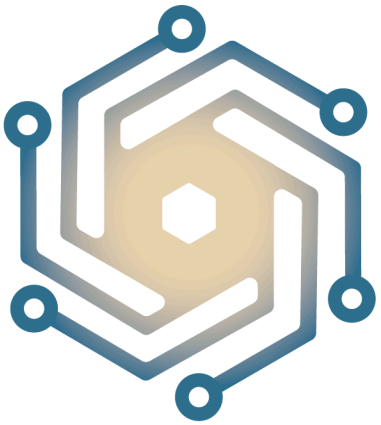


CLINT

CLIMATE INTELLIGENCE

D3.3
AI-ENHANCED EXTREME EVENTS DETECTION

October, 2024

This project is part of the H2020 Programme supported by the European Union, having received funding from it under Grant Agreement No 101003876

Programme Call:	Building a low-carbon, climate resilient future: climate action in support of the Paris Agreement (H2020-LC-CLA-2018-2019-2020)
Grant agreement ID:	101003876
Project Title:	CLINT
Partners:	POLIMI (Project Coordinator), CMCC, HZG, CSIC, SMHI, HKV, E3M, TCDF, DKRZ, IHE, ECMWF, UAH, JLU, OGC, UCM
Work-Package:	WP3
Deliverable #:	D3.3
Deliverable Type:	Document
Contractual Date of Delivery:	31 October 2024
Actual Date of Delivery:	31 October 2024
Title of Document:	AI-enhanced Extreme Events detection
Responsible partner:	CMCC
Author(s):	Enrico Scoccimarro, Antonello Squintu, Ronan McAdam, Michael Maier-Gerber, Matteo Giuliani, Niklas Luther, Leone Cavicchia, Paolo Lanteri, Guido Ascenso, Filippo Dainelli, Linus Magnusson, Felicitas Hansen, Jorge Pérez Aracil, César Peláez Rodríguez, Martina Merlo, Paolo Bonetti, Matteo Sangiorgio, Marcello Restelli, Andrea Castelletti, Yiheng Du, Elena Xoplaki, Odysseas Vlachopoulos, Katharina Klehmet, Wei Yang, Lucia De Stefano, Harilaos Loukos.
Content of this report:	Description of the final set of AI-enhanced tools for Extreme Events detection: the different AI-enhanced tools to be used to detect the different types of Extreme Events considered in WP3 are described in the present document.
Availability:	This report is public.

Document revisions		
Author	*Revision content*	*Date*
Antonello Squintu and Ronan McAdam	D3.3_v01 – First draft of chapter 2, structure of the document	07/09/2024
All authors	D3.3_v02 - Final draft of all chapters, formatting	30/09/2024
Internal Reviewers, CMCC and Chapter Authors	D3.3_v03 – Internally reviewed document	25/10/2024
Guido Ascenso	Final quality check	29/10/2024

TABLE OF CONTENTS

TABLE OF CONTENTS **5**
LIST OF FIGURES **8**
LIST OF ACRONYMS **13**
EXECUTIVE SUMMARY **16**
1 INTRODUCTION **19**
2 TROPICAL CYCLONES: INDICES **20**
2.1 Introduction 20
2.2 AI-enhanced indices for TC detection 22
2.2.1 Optimization of the ENGPI: the oGPI 22
2.2.2 AI-enhanced Genesis Potential Index (AI-GPI) 22
2.2.2.1 Data 22
2.2.2.2 Methods 24
2.2.2.3 Results 26
2.2.2.4 Future Developments 36
3 DETECTION OF TROPICAL CYCLONES **37**
3.1 Tropical Cyclone Activity 37
3.1.1 Overview 37
3.1.2 Datasets and Forecasts 37
3.1.2.1 Datasets 37
3.1.3 Skill and performance of existing forecasts 39
3.1.4 Developed algorithms 40
3.1.5 Results: AI-Enhanced forecasts and relevant drivers 43
3.1.6 Summary and Outlook 46
3.2 Extratropical Transition 47
3.2.1 Overview 47
3.2.2 Datasets, candidate drivers and target variable 49
3.2.3 Skill and performance of existing forecasts 51
3.2.4 Developed Algorithms 51
3.2.5 Results: AI-Enhanced forecasts and relevant drivers 52
3.2.6 Summary and Outlook 53

4 HEATWAVES AND WARM NIGHTS **54**
4.1 Indices and Datasets 54
4.2 ML-based identification of HW indicators for agriculture 56
4.2.1 Impact Model 56
4.2.2 Patient Rule Induction Method 57
4.2.3 Results 58
4.3 HW precursors: ML-defined weather regimes 61
4.4 Optimisation-Based Feature Selection Framework: Driver Detection & Forecasting 64
4.4.1 Identification of Regional HW Cluster Predictors 65
4.4.2 Hybrid Seasonal Forecasting 68
4.4.3 Data-Driven Seasonal Forecasting 70
4.5 Night-time extremes 75
4.6 Summary and Future Steps 76
5 DETECTION OF EXTREME DROUGHTS **77**
5.1 Introduction 77
5.2 AI-enhanced impact-based drought detection via multi-task learning 78
5.2.1 Case study and data 78
5.2.1.1 FAPAR anomaly 78
5.2.1.2 Hydroclimatic predictors 80
5.2.2 Multi-task learning drought detection 81
5.2.3 Numerical Results 82
5.2.3.1 Local models 82
5.2.3.2 Global models 88
5.3 Identification of critical drought features 91
5.3.1 Impact model 91
5.3.2 AI method 94
5.3.2.1 Synthetic generation of drought scenarios 94
5.3.2.2 Scenario discovery 95
5.3.2.3 Numerical results 96
5.4 Conclusions 98
6 COMPOUND EVENTS AND CONCURRENT EXTREMES **99**

6.1 Introduction 99
6.2 Data and methods 100
6.3 Compound Events 102
6.3.1 Relatively wet and warm late winters followed by dry and warm springs 102
6.3.1.1 Identification of large-scale patterns 102
6.3.1.2 Construction of objective thresholds 108
6.3.2 Dry winters followed by hot summers 111
6.3.2.1 The Bivariate Heat Magnitude Day 111
6.3.2.2 Nonlinear compound stress indices 113
6.3.2.3 Impacts on the energy sector 120
6.3.3 Wet warm springs 123
6.4 Concurrent Extremes 128
6.4.1 Nonparametric climate indices with an application to the SPEI 128
6.4.2 Detection of dependencies using AI-enhanced point process approaches 130
6.5 Conclusion 133
7 CONCLUSIONS 135
REFERENCES 137
APPENDIX A4 153
A4.1 Data-Driven Forecast Skill (2004-2022) 153
A4.2 Night-time heatwave clusters 153
A4.3 Seasonal Forecast skill of day and night-time extremes 154
APPENDIX A6 154
A6.1 Relatively wet and warm late winters followed by dry and warm springs 154
A6.2 Dry winters followed by hot summers 155
A6.3 Wet and warm springs 157
A6.4 Nonparametric SPEI 159

LIST OF FIGURES

Figure 1.1: Schematic representation of D3.3 and its interconnection with previously completed D and MS of WP2 and WP3.

Figure 2.1 Top: spatial distribution of TC genesis according to observed data (IBTrACS, panel A) and to the ENGPI (panel B). Bottom: Interannual variability curves for the ENGPI.

Figure 2.2 Top: spatial distribution of TC genesis according to observed data (IBTrACS, panel A) and to the oGPI (panel B). Bottom: Interannual variability curves for the oGPI.

Figure 2.3: Sub-basin domains extension.

Figure 2.4: Interannual variability curves for the Tropics domain comparing the observed cyclogenesis to the ENGPI, the oGPI, and the AI-GPI.

Figure 2.5: Spatial distribution of TC genesis in the Tropics according to (a) observed data, (b) ENGPI, (c) oGPI, and (d) AI-GPI. The coloured map indicates the average number of cyclogeneses per year per grid point.

Figure 2.6: Interannual variability curves for the North Atlantic domain comparing the observed cyclogenesis to the ENGPI, the oGPI, and the AI-GPI.

Figure 2.7: Spatial distribution of TC genesis in the North Atlantic according to (a) observed data, (b) ENGPI, (c) oGPI, and (d) AI-GPI. The coloured map indicates the average number of cyclogeneses per year per grid point.

Figure 2.8: Interannual variability curves for the Northeast Pacific domain comparing the observed cyclogenesis to the ENGPI, the oGPI, and the AI-GPI.

Figure 2.9: Spatial distribution of TC genesis in the Northeast Pacific according to (a) observed data, (b) ENGPI, (c) oGPI, and (d) AI-GPI. The coloured map indicates the average number of cyclogeneses per year per grid point.

Figure 2.10 Interannual variability curves for the Northwest Pacific domain comparing the observed cyclogenesis to the ENGPI, the oGPI, and the AI-GPI.

Figure 2.11: Spatial distribution of TC genesis in the Northwest Pacific according to (a) observed data, (b) ENGPI, (c) oGPI, and (d) AI-GPI. The coloured map indicates the average number of cyclogeneses per year per grid point.

Figure 2.12: Interannual variability curves for the South Pacific domain comparing the observed cyclogenesis to the ENGPI, the oGPI, and the AI-GPI.

Figure 2.13: Spatial distribution of TC genesis in the South Pacific according to (a) observed data, (b) ENGPI, (c) oGPI, and (d) AI-GPI. The coloured map indicates the average number of cyclogeneses per year per grid point.

Figure 2.14: Interannual variability curves for the North Indian domain comparing the observed cyclogenesis to the ENGPI, the oGPI, and the AI-GPI.

Figure 2.15: Spatial distribution of TC genesis in the North Indian according to (a) observed data, (b) ENGPI, (c) oGPI, and (d) AI-GPI. The coloured map indicates the average number of cyclogeneses per year per grid point.

Figure 2.16: Interannual variability curves for the South Indian domain comparing the observed cyclogenesis to the ENGPI, the oGPI, and the AI-GPI.

Figure 2.17: Spatial distribution of TC genesis in the South Indian according to (a) observed data, (b) ENGPI, (c) oGPI, and (d) AI-GPI. The coloured map indicates the average number of cyclogeneses per year per grid point.

Figure 3.1: (a) Mean and (b) variance of relative frequency of TC occurrence (%) calculated for 1980-2015, which is used as training period. Note that interval boundaries are not equidistant. The blue box encloses the area in the Southern Indian Ocean, for which the ML models are trained.

Figure 3.2: (a) Brier skill score (BSS) (in %) of TC activity probability with respect to the climatological model as function of lead time. (b) BS decomposition into uncertainty (grey), miscalibration (blue), and discrimination (red) for the two benchmark models. Resulting BSs are displayed by the black asterisks.

Figure 3.3: Example of U-Net, a state-of-the-art convolutional-based architecture considered for the task. Each feature is a channel of the input image and the output represents the spatial probability of occurrence of a TC. Credits for the image: (Serifi et al. 2021).

Figure 3.4: Schematic illustrating the purely data-driven (orange arrow) and hybrid (combination of red arrows) modelling approach for the example of predicting TC activity at 120h lead time.

Figure 3.5: Same as in Figure 3.2a but including the baseline ML models.

Figure 3.6: Same as in Figure 3.2a but including the CNN-based ML models (solid lines) and the hybrid model approach (dashed line).

Figure 3.7: Same as in Figure 3.2b but for the original (left) and hybrid (right) versions of the best-performing U-Net.

Figure 3.8: IBTrACS positions cyclones in the extratropical stage for April 2016-December 2022.

Figure 3.9: Brier score (BS) decomposition into uncertainty (grey), miscalibration (blue), and discrimination (red) for the two benchmark models. Resulting BSs are denoted by the vertical black lines.

Figure 3.10: (a) ROC curves for all models with AUC scores in the legend. (b) As in Figure 3.9, but including the results for the ML models sorted by BS.

Figure 3.11: Results of the sequential predictor selection applied to the logistic regression. Mean (line) and standard deviation (shading) of the negative log loss, the AIC, and the BIC as a function of the number of features. The dotted vertical line marks the optimal number of features identified for the corresponding score.

Figure 4.1: Scatter plot of the most relevant heatwave/drought indices and crop yield.

Figure 4.2: PRIM results in terms of scenario boxes navigating the trade-off between coverage (x-axis) and density (y-axis) of crop failure with respect to heatwaves/drought indices. Circled points are the scenario boxes analyzed in detail.

Figure 4.3: Composite maps of WR10, WR09, WR05 and WR12, where red isolines indicate higher-than-average pressure anomalies and blue isolines indicate lower-than-average pressure anomalies.

Figure 4.4: Histogram of extreme events (May to August, 1940-2022), defined using standard heatwave definition (left) and the EHF (right), with given WRs and corresponding tests of statistical significance (sig.level = 0.05) .

Figure 4.5: Annual number of heatwave days (May to August, 1940-2022) based on NDQ90, HW_occ and EHF severity > 1 (severe and extreme heatwaves) over the period 1940-2022 during the summer season (May to August) selected for a grid of ERA5 representing Stockholm. The bold lines represent the 10-year moving average.

Figure 4.6: HW clusters over the European domain, coloured by their average intensity (contours correspond to 0.3 ^{o}C intervals). Clu1 corresponds to no HW. Variability (in parentheses) explained by each cluster is also a measure of dataset imbalance.

Figure 4.7: Example feature selection for Clu6 (British Isles). Maps of each cluster can be found in the appendix. Letters correspond to the following climate indices and dummy variables: a - ENSO, b - NAO, c - IOD, d - atmospheric CO2 concentration, e - day of year.

Figure 4.8: Recreation of HW cluster 6 indices from 2015-2022 (test period) from optimal features input into different models (Logistic Regression and Gradient Boosting Classifier). Values in the legend correspond to F1-score (left) and correlation of total summer days each year (right).

Figure 4.9: Time series of t2m temperature anomaly in Europe cluster 2 in ERA5 (black), ECMWF-SEAS5 ensemble members (red) and ensemble median (red dashed). Units in °C. In the hybrid framework, predictor data from May onwards (black vertical line) is taken from dynamical system (e.g. ECMWF-SEAS5) forecasts.

Figure 4.10: F1-score for HW occurrence over training period (1951-2004) cross-validation and test period (2004-2022).

Figure 4.11: Correlation skill score of NDQ90 (May-July) in the period 2004-2022 between detection/forecast systems and ERA5. Detect uses only ERA5 predictors, while Hybrid replaces predictor data after May 1st with ECMWF-SEAS5 forecasts. Black stippling indicates statistical significance. For Dynamic and Hybrid, forecasts are initialised in May.

