Project: **1318**
Project title: **CLINT - climate intelligence**
Principal investigator: **Christopher Kadow**
Report period: **2023-05-01 to 2024-04-30**

DKRZ has been the main user of computing resources during the 2023-05-01 to 2024-04-30 period. These resources have been mostly employed to train infilling models to reconstruct 4 extreme indices (TX90p, TN90p, TN10p, TX10p) of the HadEX-CAM dataset that is an intermediate product of the HadEX3 dataset (https://www.metoffice.gov.uk/hadobs/hadex3/). The amount of missing values in the HadEX-CAM dataset being very large, it has been required to adopt novel machine learning approaches. In particular, DKRZ has been working on the development of a U-Net that employs partial convolutions. This technique allows for the reconstruction of large and irregular regions of missing values such as in the HadEX-CAM dataset.

The training of the infilling models necessitates performing a hyperparameter tuning that aims at determining the optimised configuration of hyperparameters (e.g., learning rate, number of layers, size of convolution filters, etc.) for the corresponding task. In total, more than 100 configurations have been explored to determine the optimal configuration. Additionally, as the training process is not purely deterministic, it is generally recommended to train multiple models for a single configuration of hyperparameters.

The training of these models has been performed using the GPU nodes at Levante. The input dataset has been created using 8 historicals models from the CMIP6 archive. The evaluation has been carried out on a test set made from unseen random samples of these 8 historical models. A cross-validation using reanalysis datasets (ERA5 and 20crv3) have also been performed. To compute the climate indices for the 20crv3 dataset, it has been required to download the full dataset at the 3-hourly resolution that took more than 20TB of storage.

An extensive analysis of the reconstructed dataset has been performed and reported in a scientific article that is currently under review in Nature Communications.

POLIMI ranks among the primary CLINT partners in terms of computing resources usage in the course of 2023-05-01 to 2024-04-30. Within the project, POLIMI has been developing machine learning algorithms to analyse tropical cyclones (TCs): one for the downscaling and bias adjustment of ERA5 extreme rainfall caused by tropical cyclones, and one for the estimation of tropical cyclone intensity from satellite images. The algorithm served as the basis for which to develop a similar algorithm to be applied to ECMWF forecast data (HRES) instead, which in turn will be used within CLINT's WP7 in the Zambesi local study.

For the adjustment of ERA5 TC-induced rainfall, we implemented a modification of the popular UNet architecture called Residual Attention UNet (RA-UNet). We also developed a novel loss function that combines a pixel-wise component (MSE/MAE) with a component that measures if peaks of rainfall are correctly localized within the output image (Fractions Skill Score). The majority of experiments carried out were aimed at optimizing the various hyperparameters of the model, including the weights of the loss function, the number of blocks in RA-UNet, and the number of filters in each convolutional block. A second large set of experiments was carried out to understand if adding more input variables (other than ERA5's precipitation) would increase the performance of the model.

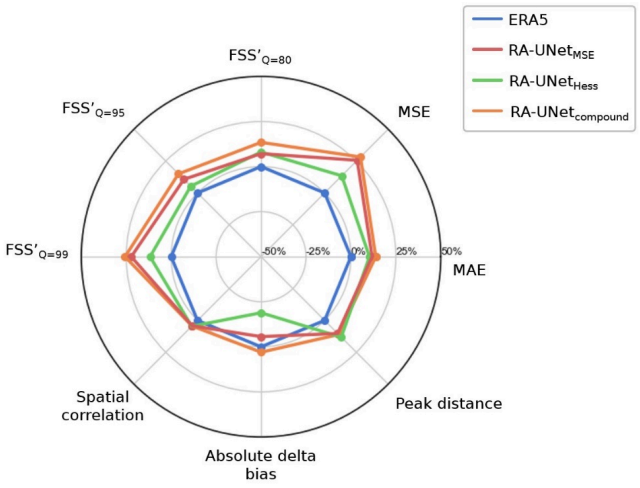

Figure 1 shows that the proposed method was not only able to reduce all biases (pixel-wise and of spatial distribution) of ERA5, but also performed better than some other similar algorithms found in the literature.

Figure 1 (rainfall adjustment) Radar plot showing how the model developed reduces all metrics of error of ERA5 and outperforms similar methods from the literature.

