Project: **1444**
Project title: **EXPECT**
Principal investigator: **Étienne Plésiat**
Report period: **2024-05-01 to 2025-04-30**

During the reporting period, DKRZ has been the primary user of the allocated computing resources. These resources were predominantly utilized to develop and train machine learning models aimed at infilling missing values in four climate extreme indices (TX90p, TN90p, TN10p, TX10p) derived from the HadEX-CAM dataset—an intermediate product of the HadEX3 dataset (https://www.metoffice.gov.uk/hadobs/hadex3/).

Due to the extensive presence of missing data in HadEX-CAM, conventional methods proved insufficient, necessitating the adoption of advanced data-driven approaches. In response, DKRZ focused on designing a deep learning architecture based on a U-Net framework enhanced with partial convolutions. This method is particularly effective in reconstructing large and irregular gaps within gridded datasets, such as those found in HadEX-CAM.

A substantial portion of the computational effort was devoted to model optimization. This involved an exhaustive search across a broad hyperparameter space—including parameters such as learning rates, convolutional kernel sizes, and network depth—to identify configurations that yielded optimal performance. In total, over 100 distinct configurations were explored. Given the non-deterministic nature of training deep neural networks, multiple iterations were run for each configuration to ensure reproducibility and consistency.

Model training leveraged GPU nodes on Levante. The input data comprised simulations from eight distinct historical models within the CMIP6 archive. Validation was carried out using withheld random subsets from the same data source. Further robustness checks included cross-validation using two major reanalysis datasets, ERA5 and 20CRv3. The latter required downloading and processing the entire dataset at a 3-hour temporal resolution, resulting in data volumes exceeding 20 terabytes.

The results of this work have undergone thorough evaluation and were published in Nature Communications in October (Plésiat et al., 2024) as illustrated by Figure 1.

Following the successful reconstruction of extreme indices in HadEX-CAM, the methodology was extended to a more demanding dataset: the Full Data Monthly Version 2022 from the Global Precipitation Climatology Centre (GPCC, https://opendata.dwd.de/climate_environment/GPCC/html/fulldata-monthly_v2022_doi_download.html). This dataset spans global monthly precipitation observations from 1891 to 2020 at a 0.5° x 0.5° resolution. Several challenges complicated this task, including:

- The high spatial resolution, which significantly increases input data size, model complexity, and computational demands (RAM, VRAM, and processing time).
- Limited training data availability at such resolution.
- The need to account for global boundary conditions.
- A lower proportion of missing values (maximum ~25%) compared to HadEX-CAM (~50%).
- Pronounced spatial heterogeneity in precipitation patterns.

To address these challenges, various model architectures, data subsets, and training strategies were systematically tested. The most promising results were achieved using a U-Net variant with 6 encoding layers and 2046 output channels in the bottleneck layer. This model was trained on HighResMIP data and employed circular padding to better accommodate global continuity constraints.

The resulting model achieved a Root Mean Square Error (RMSE) on the test and ERA5 ensemble mean datasets that was marginally superior to the performance observed with HadEX-CAM.

Further analyses of the GPCC reconstructions are ongoing and will be presented at the upcoming European Geosciences Union (EGU) General Assembly (https://doi.org/10.5194/egusphere-egu25-18816).

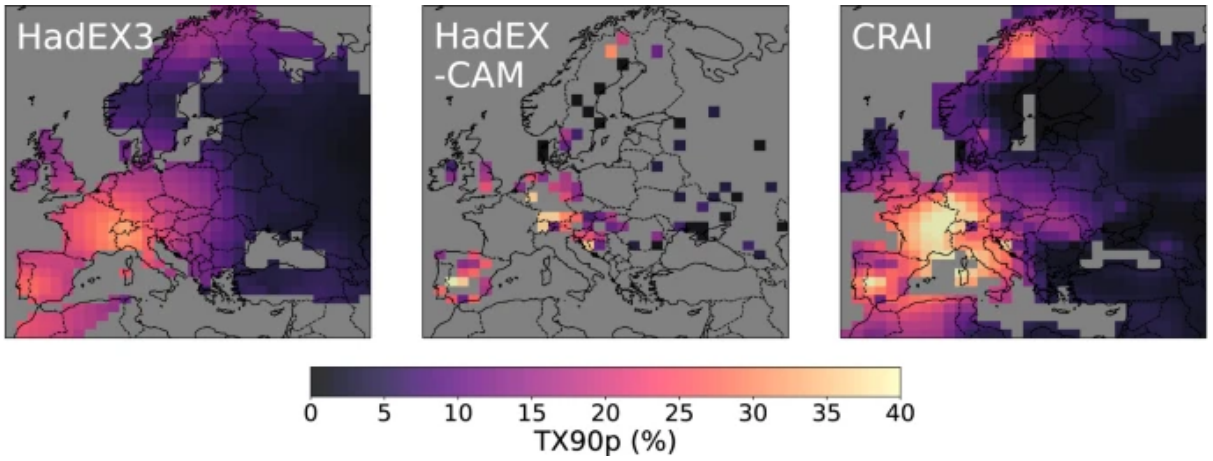

Figure 1: Reconstruction of the TX90p extreme index for a reported heatwave event (September 1911). TX90p is the percentage of days when the daily maximum temperature is > 90th percentile. Left panel: original HadEX3 dataset. Central panel: original HadEX-CAM dataset. Right panel: reconstruction using CRAI.

Over the past year, ULEI has used Levante primarily for data preprocessing, figure review, and causal inference studies in the context of the EXPECT project. Our work aimed at investigating causal relationships between atmospheric patterns and sea surface temperature (SST) variability, and to explore the drivers of heat extremes in Western Europe.

We used ERA5 reanalysis data for preprocessing tasks, including regridding and selecting specific domains over the North Atlantic and Western Europe. Figures from Carvalho-Oliveira et al. 2024 were reviewed, where causal links between the East Atlantic pattern and extratropical North Atlantic SSTs were analysed using the Tigramite PCMCI algorithm applied to ERA20C reanalysis data. Additionally, a Bachelor student working with us has been testing causal inference methods to investigate the role of soil moisture anomalies in influencing heat extremes over Western Europe.

References

Carvalho-Oliveira, J., Di Capua, G., Borchert, L.F., Donner, R.V. and Baehr, J., 2024. Causal relationships and predictability of the summer East Atlantic teleconnection. Weather and Climate Dynamics, 5(4), pp.1561-1578.

Plésiat, É., Dunn, R.J.H., Donat, M.G. et al. Artificial intelligence reveals past climate extremes by reconstructing historical records. Nat Commun 15, 9191 (2024). https://doi.org/10.1038/s41467-024-53464-2