Figure 4.12: Sub-sample of past2k target data (number of May-June-July HW days) over the period 1750-1850, defined relative to diverse climatology periods.

Figure 4.13: Example optimisation and feature selection for grid cell (East Mediterranean Sea).

Figure 4.14: NRMSE of optimised solutions across the European domain for recreation of past2k HW indicators. Left: training period cross-validation 0-1600. Right: test period 1600-1850.

Figure 4.15: Identification of selected predictors for the whole European domain. Percentage of grid points which use cluster and lag in optimal solution. Weeks from initialisation (May 1st).

Figure 4.16: Percentage of grid points which select features based on lag time.

Figure 4.17: Anomaly correlation of skill scores over Europe for the period 1993-2016 for the dynamical system ECMWF-SEAS and data-driven approach implemented in two ML models (LR - Logistic Regression; RF - Random Forest). Black stippling indicates statistical significance.

Figure 5.1: mean MAE across all sub-basins, obtained from utilising centroid (a) and average linkages (b), respectively. Both linkages are implemented within hierarchical clustering and hierarchical NonLinCTFA algorithms.

Figure 5.2: Results from local models. The hierarchical clusterings with the lowest mean MAE and the NonLinCTFA clusterings, considering centroid and average linkages, with their respective MAEs on the right. The clusterings visualized on the left are the ones generate in the first split of cross-validation, while the MAEs consider the entire reconstructed FAPAR Anomaly time series.

Figure 5.3: Impact of CMI filter and nested forward wrapper as feature selection methods on model performance metrics with increasing numbers of selected features. The linear regression models are trained on data aggregated by hierarchical clustering with average linkage and a threshold of 0.4.

Figure 5.4: Estimated number of optimal features in each region, obtained by averaging the number of optimal features for each cluster from the 17 clusterings (one for each cross-validation split).

Figure 5.5: Regions where SPEI-1, SPI-3, and SMA-1 are selected by the nested forward wrapper and the CMI filter. Each map indicates where and how many times, during cross-validation, the feature was selected.

Figure 5.6: Comparison of global models considering individual sub-basins as baseline. The maps on the left show the granularity of the input features, while the maps on the right the corresponding model's accuracy.

Figure 5.7: Comparison of local and global models on a specif area in northern Italy. The left map shows the clusters obtained in the local case, with the associated models' accuracy visualized in the middle panel. The right panel shows instead the accuracy of the global model trained on clustered features.

Figure 5.8: Accuracy of global models combined with CMI and wrapper feature selection methods in terms of average MAE (left panel) and correlation (right panel).

Figure 5.9: Most selected features for the global model, considering wrapper feature selection with 6 selected features in 17-fold cross-validation.

Figure 5.10: SSI drought index computed for inflow into Lake Como in the historical period 1946-2021. The hydrological droughts are highlighted in red.

Figure 5.11: 3-dimensional space (persistence, intensity, frequency) with three hundred LHS samples (orange circles) and points representative of the historical period (black square) (a). Examples of SSI time series for some extreme cases (b-e).

Figure 5.12. Decision tree classifier structure.

Figure 6.1: Retained spatial pattern of soft winter wheat anomalies (top panel) and corresponding time series (bottom panel).

Figure 6.2: Top panel: variable importance obtained for the 70th conditional quantile of the crop yield anomalies using the vine copula based quantile regression model. Bottom panel: preimages spatial patterns for the NPSPEI -1 and SATS-1 corresponding to the first component of the crop yield anomalies displayed in Figure 6.1.

Figure 6.3: Marginal effects of warm May (a) and spring drought (b) on yield anomalies. Interaction effect of NPSPEI-1 and the SATS-1 in May is shown in (c).

Figure 6.4: Obtained agroclimatic regions used for grouping winter wheat crop yield anomalies and the local climate variables from February to May described in the text.

Figure 6.5: Global surrogate model for the imbalanced random forest.

Figure 6.6: Schematic Visualisation of multivariate thresholds taken from Salvadori et al. 2016.

Figure 6.7: Explained Variance by regressing the nonparametric indices on grain maize anomalies using d-vine copula-based quantile regression. Blue lines correspond to the explained variance (R^2) of this model, while the superimposed orange lines describe the increase of the latter in comparison to the CSI. The x-axis displays the evaluated NUTS3 regions.

Figure 6.8: Model diagnostics from the QUINN model used for predicting grain maize anomalies. Panel (a) and (b) show Q-Q-plots of the model and (c) and (d) compare the QUINN prediction with the CSI and non-linear extension (section 6.3.2.2.1).

Figure 6.9: Variable importance for the grain maize based QUINN model utilizing the 90th Percentile. SSM denoted soil moisture in layer 1, 2, 3 and 4.

Figure 6.10: Large-scale component extracted for the analysis of outages. Panel (a) shows the component of the outages obtained from glmPCA and KRGCCA. Panel (b) and (c) correspond to preimages of the first component of the KRGCCA for the (b) soil moisture layer 4 and (c) total precipitation. The Figure on the down right shows ALEplots obtained for total precipitation and soil moisture in the fourth layer.

Figure 6.10: Large-scale component extracted for the analysis of outages. Panel (a) shows the component of the outages obtained from glmPCA and KRGCCA. Panel (b) and (c) correspond to preimages of the first component of the KRGCCA for the (b) soil moisture layer 4 and (c) total precipitation. The Figure on the down right shows ALE plots obtained for total precipitation and soil moisture in the fourth layer.

Figure 6.11: The first two extracted components from the KRGCCA approach taking the E-Hype model as output.

Figure 6.12: Variable importance for the river discharges using the QUINN model based on the 90th percentile. Whiskers indicate 95 % credible intervals

Figure 6.13: Constructed preimages for the second component of the KRGCCA analysis reflecting Total Precipitation for (a) MAM and (b) DJF. (c) and (d) correspond to preimages of maximum temperature taken in MAM and JFM.

Figure 6.14: Second order interaction of MAM Total Precipitation and JFM maximum temperature.

Figure 6.15: Q-Q-plots of the considered SPEI version for eleven representative regions in the world.

Figure 6.16: Identification of dependencies through the AI-based J-function interpreter.

Figure 6.17: All classified J-functions of figure 15 plotted with respect to their identified class. The dotted lines correspond to pointwise 10th and 90th percentiles, indicating that most of the identified functions follow their theoretical trajectories.

Figure A4.1 Correlation skill of Data-Driven HW Seasonal Forecasts over 2004-2022, for comparison with equivalents from the dynamical (ECMWF-SEA5) and hybrid systems (Fig 4.11). LR - Logistic Regression; RF - Random Forest.

Figure A4.2 Night-time HW clusters over the European domain, coloured by their average intensity (contours correspond to 0.3°C intervals).

Figure A4.3: Differences between the correlation maps of the multi-model seasonal predictions for the ATn and the corresponding correlation maps but for (a) Tmin, (b) Tn, and (c) Tmax for the 1993–2016 period in the 15MJJA season. These correlation maps are computed with ERA5 as an observational reference. Hatched areas indicate where the four individual prediction systems agree in the positive (green lines) or negative (purple lines) correlation differences. The seasonal forecasts are issued on the 1st of May.

Figure A6.1: Connection matrix used for the KRGCCA employed for the analysis of the relatively wet and warm late winters followed by dry springs.

Figure A6.2: Density histogram of correlation of residuals from the NUTS3 regions utilized for the SUR model.

Figure A6.3: Histogram for the number of observations in the NUTS3 regions.

Figure A6.4: Interaction Effect of Minimum Temperature with (a) Total Precipitation and (b) Soil Moisture Layer 4.

Figure A6.5: Q-Q-plot of the estimated QUINN model for the river runoffs for the first component displayed in (a) and (b) and the second component in (c) and (d).

Figure A6.6 Displayed in blue is the second component of discharges (Figure 6.13) and overlaid in gray shadows the reported outages from ENTSO-E for (a) 2020 and (b) 2016. R corresponds to the biserial correlation and p denotes the p-value. Both correlations have been calculated for a lag of three days for which the maximum lagged correlation is observed.Figure A6.7: Number of non-extrapolatable points of SPEI using the log-logistic distribution.

Figure A6.8: Same as Figure A6.6, but for the NPSPEI.

Figure A6.9: Difference of Anderson Darling statistics for SPEI and NPSPEI. Positive values indicate that the NPSPEI produces a smaller statistics and hence a better fit with respect to the standard normal distribution.

LIST OF TABLES

Table 2.1: Years of the dataset being part of the training set and the set for the five different folds of cross-validation.

Table 2.2: Summary of interannual and spatial correlations for the ENGPI, oGPI, and AI-GPI across different domains (Tropics and sub-basins).

Table 3.1: Qualitative summary of results of a number of the feature engineering and selection tests indicating to what extent a change has been beneficial (green plus symbols), neutral (grey circle) or detrimental (red minus symbols).

Table 3.2: Total and train-test-split number statistics of all TCs and TCs reaching extratropical stage.

Table 4.1: Quality of recreation of HW cluster occurrence (cross-validation F1-score) with optimal solution (0 lowest, 1 highest).

Table 5.1: Performance of different global models that use various sets of input features.

Table 5.2. Confusion matrix assessing the accuracy of the decision tree classifier.

Table 6.1: Scores obtained from the test set (2010-2020) using the imbalanced RF approach.

Table 6.2: Likelihood of grain maize crop yield failures. The Reference ratio is defined as the median likelihood (column 1) of observing the desired event divided through the likelihood of observed a failure in the REF scenario.

Table A6.1: Statistics for the EOF analysis performed for the Alpine region variables.

LIST OF ACRONYMS

ADNET: Adaptive Elastic Net
AI: Artificial Intelligence
AI-GPI: Artificial Intelligence enhanced Genesis Potential Index
AIC: Akaike Information Criterion
ALE: Accumulated Local Effect
AMO: Atlantic Multidecadal Oscillation
ANN: Artificial Neural Network
ATn: Apparent Temperature at night
ATS: Active Temperature Sum
AUC: Area Under the Curve
BIC: Bayesian Information Criterion
BS: Brier Score
BSS: Brier Skill Score
BVHMD: Bivariate Heat Magnitude Day
CART: Classification And Regression Tree
CCA: Canonical Correlation Analysis
CCEW: Convectively Coupled Equatorial Wave
CDF: Conditional Distribution Function
CEU: Central Europe
CI: Coupling Index
CMCC: Centro Euro-Mediterraneo sui Cambiamenti Climatici
CMI: Conditional Mutual Information
CMIP6: Coupled Model Intercomparison Project Phase 6
CNN: Convolutional Neural Network
CSI: Compound Stress Index
CSIS: Conditional Independence Screening
CVM: Cramer von Mises
DKRZ: Deutsche Klimarechenzentrum
DSC: Discrimination
DTW: Dynamic Time Warping
E-HYPE: European Hydrological Predictions for the Environment
ECMWF: European Centre for Medium-Range Weather Forecasts
ECV: Essential Climate Variables
EE: Extreme Events
EHF: Excess Heat Factor

ELM: Extreme Learning Machine
EN-GPI: Emanuel-Nolan Genesis Potential Index
ENSO: El Nino Southern Oscillation
ENTSO-E: European Network of Transmission System Operators for Electricity
EOF: Empirical Orthogonal Function
ERA5: ECMWF Reanalysis v5
ET: Extratropical Transition
FAPAN: Fraction of Absorbed Photosynthetically Active radiation anomaly
FAPAR: Fraction of Absorbed Photosynthetically Active Radiation
FFNN: Feed-forward Neural Network
FIND: Frequency INtensity and Duration
FRIDA: FRamework for Index-based Drought Analysis
FS: Feature Selection
GA: Grant Agreement
GDO: Global Drought Observatory
GHG: Greenhouse Gas
GLM: Generalized Linear Model
GLMPCA: Generalized Linear Model Principal Component Analysis
GNN: Graph Neural Network
GPI: Genesis Potential Index
HERA: High-resolution pan-European hydrological analysis
HMD: Heat Magnitude Day
HYPE: HYdrological Predictions for the Environment
HydroGFD: Hydrological Global Forcing Data
IBTrACS: International Best Track Archive for Climate Stewardship
IFS: Integrated Forecasting System
IPCC: International Panel on Climate Change
IVS: Input Variable Selection
JFM: January, February, March
KDE: Kernel Density Estimator
KDESPEI: Kernel Density Estimated Standardized Precipitation and Evapotranspiration Index
KRGCCA: Kernel Regularized Generalized Canonical Correlation Analysis
LDAS: Land Data Assimilation System
LHS: Latin Hypercube Sampling
LSTM: Long Short-Term Memory
MAE: Mean Absolute Error
MCB: Miscalibration
MCE: Meteorological Compound Event
MCMC: Monte Carlo Markov Chain
MED: South Europe/Mediterranean
MJO: Madden-Julian Oscillation
ML: Machine Learning
MOEA: Multi-Objective Evolutionary Algorithm
MPI: Maximum Potential Intensity
MS: Milestone
MSLP: Mean Sea Level Pressure
MWMOTE: Majority Weighted Minority Oversampling Technique
NAO: North Atlantic Oscillation
NEU: Northern Europe
NOAA: National Oceanic and Atmospheric Administration

NPSPEI: Nonparametric Standardised Evapotranspiration and Precipitation Index
NSGA-II: Non-dominated Sorting Genetic Algorithm
NUTS: No-U-Turn-Sampling
NUTS3: Nomenclature des unités territoriales statistiques 3
NWP: Numerical Weather Prediction
oGPI: optimized Genesis Potential Index
OLR: Outgoing Longwave Radiation
PCA: Principal Component Analysis
PDO: Pacific Decadal Oscillation
PIT: Probability Integral Transform
PMIP4: Paleoclimate Modelling Intercomparison Project phase 4
PV: Potential Vorticity
QBO: Quasi Biennial Oscillation
QUINN: Quantile Regression using I-spline Neural Network
REDS: Reflection Via Data Splitting
RF: Random Forest
RH: Relative Humidity
ROC: Receiver operating characteristic
RWB: Rossby Wave Breaking
SA: Simulated Annealing
SANDRA: Simulated Annealing and Diversified Randomisation
SATS: Standardised Active Temperature Sum
SDTW: Soft-Dynamic Time Warping
SIC: Sea Ice Concentration
SIS: Sure Independence Screening
SK: Survival Kendal Function
SMA: Soil Moisture Anomaly
SMHI: Swedish Meteorological and Hydrological Institute
SPEI: Standardised Precipitation and Evapotranspiration Index
SPI: Standardised Precipitation Index
SSI: Standardised Streamflow Index
SST: Sea Surface Temperature
SUR: Seemingly Unrelated Regression
SVR: Support Vector Regression
TC: Tropical Cyclone
TCG: Tropical Cyclone Genesis
TC-GPI or TCGI: Tippett Genesis Potential Index
Th: Thickness asymmetry
TNR: True Negative Rate
TPR: True Positive Rate
UNC: Uncertainty
WAIC: Widely Applicable Information Criterion
W-QEISS: Wrapper for Quasi-Equally Informative Subset Selection
WM-GPI: Wang & Murakami Genesis Potential Index
WP: Work Package
WPS: Web Processing Service
WR: Weather regime

EXECUTIVE SUMMARY

Detection of extreme events is of primary importance in climate science. The concept of EE detection includes their definition based on physical processes or impacts, the identification of features and phenomena which determine or pre-date them, and forecasting - detection with early warning. All these aims are linked; identifying the features that contribute to the occurrence of these phenomena can impact how we describe and measure EE, and can ultimately help improve early actions and prompt communication to institutions and stakeholders. On the one hand, new studies on EE can corroborate expected relations. On the other hand, they may shed light on unexplored behaviours among various features, such as the events themselves, the indices that are used to define them, and the (observed or forecasted) values of weather variables at various temporal and spatial scales.

