Project: **1318**
Project title: **CLINT - climate intelligence**
Principal investigator: **Christopher Kadow**
Report period: **2024-05-01 to 2025-04-30**

Over the last year POLIMI used Levante computational resources mainly to work on the development of a machine learning model to advance tropical cyclogenesis (TCG) detection. The two frameworks developed were a group of Convolutional Neural Networks (CNNs) to estimate the cyclogenesis occurrences given the same inputs of state-of-the-art methods and a feature selection process followed by the training of a 1-layer neural network having the same target.

State-of-the-art numerical indices like the Emanuel and Nolan Genesis Potential Index (ENGPI) combine atmospheric and oceanic variables through fixed functional forms but struggle to capture interannual variability and future trends in TCG. To address these limitations, we developed CNN models trained to estimate the monthly number of TCG events from large-scale environmental inputs used in traditional GPI calculations. The dataset includes reanalysis data (ERA5) and observed TCG events from the IBTrACS archive, covering the period 1980–2021, on a global 2.5° grid. The processing of the ERA5 data to obtain the training dataset was performed on Levante cpu nodes. The model is implemented at both global and tropical sub-basin levels. The hyperparameters tuning of the different CNNs and the training of the final architecture of the models was performed on Levante gpu nodes. Predictions are spatially distributed using a probabilistic method based on observed frequency maps. This approach demonstrates improved skill in capturing the spatial-temporal structure of TCG.

In parallel, we explored the potential of alternative predictors for TCG detection through a dedicated feature selection framework. The candidate pool of features included both atmospheric and ocean variables, as well as climate indices, with the environmental variables derived from the ERA5 reanalysis dataset. To reduce the spatial dimensionality, these gridded fields were clustered prior to the feature selection process. The feature selection algorithm employed is based on the metaheuristic Coral Reef Optimization algorithm, designed to identify the most informative subset of predictors for TCG estimation. Multiple simulations of the selection process were conducted on both cpu and gpu nodes on Levante. Using the selected features, we then trained a final 1-layer neural network to estimate the monthly number of TCG events. As with the CNN pipeline, all data preparation and pre-processing steps for this framework were executed on Levante's cpu nodes. This approach provides a flexible and data-driven method to explore new environmental and climatic signals potentially relevant for improving cyclogenesis detection.

CMCC has used Levante to develop a machine learning (ML) data-driven seasonal forecasting system of heatwaves over the European domain. Within the CLINT project, CMCC has contributed to the development of a novel feature selection framework which identifies the variables, locations and lag times which provide optimal skill for predicting temperature extremes. The description of this methodology is soon to be published in Weather and Climate Extreme, while the application to seasonal forecasting, described below, is in review in Nature Communications Earth and Environment.

The seasonal forecast system combines two steps: (1) the application of the feature selection to individual model grid points over the European domain; (2) the use of selected features in a range of ML models. Both steps are performed on CPUs. The first step is the most computational expensive task, requiring roughly 1000 CPU hours (1 hour per optimization over 1000 grid points). Input data used is the MPI-ESM paleo-simulation “past2k” hosted on the Levante server. The second step is relatively light, as the feature selection outputs a low number of optimal predictors that are used as

input for relatively simple models such as Random Forest. The resulting forecasts and list of predictors require 10-100 of GBs.

The results include a list of key atmospheric and oceanic predictors of European summer heatwaves - a key scientific finding - and purely data-driven forecasts of heatwaves several months ahead, which match the skill of the current state-of-the-art in dynamical seasonal forecasting (Fig. 1).

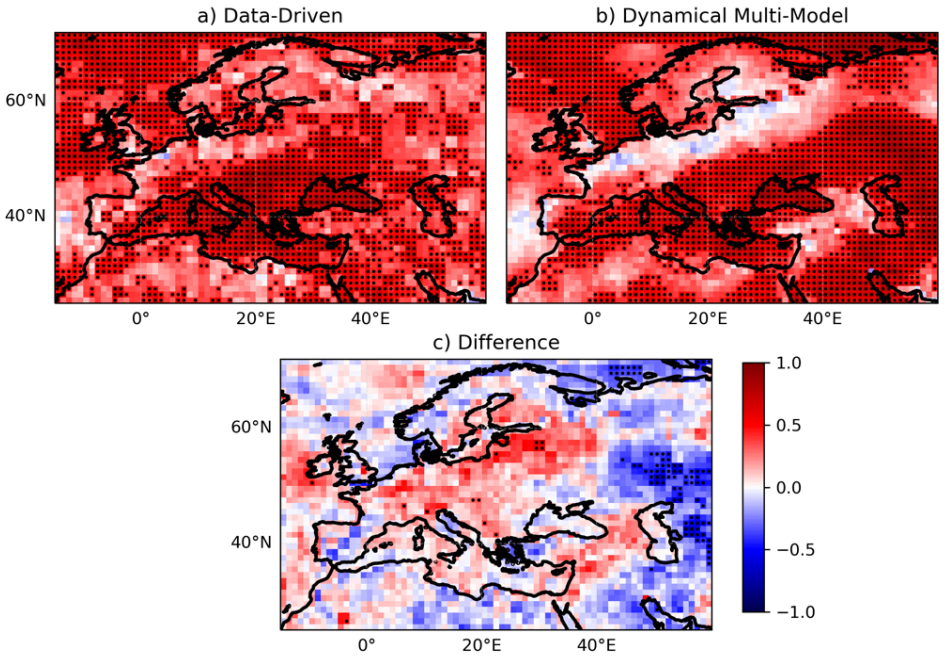

Figure 1: Correlation skill score of forecasts of the number of heatwave days between May and July in the data-driven (a) and Copernicus Climate Change Service multi-model ensemble (b) forecasts over the forecast test period 1993-2016, validated against ERA5. Black stippling represents statistically significant correlation (a & b) or correlation difference (c) at the 95% confidence interval.

As part of the project, a portion of the allocated CPU and GPU node hours was dedicated to organizing a workshop (“Introduction to Deep Learning for Climate Scientists”), which took place at DKRZ from March 19 to March 21. The aim of the workshop was to provide participants with both theoretical foundations and practical, hands-on tutorial sessions to introduce the main concepts of deep learning and demonstrate their applications in the context of climate science. The event brought together over 25 scientists from more than 10 institutions, including members of the Max Planck Society for Meteorology, Helmholtz Zentrum Hereon and Alfred Wegener Institute. The workshop leveraged JupyterHub capabilities to provide an interactive and scalable learning environment, enabling hands-on sessions and seamless access to computing resources and climate data.