A data-driven parametrisation of atmospheric small scale processes

Janika Rhein

***Abstract*— Although it is not required to write a report about this project, this will give an overview about the motivation, approach and results. Currently some models are still training on the full dataset (or at least a much larger subsample). I will add the results from this here in the near future. For now results can be found mainly in "working" notebook...**

I. INTRODUCTION

Climate modeling is essential for understanding and predicting the effects of climate change. However, the complexity of the climate system requires simplifications in numerical models to maintain computational feasibility. One key challenge is the representation of small-scale processes such as turbulence, cloud formation, and radiation, which cannot be explicitly resolved in long-term climate simulations. This project explores the potential of artificial intelligence (AI) to improve the parameterization of these processes, aiming to enhance the accuracy and efficiency of climate models.

A. opencampus and Kaggle

This project was developed as part of the "Intermediate Machine Learning" course at opencampus SH, a non-profit organization focused on startups and machine learning. opencampus offers hands-on courses to complement the theoretical approach of university education. These classes are open to everyone and are free of charge for students, retirees, and unemployed individuals. The organization's main goals include providing free education, building professional networks, and helping participants connect with potential employers while broadening their horizons.

The "Intermediate Machine Learning" course aims to deepen students' knowledge of machine learning by focusing on state-of-the-art practices, enabling them to work in AI-related fields. Each course at opencampus is accompanied by a semester project, which students complete either individually or in groups. For this course hese projects are based on Kaggle competitions—Kaggle being an online AI community that hosts challenges proposed by individuals or companies to encourage problem-solving through AI solutions. While these are competitions, knowledge exchange and collaboration among participants are encouraged.

Students in the course are free to choose a competition to work on, even if it is no longer active. Kaggle provides the dataset, task description, and evaluation metrics. Given my background in climate sciences, I selected a related project: LEAP - Atmospheric Physics using AI. This competition's dataset originates from the atmospheric component of an operational climate model. The objective was to train an AI-based parameterization for atmospheric small-scale processes.

B. Numerical Models and Parametrisations

In addition to observational records, climate science heavily relies on numerical models. A climate model is a multi-component representation of the climate system, typically consisting of the atmosphere, ocean and sea ice, land, biogeochemistry, and possibly other components. A key difference between weather forecasting and climate projections is the timescale.

Weather forecasts are typically made for up to two weeks, with the highest confidence for short-term predictions. Because these simulations are not required to maintain long-term stability, they can be performed at extremely high spatial resolutions. Long-term conservation of energy and mass is also less critical in short-term weather forecasting.

In contrast, climate simulations run for several decades or even centuries, requiring long-term stability and conservation of physical properties. To make these long simulations computationally feasible, climate models use longer time steps and coarser spatial resolutions. This means that only large-scale processes are explicitly simulated. However, we know that synoptic-scale (mesoscale in the ocean) processes contribute significantly to the energy dynamics of the system, influencing and interacting with large-scale patterns. To account for the influence of unresolved small-scale processes, climate models rely on parameterizations—simplified mathematical representations of subgrid processes that are not explicitly resolved. These parameterizations are often based on empirical relationships or statistical approximations. However, they are typically sensitive to parameters, which can lead to significant variability in results and introduce large uncertainties. Over long simulations, these uncertainties accumulate, making climate projections used by organizations like the Intergovernmental Panel on Climate Change (IPCC) less reliable for policymakers.

In recent years, AI has emerged as a promising tool in climate science. Replacing traditional parameterizations with data-driven approaches offers the potential to improve accuracy while maintaining computational efficiency. Even a single high-resolution model run—though short in climate timescales—produces a vast amount of data, sufficient for training neural networks. This project explores how AI can be leveraged to develop improved parameterizations, ultimately contributing to more precise and computationally efficient climate simulations.

C. climSim

The ClimSim dataset [1] was developed to facilitate the application of machine learning (ML) in climate modeling by providing high-resolution atmospheric simulation data for

training AI-based parameterizations. This dataset is derived from the Energy Exascale Earth System Model (E3SM), a next-generation climate model developed by the U.S. Department of Energy (DOE).

A major challenge in climate modeling is the accurate representation of small-scale atmospheric processes, such as convection, cloud formation, and turbulence. These processes influence large-scale weather and climate patterns, but their direct simulation is computationally infeasible for long-term climate projections. The E3SM model, specifically its Multiscale Modeling Framework (MMF), addresses this challenge by embedding high-resolution cloud-resolving models (CRM) within a coarser global model.

II. DATA & METHODS

A. Dataset and EDA

The dataset contains about 10 million samples each with 556 input features and 368 target features. Altogether this makes for 25 input variables, 9 with vertical resolution of 60 atmospheric levels, and 14 target variables, here 6 with a vertical resolution of 60 levels.

B. Data preparation

The data preparation are mainly three steps:

- feature selection (input to target mapping),
- normalisation,
- transforming to tensors.

C. Model Arcitecture

MLP Transformer

D. Training and Evaluation

TO BE FILLED

E. evaluation and R2 score

TO BE FILLED

III. RESULTS

TO BE FILLED

A. MLP results

TO BE FILLED

This will show the results from training on teh larger dataset

Fig. 1. prediction versus true targets

IV. CONCLUSIONS

TO BE FILLED

APPENDIX

show the correlation results here?

REFERENCES

[1] Yu, Sungduk, et al. "Climsim-online: A large multi-scale dataset and framework for hybrid ml-physics climate emulation." *arXiv preprint arXiv:2306.08754* (2023). https://arxiv.org/abs/2306.08754.