The role of WP3 in CLINT has been to advance traditional EE detection methods through the use of novel ML algorithms and tools developed (in WP2), and to then generate AI-enhanced forecasts (to be assessed in WP6-7). This chain of work has included the refinement of existing detection techniques/indices using ML, building on and creating knowledge of the complex relationships between EE and large-scale fields, the development of new (data-driven) forecasting systems or the enhancement of existing systems. The focus has been EE at short-term to S2S to seasonal time scales based on the most extensive climate data records available (e.g. ERA5, S2S forecasts, CMIP5/6 simulations).

This deliverable presents the main highlights of work performed on the selection of drivers and machine learning algorithms to detect and predict the four types of climate Extreme Events (EE) considered in CLINT: tropical cyclone genesis and extratropical transitions; heatwaves and warm nights; extreme droughts; and compound events and concurrent extremes.

For each type of EE, a description of the datasets used, considered indices, and inspected models is provided. This is followed by a discussion on which of the candidate features indicated in D3.1 have been found to be the most relevant and effective. The skills of the implemented methods were compared to pre-existing ones and climatological baselines, obtaining indications about which methods to select and how to implement them most effectively. Finally, it is possible to highlight the implications of these findings on the physical understanding of the phenomena.

The advancement in the detection of Tropical Cyclone Genensis (TCG) at long time scales has focused on developing new Genesis Potential Indices (GPI). The original Emanuel and Nolan Genesis Potential Index (ENGPI) has been enhanced to an optimized version, oGPI, which has shown improvements in both spatial and temporal correlations with observed TC activity. Moreover, a new machine learning based index has been developed, the Artificial Intelligence enhanced Genesis Potential Index (AI-GPI). This model utilizes data from the

ERA5 reanalysis dataset and observed TC genesis data from the IBTrACS project, employing a Convolutional Neural Network (CNN) paired with a spatial redistribution operator to predict monthly TC activity. The AI-GPI outperforms the ENGPI and, in certain sub-basins, the oGPI, particularly in terms of spatial correlation and identifying cyclogenesis hotspots. However, challenges remain in matching the interannual variability with observed data, notably in sub-basins with fewer cyclogenesis events, which impacts the training dataset's robustness.

To enhance the detection of TC activity on the medium range, efforts have been made to test a variety of AI model architectures on a large pool of feature variables. Despite the provision of various influencing variables and the optimisation of their representation, the largest skill improvements have resulted from adding previous predictions of the target variable, near real-time observations, or from using daily-averaged instead of instantaneous feature values. Among the purely data-driven approaches, the U-Net architecture has turned out to be most useful, exceeding the skill of climatological predictions out to day four. This architecture has been further tested in a hybrid mode, where the ERA5-trained model for predicting TC activity has been fed with IFS control forecasts as input. This approach has resulted in a much slower decline in prediction skill with lead time, and therefore paves the way for future developments.

For existing TCs in the North Atlantic, the probability of extratropical transition is another application for which ML models have been developed. Their performance has been evaluated against forecasts based on the ECMWF ensemble and climatological probability. The decomposition of the Brier score has revealed why no ML model is able to outperform the ECMWF ensemble. Even though the ML models have been all better calibrated, they considerably lack discriminative ability with respect to the binary outcome. The genesis position has been identified as the most relevant predictor and logistic regression as the best model, indicating that non-linear dependencies are not yet sufficiently represented in predictor data and/or modelling approaches.

The work on extreme temperature events has employed traditional and ML-based methods to tackle a diverse range of problems. First, alternative health-based heatwave indices, such as humidity-based night-time temperatures (warm nights), have been studied to complement existing knowledge on (daytime) heatwaves. Driver detection and seasonal forecast validation analysis. ML approaches have been used to define potential predictors (such as weather regimes) as well as identifying indicators which explain agricultural impacts. Analysis of HWs in a paleo-climate simulation has supported ML-based approaches which required longer-term training datasets than available in reanalysis. Meanwhile, a spatio-temporal optimization-based feature selection framework, developed in T2.4, has been applied to three problems: (1) the identification of HW predictors, (2) production of a data-driven and (3) a hybrid seasonal forecast systems, which both add value over existing dynamical systems either through computational resources used or forecast skill.

The detection of Extreme Droughts has been the subject of extensive research in the past decades, with the development of a comprehensive set of indices (e.g., SPI, SPEI, and SMA). However, these methods failed to properly reproduce drought impacts, such as the Fraction of Absorbed Photosynthetically Active Radiation (FAPAR) used to capture drought-induced stress on crops. The analysis considered 35,408 sub-basins in the pan European domain. The FRamework for Index-based Drought Analysis (FRIDA) has been used to construct new composite drought indices that consider drought impacts. In particular, it has helped identify the basin characteristics and extract the most relevant features. This process has helped detect the relevant drivers for each cluster and enhanced the impact-based drought indices.

Our studies have examined the impact of compound climate events on Europe's food, water, and energy sectors. A case study on winter wheat in France has revealed significant crop losses caused by a combination of wet and warm conditions in January and February, followed by dry conditions in April and warmth in May. Random forests and decision trees have provided accurate local-scale predictions, with critical thresholds often being lower than traditional extreme event thresholds, indicating that severe damage can result from non-extreme conditions. We have shown that the impact of hot summers on agriculture and energy can be significantly accelerated when preceded by dry winters. The risk of crop failure is notably higher with the inclusion of dry winter soils. The state of winter soils also increases the predictability of hydropower availability. Finally, our third study has revealed that extreme flood risk in the Alpine region increases when wet springs follow cold winters.

The analysis of concurrent extremes has focused on detecting dependencies between heatwaves and droughts, which are known to cause severe impacts across agriculture, energy, water resources, and health. Leveraging methods from point process theory and Deep Learning, we analyse global interdependencies of these extreme events, accounting for their changing frequencies due to climate change. This developed tool enables the analysis of interdependencies between thousands of extreme events within seconds in a changing climate, demonstrating that European droughts and heatwaves are connected to those events in many other regions of the world. Additionally, a nonparametric version of the Standardized Precipitation and Evapotranspiration Index (SPEI) has been proposed, offering enhanced extrapolation during reference periods and improved performance for extreme event detection.

Finally, many of the detection methods presented here are being applied across the project in either pan-European (WP6) or local-scale (WP7) case studies. Others are being prepared for deployment in the Climate Services Information Systems (WP8), or in the demonstration of prototype operational prediction services (WP9). The application and operationalisation of these tools will ensure their continued development within and beyond the project.

1 INTRODUCTION

The detection of extreme events is an important step towards understanding the mechanisms that drive these phenomena. This includes the exhaustive identification of indices that help define the events or their potential occurrence, and the analysis of which weather variables influence the development of the phenomena themselves. ML algorithms help these processes, enhance the existing methods and develop new approaches that can improve detection and prediction of extreme events. ML models require a thorough prior inspection of which drivers and algorithms to use. This helps to avoid overparameterization and reduces the computational time.

Good ML models need to be trained on large and consistent datasets. In climate sciences these characteristics are provided by reanalyses (e.g., ERA5). Although reanalyses are partly built upon model approximations, they provide spatial and temporal consistency which is not guaranteed by observations. At the same time, some other variables which can provide more specific information about the EE can be retrieved from specific datasets (e.g., IBTrACS and FAPAR). They can also be used as a benchmark for the training of ML models.

As a first step in ML development, it is fundamental to properly select the drivers (or predictors) that are given as inputs to the models and to assess which model is the most appropriate for the characteristics of the problem at hand. A thorough selection of these aspects allows one to build a solid algorithm to avoid over-parameterization and reduce the computational time. The resulting models may improve or compete with existing models.

Furthermore, the training process benefits from an accurate driver selection. Identifying the best drivers requires a different approach according to the physical and statistical characteristics of each phenomenon considered. In addition, the detection and the prediction of events imply the use of different drivers. While the former requires only local predictors, the latter also needs the inclusion of temporally and spatially remote predictors. In this deliverable, this selection is discussed, in combination with the evaluation of the algorithms and ML techniques that performed best in each case.

For each of the following Extreme Events (EE):

- Tropical cyclones: in terms of genesis and activity on different timescales (Chapter 2) and extratropical transitions (Chapter 3).
- Heatwaves and warm nights (Chapter 4)
- Extreme droughts (Chapter 5)
- Compound events and concurrent extremes (Chapter 6)

The report provides an overview of the problem, summarises the data used, describes the features of the inspected algorithms, explains the results, and finally analyses the physical and statistical implications.

This report builds on the previous milestones and deliverables from WP2 and WP3, as illustrated in Figure 1.1.

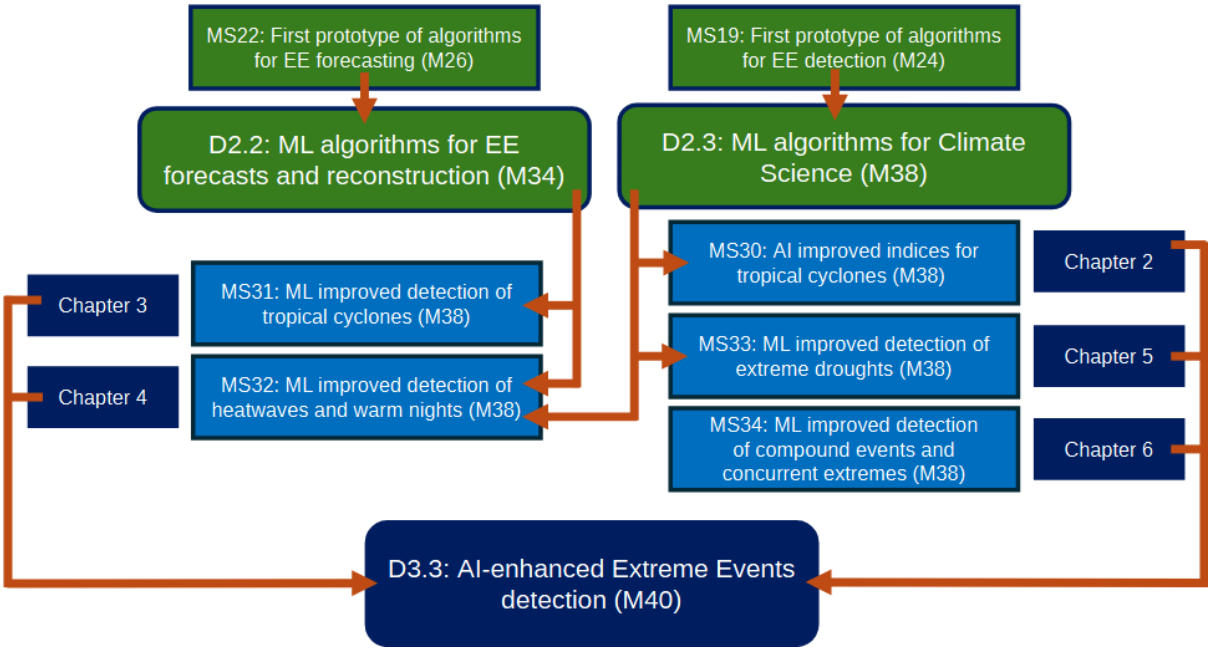


Figure 1.1: Schematic representation of D3.3 and its interconnection with previously completed D and MS of WP2 and WP3.

2 TROPICAL CYCLONES: INDICES

2.1 Introduction

Tropical Cyclones (TCs) form at a rate of about 80-90 per year globally in the tropical latitude bands on both sides of the Equator (Walsh et al., 2016). TCs making landfall are among the costliest and deadliest natural disasters, due to the strong winds, heavy precipitation and risk of storm surges associated (Mendelsohn et al., 2012). Therefore, it is of paramount importance to be able to accurately predict their activity at several timescales, ranging from a few days, to the seasonal, up to the climate projections scale. Unfortunately, a complete theory of TC formation is so far lacking.

Over the years, several indices have been developed to either directly forecast the genesis of TC, or estimate how prone the atmosphere in a given region is to the formation of TCs (see Milestone MS6, "*Indices, Datasets, and Candidate Drivers for Tropical Cyclones*", Table 1, for an overview). The most widely used of these indices—developed by (Emanuel & Nolan, 2004)—is known as the Emanuel and Nolan Genesis Potential Index (ENGPI) and aims at describing the climatological distribution and seasonal variations of TCs. Its functional form is given by:

$$ENGPI = \left|10^5 \eta\right|^{3/2} \left(\frac{H}{50}\right)^3 \left(\frac{MPI}{70}\right)^3 (1 + 0.1\,V)^{-2} \quad (2.1)$$

where η is the absolute vorticity at 850 hPa, H the relative humidity at 600 hPa, V the vertical wind shear between 200 hPa and 850 hPa, and MPI (Maximum Potential Intensity) a theoretical estimate of the maximum sustained wind speed a TC could reach in a given environment.

As discussed in D3.2, "*Preliminary AI-enhanced extreme events detection"*, the ENGPI performs well in terms of spatial correlation (Cavicchia et al. 2023), with observed data (i.e., it correctly identifies zones prone to the formation of TCs), but poorly in terms of interannual correlation (i.e., it fails to correctly estimate the exact number of TCs forming each month) in both historical simulations and future projections (Figure 1).