JLU is leading the tasks for compound events and concurrent extremes as well as the development of AI-enhanced climate services for the food sector. The main objective of the **concurrent extremes** task is to exploit the global interconnectivities of large-scale heatwaves and droughts. New indices for droughts and heatwaves were derived to better describe the co-variability of these phenomena. For that purpose, a nonparametric version of the Standardized Precipitation and Evapotranspiration Index (NP-SPEI) was derived based on local likelihood procedures. For heatwaves, a Bivariate Heat Magnitude Day (BVHMD) index including minimum and maximum temperature was developed. The covariability of the two indices was shown to be enhanced in terms of Pearson correlation (not shown) in comparison to the classical indices.
After the derivation of those indices, it was investigated whether those indices can be used to derive stress indices modelling climate-related impact in the agricultural sector with the aim of improving the Compound Stress Index (CSI). We focussed on deriving the impacts of climate on grain maize production in Europe, which is the most important summer crop in Europe. Using D-vine-copula based quantile regression as well as the NP-SPEI and BVHMD we were able to verify that the impacts of climate on grain maize were substantially higher than with the CSI demonstrating the added value of our new indices. Finally, the new nonparametric version of the SPEI was shown to be superior in terms of performance to the classical SPEI yielding smaller Cramer-von-mises statistics (not shown). Furthermore, we were able to verify that the NP-SPEI can be used to construct droughts indices based on reference periods (not shown), for which the original SPEI had substantial problems as it was not able to extrapolate values that were not observed in the reference period.
Another objective is to build seasonal forecasts of concurrent extremes. The methods were applied to the Lake Como region as a case study of the CLINT project. Building on the nonparametric climate indices described above a new index for concurrent extremes named concurrent extreme event index (CEEI) was developed. The index condenses the information of heatwave and drought indices and the drivers of those through a novel non-linear kernel regularized generalized canonical correlation analysis (KRGCCA) resulting in an index describing interconnectivities of large-scale droughts and heatwaves. After the development of CEEI, a Bayesian Neural Network based on the QUINN framework is used to estimate the conditional distribution of the CEEI based on meteorological variables with lags of one to six months. The lead time of one month then allows to derive one-month ahead forecasts; other lead times can be chosen depending on the applications.

Compound Events

The analysis of the compound events has been conducted for wet and warm later winters followed by dry and hot springs focusing on their impact on agriculture. Using a KRGCCA, our results indicate that wet and warm conditions in February and March as well as the following dry and warm April mainly contribute to high and extreme agricultural impacts, with the spring drought being the most important variable for the latter.
Based on those insights, compound events capturing the necessary meteorological conditions as well as the impacts can be defined, allowing events to be labelled and a classification problem to trained. For this purpose, random forests based on the q*-classifier were employed to take the imbalances into account showing desirable performance (G-Mean on test set: 0.89). Using those forests, approximate decision trees are obtained giving thresholds of around 0.1 for the winter variables, while it was -0.76 for the NP-SPEI in April. Thus, the bounds are much lower than one standard deviation suggesting the considered events are not extreme at all, while being often associated with high impacts. The latter demonstrates that while the individual climate events may not be individually extreme and likely unharmful, but they can lead to substantial agricultural damage when combined.

AI-enhanced climate services for the food sector

The ECroPS crop growth model has been extensively utilized for the European domain and the simulation of grain maize growth for yield output, for 31 years (1993 to 2023) with a spatial resolution of 0.25 degrees. The ECroPS model is fed with soil gridded input data, crop-specific and agromanagement data and it is forced with 6 daily ERA5 weather variables. ECroPS also requires three types of water balance evaporation variables calculated with the modified Penman approach and the Penman-Monteith approach.
The constructed database consists of the sample points deriving from each independent run of the ECroPS model for each grid cell and each year. For the AI surrogate model, we have performed benchmarking model runs with ensembles of Random Forests and ensembles of XGBoost runs.
Continuing towards the building of the actual emulator, since the feature space is very large, we tested feature selection with (I) Recursive Feature Elimination with Cross-Validation algorithm forced with Extreme Learning Machines (ELM), (II) an optimization pipeline using the minimize functions from scikit-optimize forced again with ELMs and (III) the Coral Reef Optimization with Substrate Layers (CRO-SL) algorithm, further reducing the feature space using Principal Components Analysis. Finally, the initial tests for the actual emulator include shallow feedforward neural networks, AI pipelines that rely on recurrent neural network architectures consisting of LSTM components and we have initiated efforts towards the more robust probabilistic approaches of BNN.