-driven digital twin model of the atmosphere for environmental applications using 70 years of observational data

- big-data handling
- ML expertise
- Earth science expertise
Learning from observations

Use 500 TB of publicly available Earth data observations

70 years of ERA5 = lots of data about short to medium term phenomena
Learning from observations

70 years of ERA5 = lots of data about short to medium term phenomena

Exascale training on supercomputer using newly introduced machine learning architectures

Transformers ML architectures (scalable to large datasets, adaptable)
Learning from observations

Outcome: global scale representation learning of the atmosphere dynamics

Transformers ML architectures (scalable to large datasets, adaptable)

70 years of ERA5 = lots of data about short to medium term phenomena

500 TB

ERA5 reanalysis

large scale machine learning

learned spatio-temporal representation of atmospheric dynamics

First ML-based global environmental core model of the atmosphere in 4D

Copernicus

Historic measurements

Copernicus

Jülich Forschungszentrum
Environmental modelling and societal challenges

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

Use the model as backbone for a set of global and local applications

Digital Twin

Technology

Operations

Collect

Aware

Predict

Respond

Compute

Visualize

Visualization Tools

Big Data Analytics

Artificial Intelligence
Applications: local scale program

- **HSE Unit @ CERN**
  - **Occupational Health & Safety and Environmental Protection Unit (HSE)**
  - Application: Understand the most effective placement of a wind park and the potential energy output.
  - Validation: Use data from the CERN meteorological stations to generate predictions for *CERN environmental monitoring applications*.

- **Local communities in the Geneva area, e.g. regional administrations**
  - Application: Understand the most effective placement of a wind park and the potential energy output.

- **UN satellite agency (UNOSAT) at CERN, insurance companies**
  - Application: Improve the accuracy of the *risk assessment* and intervention plans, e.g. for *flooding or wildfires*.
  - Validation: Use data from the CERN meteorological stations to generate predictions for *CERN environmental monitoring applications*.
Applications: Global scale impact

Improve super-resolution of simulations ⇒ Model downscaling

Improve weather forecasts ⇒ e.g. earlier and more accurate prediction of extreme events

Improve classification of extreme events ⇒ e.g. inform risk management and policy decisions

Disaster mapping + aid ⇒ shifting local regions; link between climate and, e.g., outbreaks, food security (longer term)

improve alert systems and/or better inform risk management and policy decisions through future IPCC related studies

Collaboration with Climate Scientists within the Climate21 group
Visual Interface: CS4OD

Platform for data exploration
⇒ Facilitate data exploration and results visualisation to diverse stakeholders

Platform for collaborations
⇒ Facilitate scalable linking, collection, sharing, and storage of environmental data from open access repositories, crowd-sourced projects, institutional repositories

Platform for open data
⇒ accessibility of data from different communities, based on FAIR principles and data sharing best practices
Ideally: first results available already in summer 2023.

~12 months: First proof-of-concept on partial dataset 
& model reliability tested on local applications

~15 months: First results on full dataset 
+ Start testing on global applications

Dedicated effort and funds from the CIPEA programme:
- Estimated at the level of one graduate researcher for the duration of the program (2 years)
- Development of the environmental modeling core and integration in CS4OD
- Validation and application of the models on the use cases (local and global)
⇒ **Final goal:** develop a proof-of-concept of Environmental Digital Twin for policy making

⇒ Multidisciplinary project involving CERN, climate scientists and computer scientists

⇒ **Modeling and prediction core** based on large scale machine learning architecture trained on supercomputing infrastructure.

⇒ First proof-of-concept of large scale training on 70 years of observational data!

⇒ Exploit existing ongoing work like interTwin project and CS4OD within the CERN IT Department

⇒ Test on a set of applications relevant to addressing climate change at a local and global scale!
Backup
Learning from observations

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

Single environmental modeling core as common backbone for a wide range of applications.

Global coverage

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

Flexible task-specific adaptation of the model realised, for example, through smaller and simpler neural networks.
Learning from observations

Use the model as backbone for a set of global and local applications

70 years of ERA5 = lots of data about short to medium term phenomena

Transformers ML architectures (scalable to large datasets, adaptable)

First ML-based global environmental core model of the atmosphere in 4D

ERP5 reanalysis

large scale machine learning

learned spatio-temporal representation of atmospheric dynamics

Applications

- Downscaling
- Predictability
- Classification
- Missing data
- Disaster support
Timeline & Resources

Ideally: first results available already in summer 2023.

~12 months: First proof-of-concept on partial dataset & model reliability tested on local applications

~15 months: First results on full dataset + Start testing on global applications

Longer timescale
- Include ocean data or use raw data (e.g. satellite data from ESA, EUMETSAT)
- Broad impact applications: e.g. improve simulations, disaster mapping

Develop Visual Interface and finalise applications with stakeholders
Applications

**Downscaling** of atmospheric data, e.g. to assess the **local impact of climate change**.

**Classification** of **extreme events**. Study how their **statistics** change over time and how it is affected by climate change.

**Improve disaster mapping and aid** through increased dynamical understanding and targeted satellite monitoring.

**Interpolation** of **sparse or corrupted measurements**, e.g. airplane data and measurements occluded by clouds.

**Improve numerical weather forecasting**, and error analysis in **existing weather and climate simulation models**.
AtmoRep: Learned spatio-temporal representation

Learn where to look to facilitate image processing and interpretation

Longer term applications: Improve disaster mapping and aid through increased dynamical understanding and targeted satellite monitoring

Satellite data

Rainfall Distribution

Malaria outbreak support

21
The Team

CERN IT Department:
- Anna Ferrari
- Alberto Di Meglio
- Sofia Vallecorsa

Researchers:
- Christian Lessig - Magdeburg University
- Ilaria Luise - Stony Brook University, CERN
- Martin Schultz - Juelich Supercomputing Center
- Maike Sonnewald - Princeton University - Washington University
- Aneesh Subramanian - University of Colorado Boulder

PhD students and post-docs:
- Clara Betancourt - Juelich Supercomputing Center
- Bing Gong - Juelich Supercomputing Center
- Sebastian Hoffmann - Magdeburg University

Affiliated researchers: Claire Monteleoni (UC Boulder), Donata Giglio (UC Boulder), Abigail Bodner (NYU-Courant), Yi Deng (GeorgiaTech), Julien Brajard (NERSC)