The findings reported in this Chapter will be part of Dainelli et al. (in preparation).

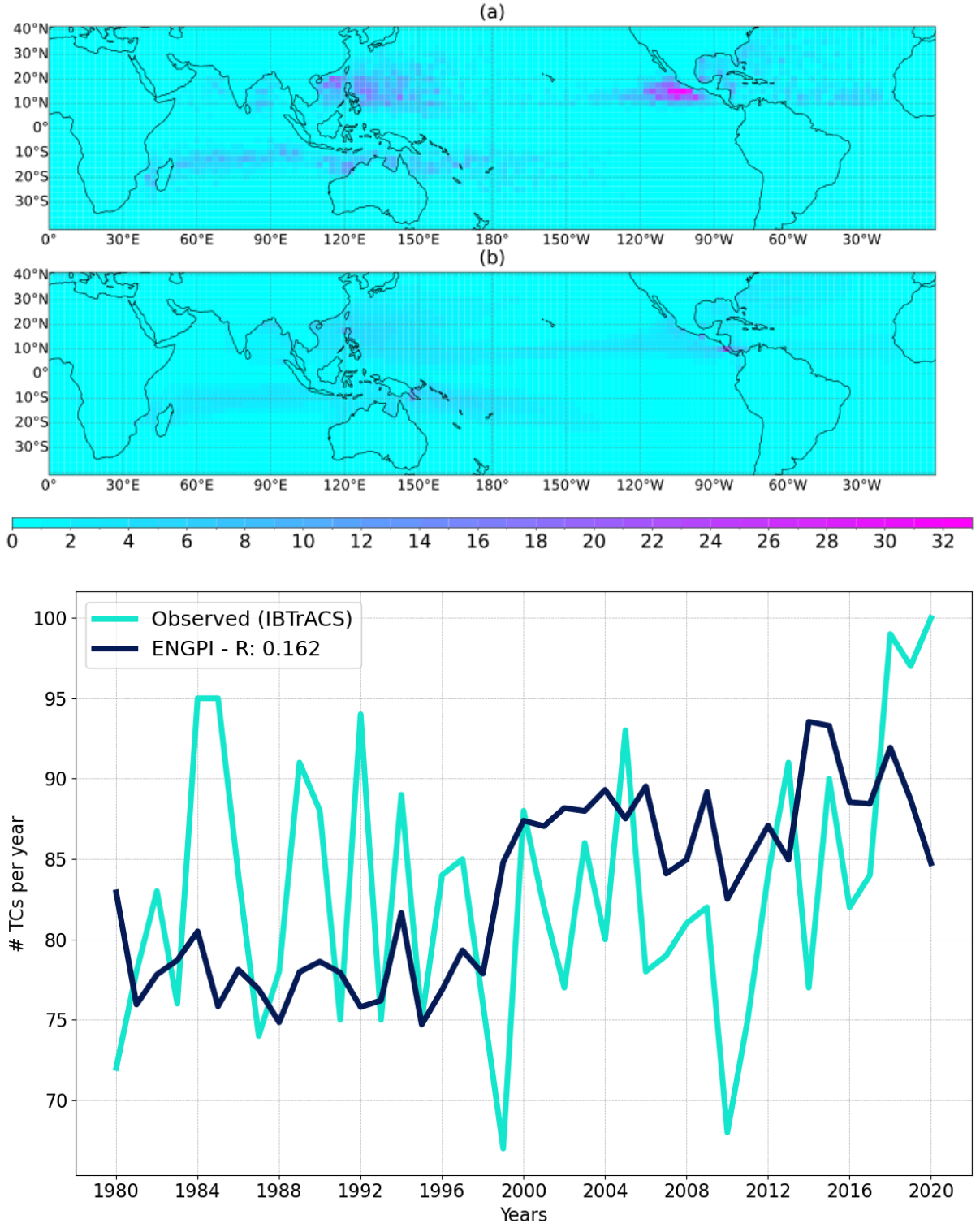


Figure 2.1 Top: spatial distribution of TC genesis according to observed data (IBTrACS, panel A) and to the ENGPI (panel B). Bottom: Interannual variability curves for the ENGPI.

2.2 AI-enhanced indices for TC detection

2.2.1 Optimization of the ENGPI: the oGPI

As already described in D3.2, our first step towards the formulation of AI-enhanced indices for TC detection is to use a genetic algorithm (NSGA-II) to optimize the coefficients that enter the definition of the ENGPI in Eq. (X), as well as the pressure levels at which the variables are taken. We refer readers to D3.2 for further details on this method, and limit the discussion here to just the salient points regarding the performance of this method, to be compared with the next method we developed.

The best combination of parameters for the improved ENGPI formula is identified by selecting those that showed a good improvement in spatial and temporal correlations; we call the resulting optimized index *oGPI* (Ascenso et al., 2023). Its functional form is given by:

$$oGPI = \left|10^{5}\eta_{600}\right|^{2.01}\left(\frac{H_{700}}{43.67}\right)^{3}\left(\frac{MPI}{68.39}\right)^{1.77}(98.9 + 23.51\,V)^{-2.91}\,e^{2.98} \quad (2.2)$$

The oGPI performs better than the original ENGPI in terms of both spatial and interannual correlation with observations (Figure 2.1). However, the interannual correlation is still relatively poor. Therefore, we moved forward to more advanced ML methods to further improve these initial results.

2.2.2 AI-enhanced Genesis Potential Index (AI-GPI)

This work aims at further improving the detection of TC occurrence, and in particular their interannual variability and future trends, by foregoing the more classical numerical formulation of indices and replacing it with a ML algorithm.

2.2.2.1 Data

As data to train our ML-based model, we considered the monthly mean maps of the environmental factors used to compute the ENGPI and the monthly maps of observed Tropical Cyclone Genesis (TCG). The source for the predictors of our algorithm is the ERA5 reanalysis dataset (Hersbach et al., 2020, Scoccimarro et al., 2024), with a spatial resolution of 2.5°x2.5°. The source for the predictand is the best-track data from the International Best track Archive for Climate Stewardship (IBTrACS) project (Knapp et al., 2010), at a temporal resolution of 3h, from 1980 to 2021. We refer readers to section 3.2.4 of D2.3 for more details regarding the data and their processing.

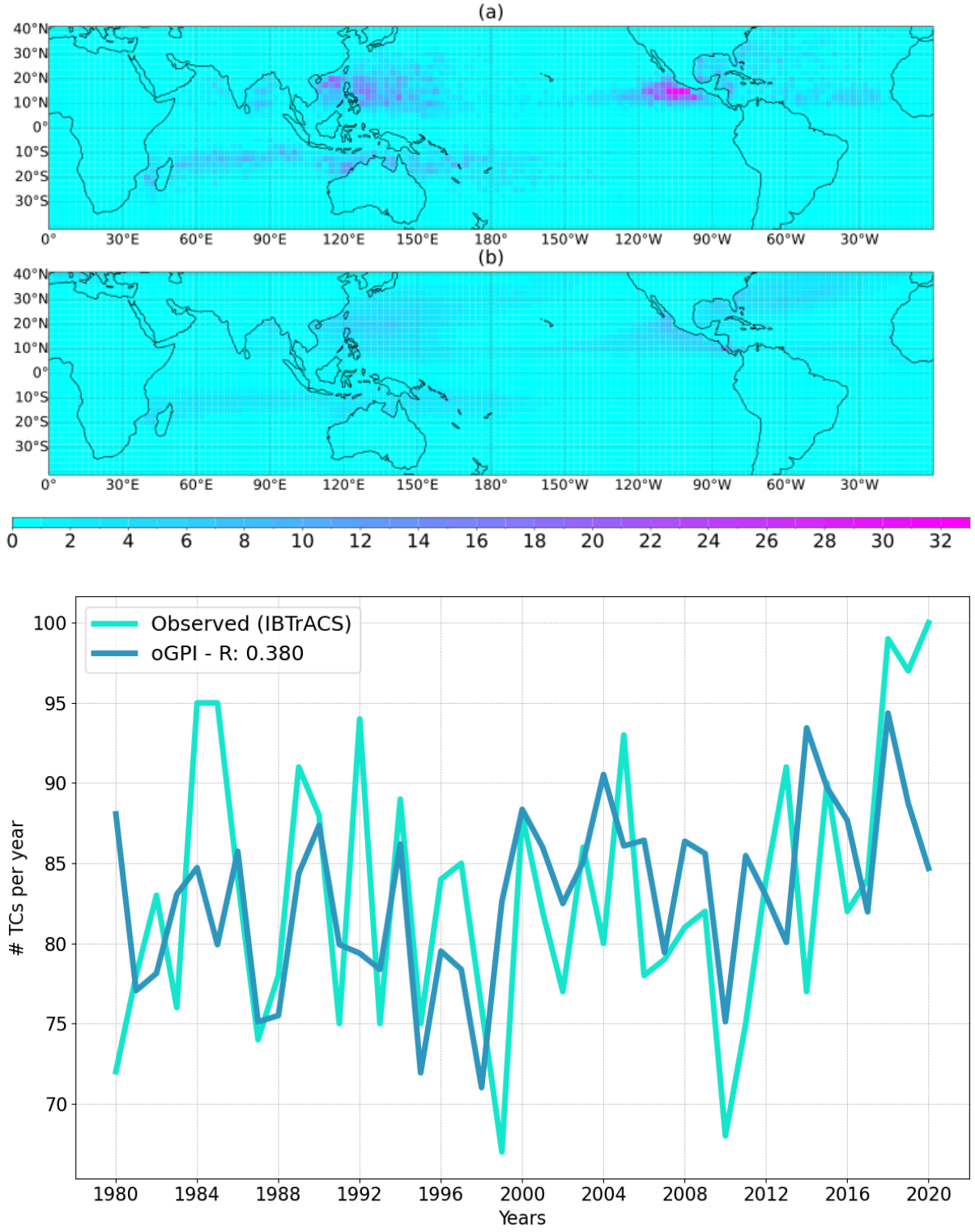


Figure 2.2 Top: spatial distribution of TC genesis according to observed data (IBTrACS, panel A) and to the oGPI (panel B). Bottom: Interannual variability curves for the oGPI.

2.2.2.2 Methods

The ML-based model we develop is composed of two components, a Convolutional Neural Network (CNN) and a spatial redistribution operator. The CNN performs a linear regression task by estimating the total amount of TC occurrences in a specific month given the predictor maps. The operator spatially redistributes this total number across the domain by a random assignment of the position based on the monthly frequency maps of cyclogenesis. We have developed seven different models: one covering the global region of the tropics (40°S-40°N) and the rest covering each tropical sub-basin experiencing cyclone activity. The boundaries of the sub-basins are defined following Fudeyasu (2014) and shown in Figure 2.2. We refer readers to section 3.2.4 of D2.3 for more details regarding the CNNs architectures and the details of the spatial redistribution operator.

In line with the methodologies outlined in D2.3, we further develop the model training process by integrating K-fold cross-validation to partition the dataset into training and test sets. This cross-validation technique is employed to address the shortcomings of calculating the interannual correlation between two time series consisting of seven values each, thereby improving the metric's robustness. This method enables us to calculate the interannual correlation between observations and the AI-GPI's predictions across the entire dataset spanning from 1980 to 2021, excluding training data from the metric computation. We divide the dataset into five folds and train five distinct models on different years. Table 2.2 details the years included in the training and test sets for each fold. By adopting the approach of K-fold cross-validation for each of the seven basin we developed an ensemble of 5 models.

***Table 2.1**:* Years of the dataset being part of the training set and the set for the five different folds of cross-validation.

	Fold 1	**Fold 2**	**Fold3**	**Fold 4**	**Fold 5**
Years in the training set	1989-2020	1980-1988 1997-2020	1980-1996 2005-2020	1980-2004 2013-2020	1980-2012
Years in the test set	1980-1988	1989-1996	1997-2004	2005-2012	2013-2020

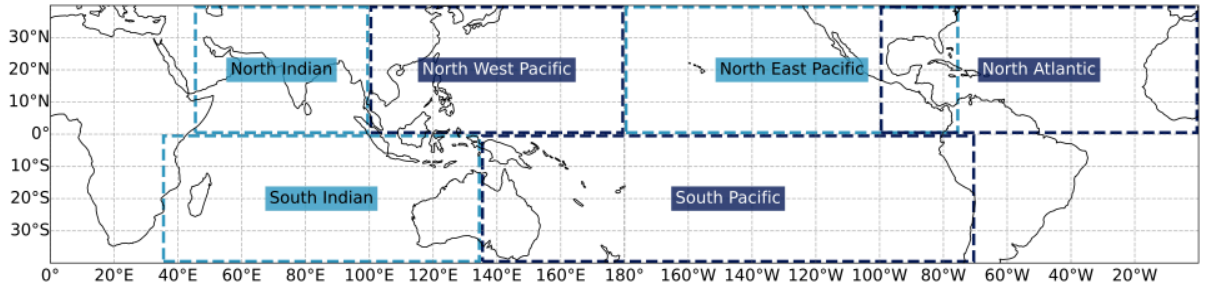


***Figure 2.3**:* Sub-basin domains extension.

***Table 2.2**:* Summary of interannual and spatial correlations for the ENGPI, oGPI, and AI-GPI across different domains (Tropics and sub-basins).

Basin	R	ENGPI	oGPI	AI-GPI
Global	Annual	0.162	0.380	0.262
	Spatial	0.446	0.490	0.771
North Atlantic	Annual	0.720	0.436	0.530
	Spatial	0.320	0.324	0.681
Northeast Pacific	Annual	0.616	0.669	0.563
	Spatial	0.472	0.628	0.878
Northwest Pacific	Annual	-0.114	0.088	0.311
	Spatial	0.480	0.428	0.693
South Pacific	Annual	0.089	0.196	0,305
	Spatial	0.372	0.429	0.606
North Indian	Annual	0.205	0,154	0,118
	Spatial	0.312	0.324	0.562
South Indian	Annual	0.073	0.105	0.267
	Spatial	0.454	0.509	0.690

2.2.2.3 Results

We evaluate the AI-GPI by considering its performances in terms of spatial and interannual correlation with respect to the observed cyclogenesis on the dataset (1980-2020), which we then compare to that of the ENGPI and oGPI globally and on a per-basin level (Table 2.1). The spatial correlation is defined as the cell-by-cell correlation between the maps of mean TC genesis per year of the AI-GPI and the observations. The interannual correlation is defined as the correlation between the curves of yearly TC genesis of the AI-GPI and the observations. It is important to highlight that to calculate the annual mean TC genesis map and the yearly TC genesis curve, we utilize predictions derived from the concatenated outputs of the five ensemble models. Specifically, for each test set split, we calculate the monthly TCG likelihood maps using the model trained on the corresponding training set. We then concatenate these monthly predictions in accordance with the dataset's chronological sequence, resulting in a time series of monthly TCG likelihood maps. From this time series we obtain the mean genesis map and the yearly genesis curve.

Figure 2.4 reports the time series of yearly TC genesis in the Tropics based on the observed data, the ENGPI, the oGPI, and the AI-GPI. Additionally, the figure includes the interannual correlations for each index. The AI-GPI performance is higher than the ENGPI in terms of interannual correlation, though it is lower than that of the oGPI. Generally, all three indices capture the overall trend observed in the data. Our method more accurately aligns with the peaks and troughs of the observation curve, despite some variations in magnitude. It is crucial to note that the oGPI is optimised using the entire dataset simultaneously, whereas our model ensembles are trained fold by fold, without ever analysing the complete dataset entirely. We believe this approach contributes to the superior performance of the oGPI on a global scale. The last three years of the test set present significant challenges for the network to model, as these years exhibit the highest cyclone activity in our dataset (refer to the observation curve in Figure 2.1).

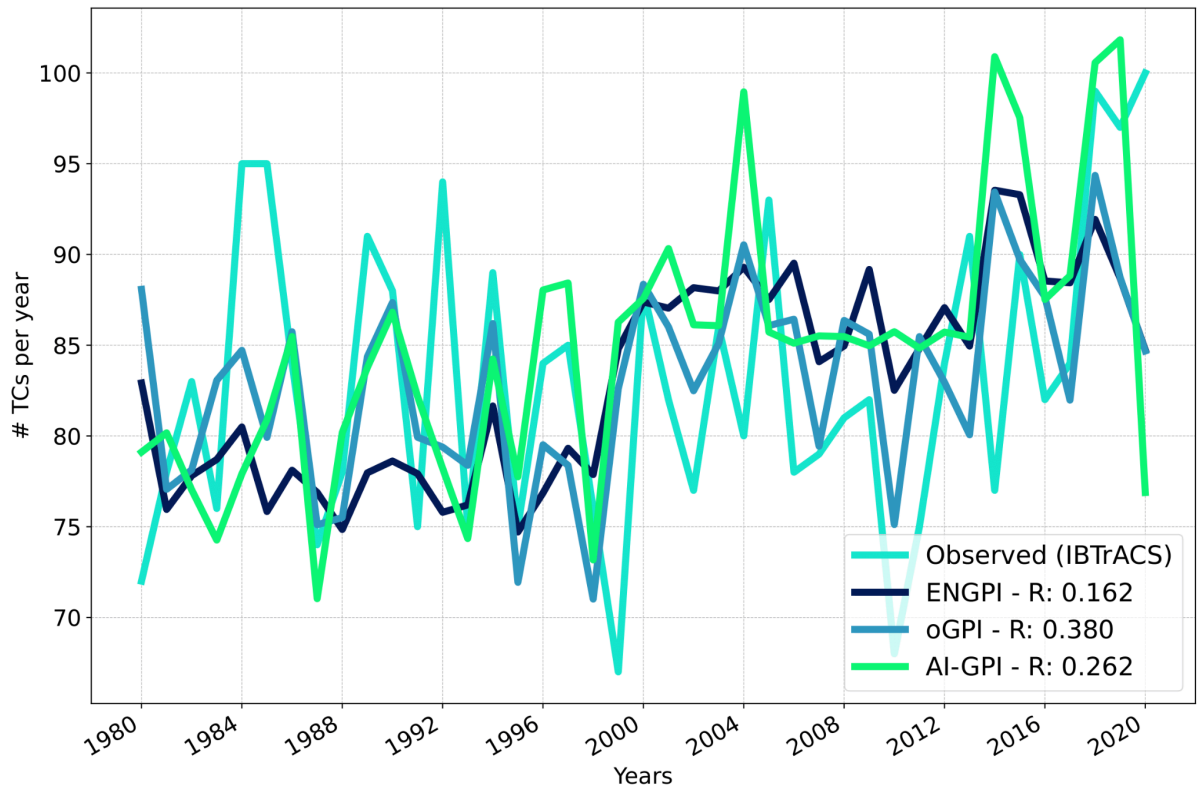


***Figure 2.4**:* Interannual variability curves for the Tropics domain comparing the observed cyclogenesis to the ENGPI, the oGPI, and the AI-GPI.

Figure 2.5 displays maps of the average number of cyclogeneses per year per grid point for (a) observed data, (b) the ENGPI, (c) the oGPI, and (d) the AI-GPI. The figure also includes the spatial correlation for each index. The AI-GPI shows superior performance in terms of spatial correlation compared to the other two methods. The maps reveal that the AI-GPI more accurately identifies the regions and hotspots of higher mean cyclogenesis. The extent of these areas and the magnitude of the average cyclogenesis frequency show a closer alignment with observed data compared to the ENGPI and oGPI.

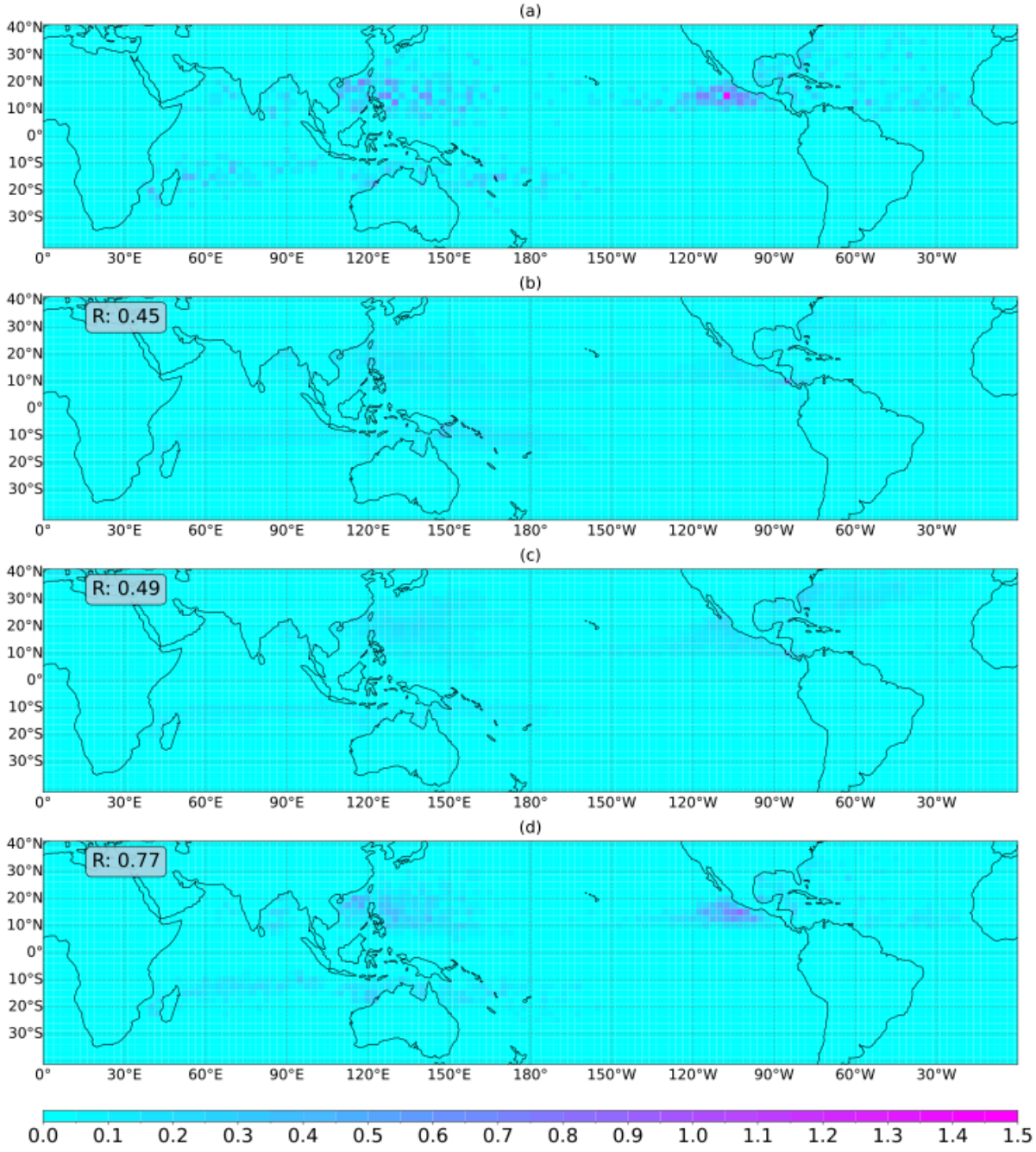


Figure 2.5: Spatial distribution of TC genesis in the Tropics according to (a) observed data, (b) ENGPI, (c) oGPI, and (d) AI-GPI. The coloured map indicates the average number of cyclogeneses per year per grid point.

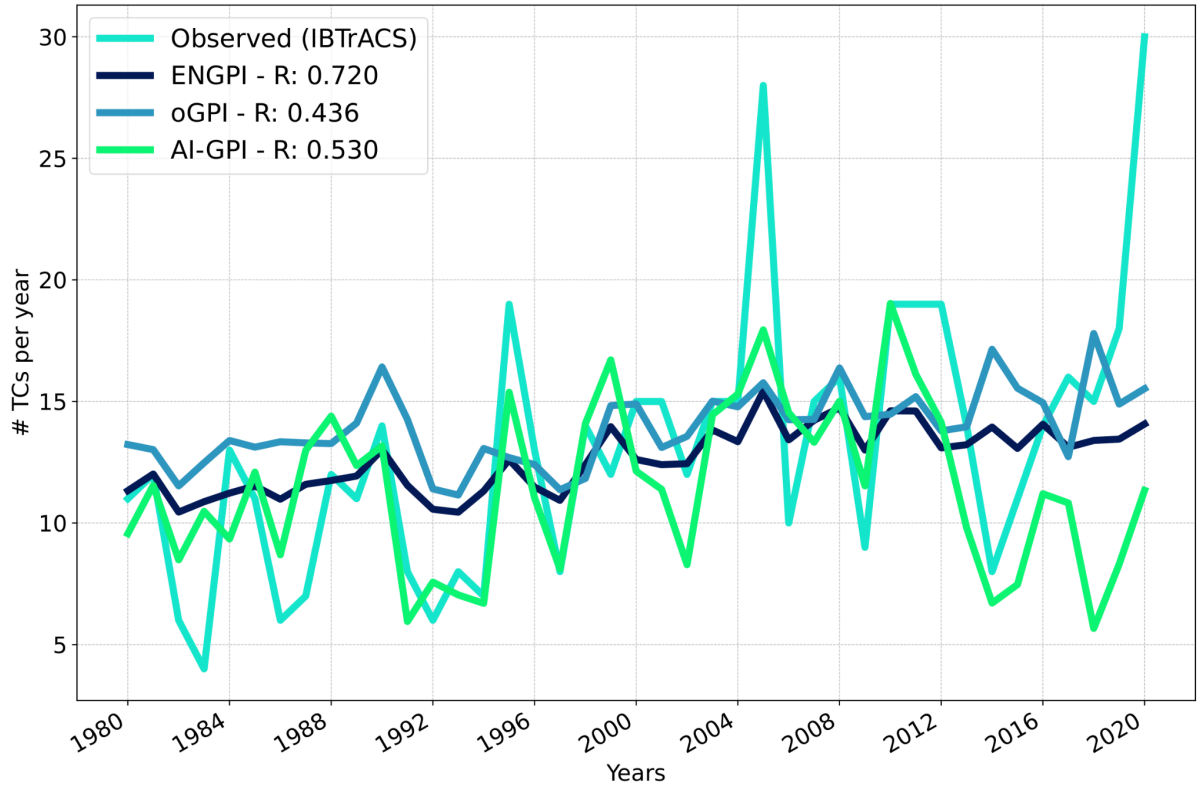


Figure 2.6: Interannual variations of cyclogenesis in the North Atlantic domain comparing the observed cyclogenesis to the ENGPI, the oGPI, and the AI-GPI.

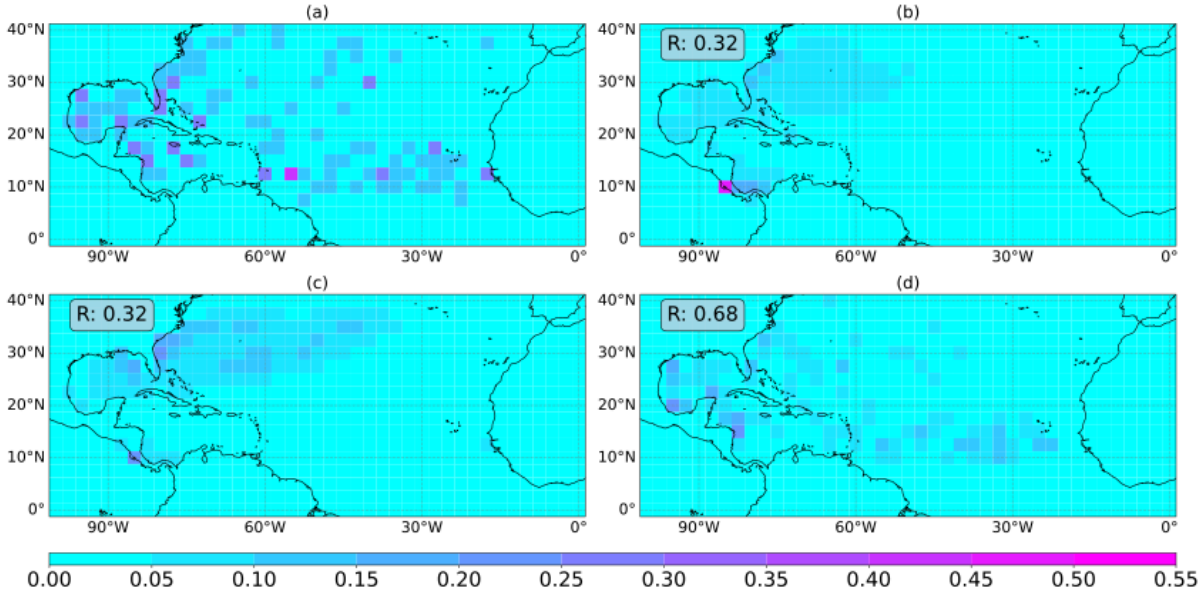


Figure 2.7: Spatial distribution of TC genesis in the North Atlantic according to (a) observed data, (b) ENGPI, (c) oGPI, and (d) AI-GPI. The coloured map indicates the average number of cyclogeneses per year per grid point.

Figures 2.6 and 2.7 present the yearly curves and annual mean maps of TC occurrences for the North Atlantic. Here, the AI-GPI outperforms the oGPI in terms of interannual correlation with observations. Although the ENGPI exhibits higher interannual correlation overall, the AI-GPI more accurately captures the peaks and troughs of the curve, particularly in the years 1991 to 1993, as well as in 1997, 2008, and 2009. The AI-GPI shows superior performance also in terms of spatial correlation. Our method more accurately captures the irregular distribution of genesis across the sub-basin and locates the regions with higher frequency of cyclogenesis in the southeastern region and in the Caribbean Sea.

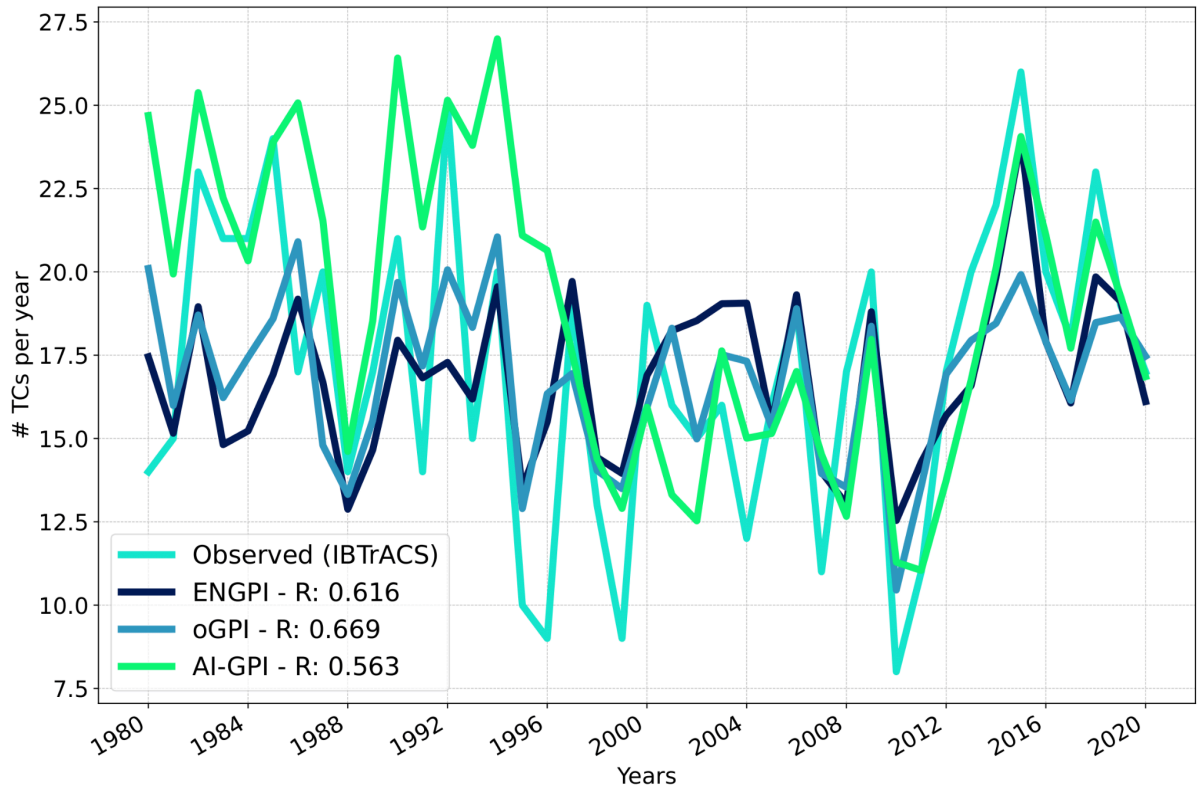


***Figure 2.8**: Interannual variability curves for the Northeast Pacific domain comparing the observed cyclogenesis to the ENGPI, the oGPI, and the AI-GPI.*

Figures 2.8 and 2.9 present the yearly curves and annual mean maps of TC occurrences for the Northeast Pacific. In this sub-basin, all three indices demonstrate fair performance in terms of interannual correlation. Both the ENGPI and oGPI exhibit higher values than the AI-GPI. However, the AI-GPI successfully captures the overall trend identified in the observations, which shows stronger cyclone activity from 1980 to approximately 1995, a decline in activity until around 2013, followed by an increase in the final years of the dataset. Considering the spatial correlation, the AI-GPI continues to excel, correctly representing the cyclogenesis hotspot west of Mexico.

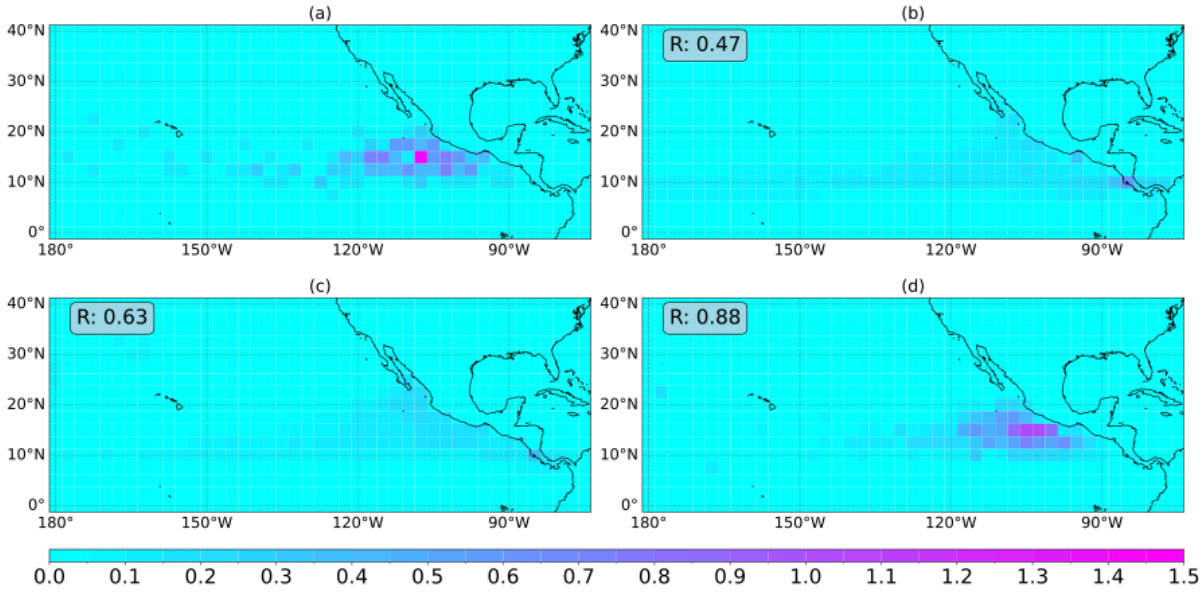


***Figure 2.9**:* Spatial distribution of TC genesis in the Northeast Pacific according to (a) observed data, (b) ENGPI, (c) oGPI, and (d) AI-GPI. The coloured map indicates the average number of cyclogeneses per year per grid point.

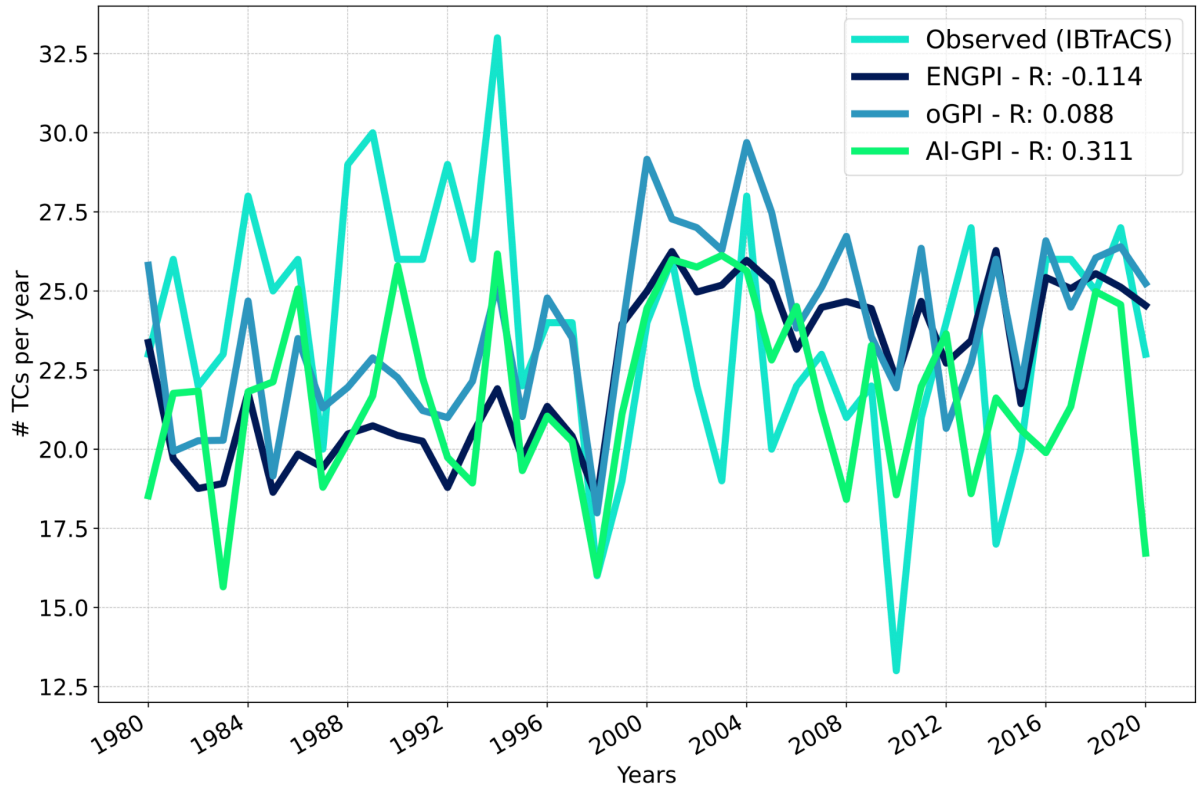


Figure 2.10 Interannual variability curves for the Northwest Pacific domain comparing the observed cyclogenesis to the ENGPI, the oGPI, and the AI-GPI.

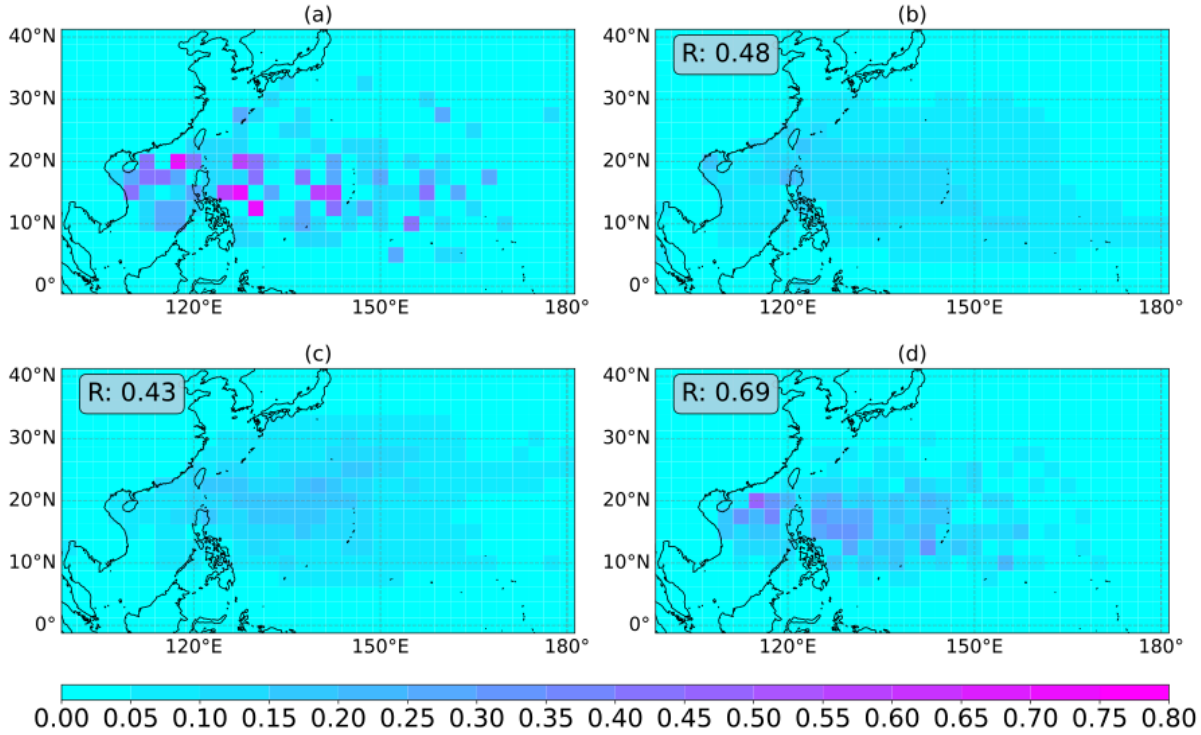


Figure 2.11: Spatial distribution of TC genesis in the Northwest Pacific according to (a) observed data, (b) ENGPI, (c) oGPI, and (d) AI-GPI. The coloured map indicates the average number of cyclogeneses per year per grid point.

Figures 2.10 and 2.11 present the yearly curves and annual mean maps of TC occurrences for the Northwest Pacific. The AI-GPI outperforms the other two methods in terms of interannual correlation. From 1994 to 2010, the AI-GPI more accurately represents the trend, displaying a pattern of decreasing, increasing, and again decreasing cyclone activity. Considering the spatial correlation, the AI-GPI continues to outperform the other method, with a more precise representation of the spatial distribution of the annual mean maps of cyclogenesis.

Figures 2.12 and 2.13 present the yearly curves and annual mean maps of TC occurrences for the South Pacific. The AI-GPI outperforms the ENGPI and the oGPI in terms of interannual correlation, as it captures the decreasing trend observed in the first 10 years of observations more accurately. Also, the AI-GPI is able to match some of the local peaks of the curve, as in years 1983 and 1987. The AI-GPI demonstrates superior spatial correlation, more accurately capturing the irregular distribution of genesis across the sub-basin and more precisely representing the frequency and magnitude of cyclogenesis within the domain.

Figures 2.14 and 2.15 present the yearly curves and annual mean maps of TC occurrences for the North Indian sub-basin. Considering the interannual correlation, the oGPI and ENGPI outperform the AI-GPI, as the AI-GPI curve deviates more significantly from the observed data than the other two. This performance gap likely stems from the lower record of cyclogenesis in the North Indian sub-basin relative to others. The reduced dataset for

training has limited the CNN's ability to accurately predict TC occurrences. The limitations are also evident in the spatial distribution, where none of the methods accurately match the observations.

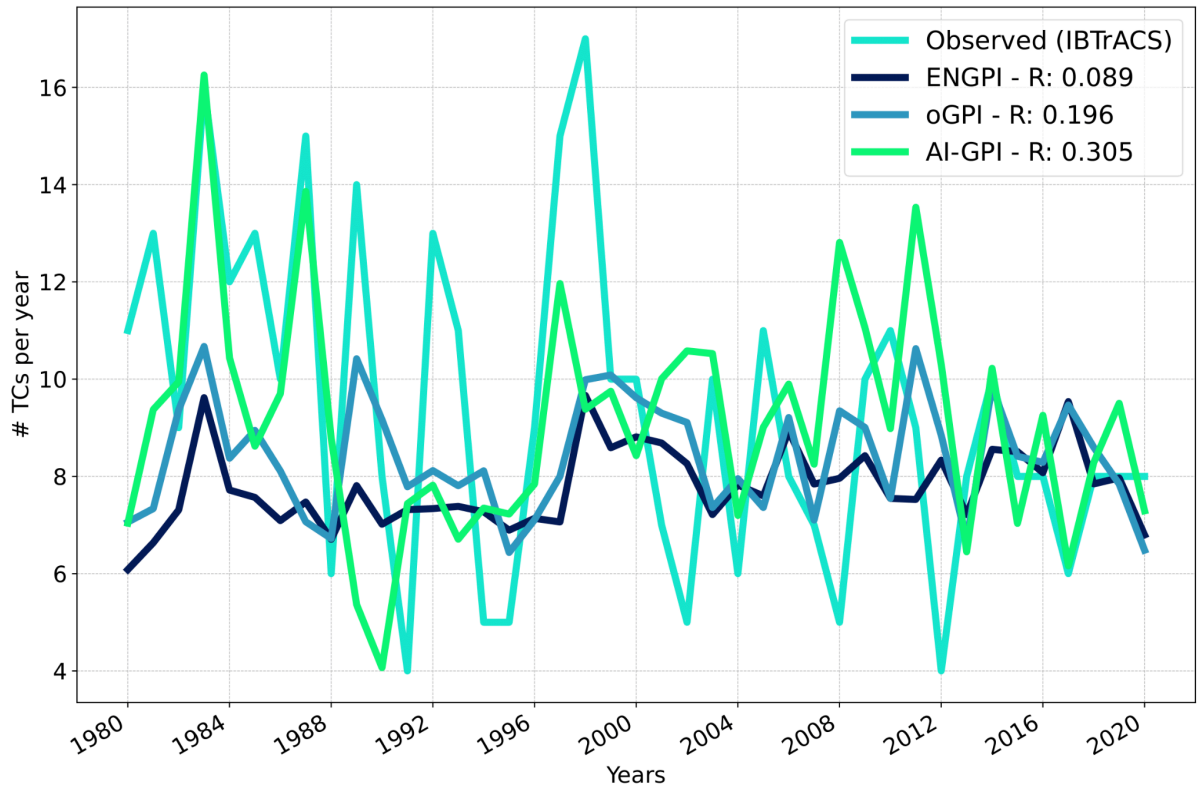


***Figure 2.12**:* Interannual variability curves for the South Pacific domain comparing the observed cyclogenesis to the ENGPI, the oGPI, and the AI-GPI.

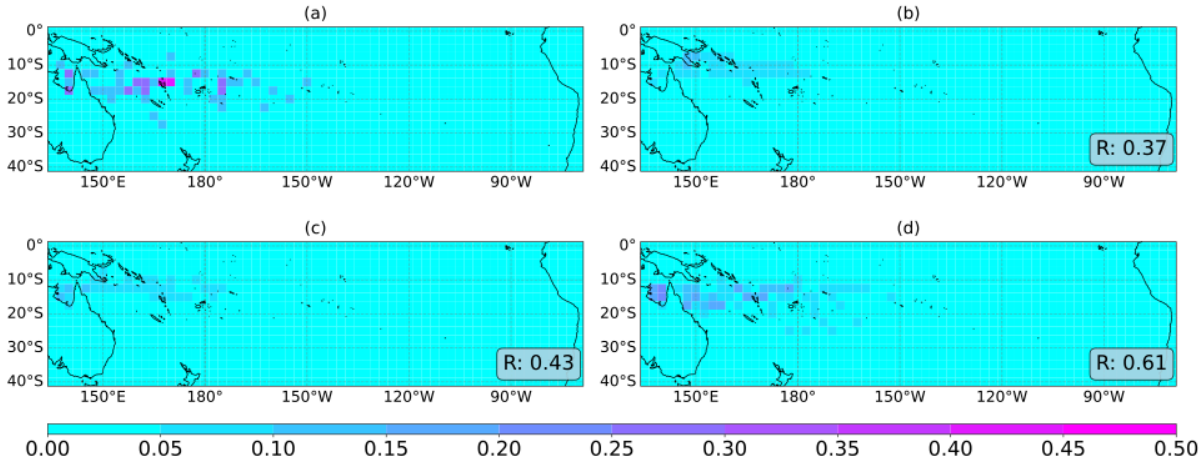


Figure 2.13: Spatial distribution of TC genesis in the South Pacific according to (a) observed data, (b) ENGPI, (c) oGPI, and (d) AI-GPI. The coloured map indicates the average number of cyclogeneses per year per grid point.

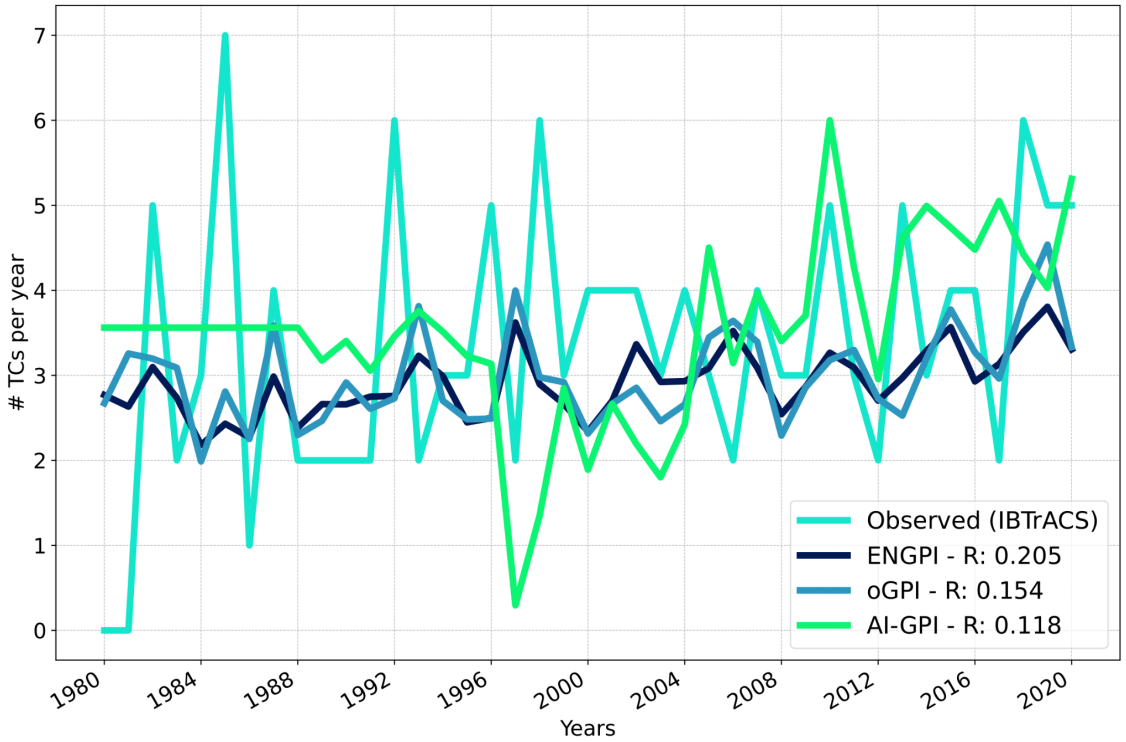


Figure 2.14*:* Interannual variability curves for the North Indian domain comparing the observed cyclogenesis to the ENGPI, the oGPI, and the AI-GPI.

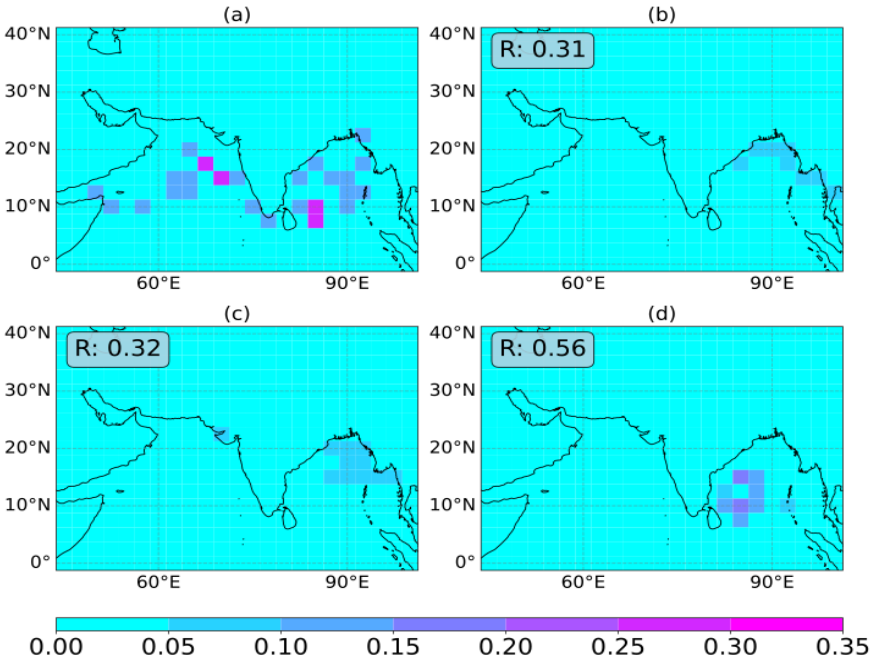


Figure 2.15: Spatial distribution of TC genesis in the North Indian according to (a) observed data, (b) ENGPI, (c) oGPI, and (d) AI-GPI. The coloured map indicates the average number of cyclogeneses per year per grid point.

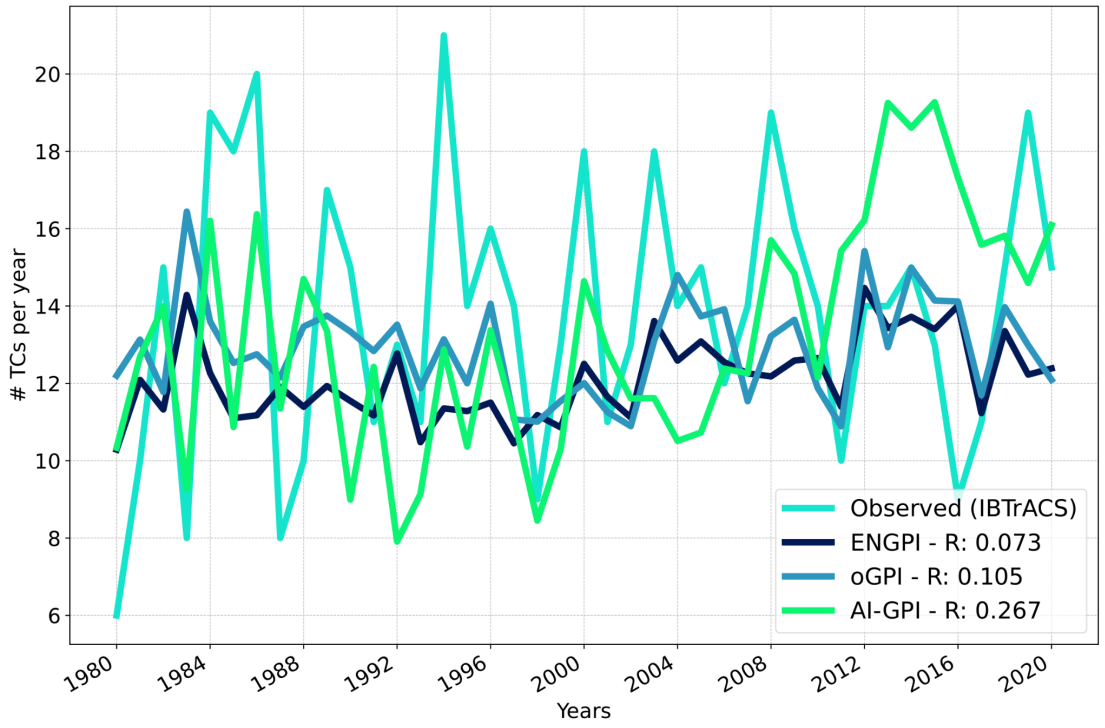


Figure 2.16*:* Interannual variability curves for the South Indian domain comparing the observed cyclogenesis to the ENGPI, the oGPI, and the AI-GPI.

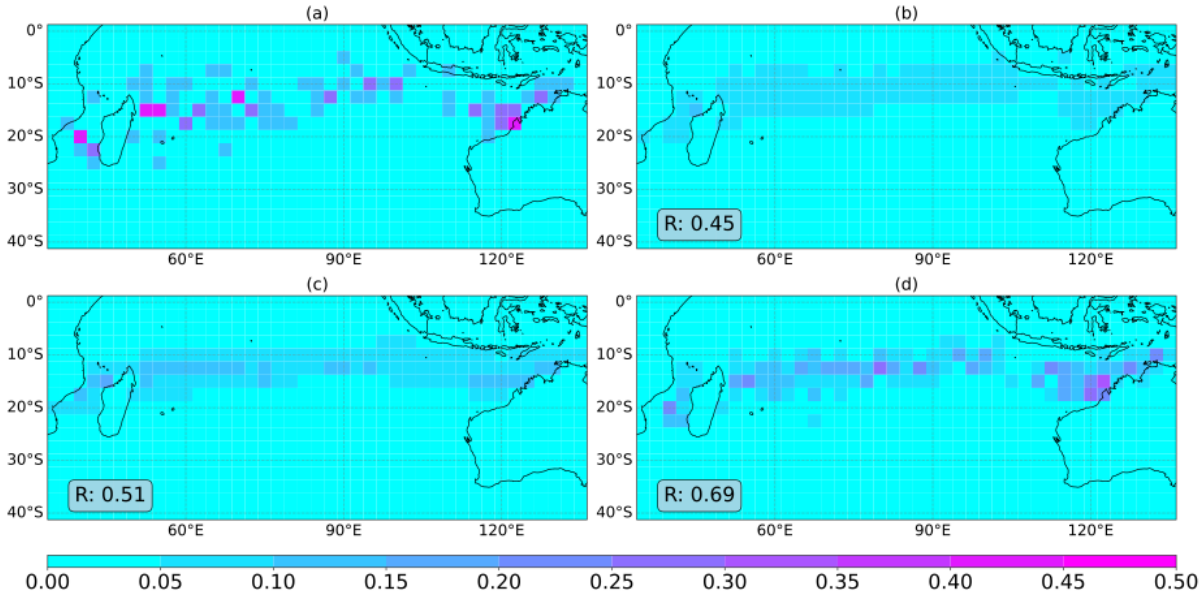


***Figure 2.17**: Spatial distribution of TC genesis in the South Indian according to (a) observed data, (b) ENGPI, (c) oGPI, and (d) AI-GPI. The coloured map indicates the average number of cyclogeneses per year per grid point.*

Figures 2.16 and 2.17 present the yearly curves and annual mean maps of TC occurrences for the South Indian sub-basin. The AI-GPI outperforms the other two methods in terms of interannual correlation. Also, the AI-GPI excels in spatial correlation, more accurately capturing the irregular distribution of genesis across the sub-basin and identifying regions with higher cyclogenesis frequency, with a similar magnitude to the observations.

In summary, the performance of the AI-GPI, oGPI, and ENGPI indices varies across different metrics and regions. The AI-GPI demonstrates superior spatial correlation, accurately capturing the irregular distribution and frequency of cyclogenesis in several sub-basins. On a global scale, across the entire Tropics, the AI-GPI shows a higher interannual correlation with observations than the oGPI. We attribute this limitation to the capability of the oGPI to access the entire dataset for the optimization process. When examining individual sub-basins, no index consistently outperforms the others. However, in sub-basins with higher cyclone activity, such as the Northwest Pacific, the AI-GPI shows better performance in terms of interannual correlation. Additionally, in other sub-basins with frequent occurrences of TCG, such as the South Indian and the Northeast Pacific, our method exhibits notable performance. The AI-GPI's limitations are particularly evident in the North Indian sub-basin, where a reduced training dataset has diminished its predictive accuracy.

2.2.2.4 Future Developments

Our future development of the AI-GPI will focus on two goals: evaluating additional predictors for estimating the number of cyclogenesis events per month and developing a multi-spatial input to spatial output model.

For the first goal, we plan to implement a Feature Selection (FS) algorithm to identify the most significant factors influencing cyclogenesis estimation in each basin. We intend to include candidate variables such as atmospheric, oceanic factors and climate indices that have been previously studied in the literature or that represent processes linked to cyclogenesis.

For the second goal, we plan to develop a more advanced ML model that better aligns with the complexity of the problem. Our goal is to create a model capable of handling diverse spatial inputs and producing a spatial output as a monthly map of TC genesis likelihood. To achieve this, we will eliminate the use of monthly frequency maps and focus on developing a more sophisticated architecture than the two-component model we have used so far. Specifically, we aim to build an Autoencoder (Goodfellow et al., 2016), initially using Convolutional Neural Networks (CNNs) for the encoder and decoder (LeCun et al., 1989), with the intention of later incorporating Graph Neural Networks (GNNs) (Kipf & Welling, 2017).

As a final step, we aim to merge these two efforts by integrating the FS algorithm with the Autoencoder-based model, resulting in a final version of the AI-GPI that directly estimates monthly maps of TC genesis leveraging the information of the most significant variables per basin.

3 DETECTION OF TROPICAL CYCLONES

3.1 Tropical Cyclone Activity

3.1.1 Overview

Tropical cyclones (TC) form at a rate of approximately 80-90 times per year globally, with about 10 of them occurring in the Southern Indian Ocean (here defined as 0°-30°S, 20°-90°E, and referred to as “SIO” hereafter). With a focus on the medium-range (i.e., up to two weeks lead time) time scale, the goal of subtask 3.1.2. was to leverage AI to enhance the prediction of TC activity. A number of large-scale atmospheric and oceanic factors are known to be relevant for TC formation, such as a moist enough atmosphere in the lower and middle troposphere, an incipient low-level vortex, moderate vertical wind shear, and SSTs high enough for the TC to fuel its convection with energy. However, a comprehensive theory of TC formation is still lacking so far. A useful predictor for TC activity is one that has a reasonable correlation with the target and shows predictive skill. The latter certainly depends on the predictor chosen and will also vary with lead time.

In the following, results are shown for the western part of the SIO, since heavy rainfall events associated with TCs in this area can pose a major threat to the Zambezi River basin

(Scoccimarro et al. 2024), one of the four selected climate change hotspots under investigation in CLINT (WP7).

3.1.2 Datasets and Forecasts

3.1.2.1 Datasets

The IBTrACS dataset version 4 (Knapp et al., 2010, 2018), which collects estimates of TC position and intensity for different ocean basins, is used in three ways, namely to i) systematically verify the forecasts of all models considered, ii) serve as the target for training ML models, and iii) calculate a reference model predicting climatological probabilities. In addition, ECMWF ensemble forecasts provide an NWP-based benchmark and constitute the dynamical component of the tested hybrid modelling approach. More details about the individual datasets are given in deliverable D3.1.

Candidate drivers

Subtask 3.1.2 developed ML models to enhance TC activity detection and prediction on the medium range. Such a goal requires defining candidate drivers that refer to local or remote conditions and can be related to atmospheric or oceanic variables. In particular, local drivers can include variables such as relative humidity, vorticity, and SST. On the other hand, large-scale candidate drivers include tropical, equator-tied waves such as the convectively coupled equatorial waves (CCEWs; Frank and Roundy 2006, Matsuno 1966, Kiladis et al. 2009, Schreck et al. 2012, Maier-Gerber et al. 2021, Lawton et al. 2022, Schreck et al. 2011), empirical wave-like phenomena such as the Madden-Julian Oscillation (MJO, Klotzbach, 2014), African Easterly waves and, even though often rejected (e.g., Leroy and Wheeler 2008, Henderson and Maloney 2013), the Quasi-Biennial Oscillation (QBO, Gray 1984). Other effects that have a role in TC genesis can be Rossby wave breaking (RWB, Zhang et al. 2016, 2017, Wang et al. 2020) or the baroclinic influence associated with the presence of an upper-level trough. The former may be represented by a predictor that describes the extratropical influence (for example, upper-level layer-averaged PV), and the latter could be considered using the coupling index (CI; Bosart and Lackmann 1995). In addition, the Q-vector convergence (Q) and the lower-level thickness asymmetry (Th), used by McTaggart-Cowan et al. (2008, 2013), can be included to describe baroclinically influenced development pathways of TC genesis. Finally, oceanic drivers present slow varying features, where possible candidates are local SSTs and teleconnections (e.g., ENSO; Gray 1979, Gray 1984, Song et al., 2022). A more extensive discussion of the list of candidate drivers for this problem formulation can be found in deliverable D3.1.

Target variable

The target variable is derived from the IBTrACS dataset by evaluating, at every grid point on a 2.5°x2.5° grid separately, whether at least one TC occurred within a 48 h period and radius of 300 km, to be consistent with the definition of TC activity used at ECMWF. The evaluation

of occurrence is based on the original 3-hourly temporal resolution and only considered cyclones that reached tropical storm intensity (i.e., ≥17 m/s). Given the extreme nature of TCs, the ratio of grid points at which TCs are active and non-active should not be too imbalanced for ML models to be able to learn a meaningful relationship between predictors and the target variable. The regional focus in this study is on the SIO, where TCs occur over ocean grid points at a rather low mean relative frequency of 0.56% (Figure 3.1a) but are subject to a distinct annual variability (Figure 3.1b). As shown in many applications, such a ratio in the target variable should still be sufficient to train meaningful models.

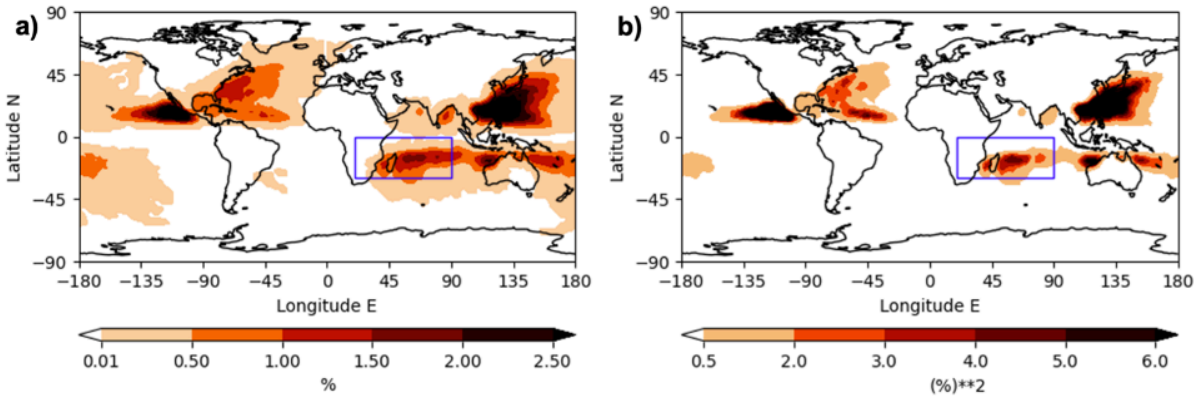


Figure 3.1*: (*a) Mean and (b) variance of the relative frequency of TC occurrence (%) calculated for 1980-2015, which is used as training period. Note that interval boundaries are not equidistant. The blue box encloses the area in the Southern Indian Ocean, for which the ML models are trained.

3.1.3 Skill and performance of existing forecasts

Since the target variable considered here is binary, all benchmarks and ML-based models are defined to output probability values, which convey more information than if the models predicted just the binary labels directly. However, this implies that the verification of forecasts is more complex. To evaluate the predictive model performance on the test set, different tools are combined to address various verification aspects. Receiver operating characteristic (ROC) curves, which display the true positive rate as a function of the false positive rate, allow the assessment of the potential predictive ability of a given model, with the best (no) skill indicated by an area under the curve (AUC) of 1 (0.5). Because ROC curves are insensitive to miscalibration, it is possible to obtain good performance even when the distribution of the forecasts is statistically inconsistent with the observations. Therefore, we also include the Brier score (BS), which averages the quadratic error over all forecasts, and thus considers the calibration aspect in the evaluation. This means that the expected score can only be minimised by predicting the underlying observed distribution. Following Dimitriadis et al. (2021), we further decomposed the BS into three additive measures: MCB represents the forecast miscalibration, DSC assesses the ability of the (re)calibrated forecasts to better discriminate between the outcome of the target compared to the performance of a climatology-based forecast, and UNC expresses the uncertainty inherent in the forecasting

problem. The BS, MCB, and UNC are negatively oriented (i.e., lower is better), whereas the DSC is positively oriented (i.e., higher is better).

In the past, TC activity forecasts were heavily based on dynamical models in the medium range, while statistical models were mostly used for the seasonal predictions. Since Subtask 3.1.2 targets the medium range, ECMWF's ensemble predictions (ENS) serve as a first benchmark. The ensemble system consists of 50 perturbed ensemble members and one unperturbed control member, all having a horizontal resolution of 18 km up to 15 days ahead of the considered training and testing periods. In each ensemble member, TCs were tracked (see Magnusson et al. (2021) for tracker description), including cases of genesis during the forecast. Based on the TC tracks in each ensemble member, a gridded field of probability of TC activity is calculated in the same manner as for the target variable.

A second type of benchmark is generated from TC activity statistics over the training period from the climatology of the target variable, referred to as the climatological model (CLIM). The simplest approach to generate a climatological forecast would have been to average over the entire training period (as shown in Figure 3.1a). However, from the variance signal in Figure 3.1b and previous studies (e.g., Maier-Gerber et al., 2021), it seems advantageous to calculate climatological probabilities separately for each day of the year to reflect seasonal variations. A 30-day window is then applied to the day-of-year dimension to smooth out discontinuities resulting from undersampling issues. Because these statistics are calculated over a set of past realizations drawn from the observational distribution, climatological forecasts are inherently independent of the current state of the atmosphere, unbiased if trends and/or regime changes are negligible, and thus, independent of lead time. The resulting BS for CLIM is constant, which makes it a good choice as a reference for any skill score, as it allows an easy comparison of predictive skill of models across lead times.

The years 1980-2015 serve as the period from which the climatological forecast probability is derived and on which ML models are trained. All models are evaluated on the period 15 April 2016 to 31 December 2022 in terms of various aspects of their forecast performance. All grid points in the SIO region are pooled for verification so that the conclusions drawn are more robust. Grid points over land are not considered, as their inclusion would have further worsened the existing imbalance in the target dataset.

The dynamical forecasts turn out to perform better than the climatological forecasts by more than 40% in BS at 0-1 days lead time (Figure 3.2a). With increasing lead time, skill decreases continuously and drops below the climatological reference beyond day 9. The decomposition of the BS reveals that the DSC and MCB terms for CLIM are of the same order as the UNC term, but almost cancel each other out, so that the BS is slightly lower (i.e., better) than the UNC (Figure 3.2b). In contrast, the ENS model results to be well calibrated overall, except over the first two lead days, but it exhibits a good discriminative ability that yields good BS values over the first week mentioned before. The generally low UNC term results from the high imbalance of the target variables, which means that even the trivial approach of always

predicting a zero probability (referred to as *ZEROS*) does not perform much worse than the CLIM reference model.

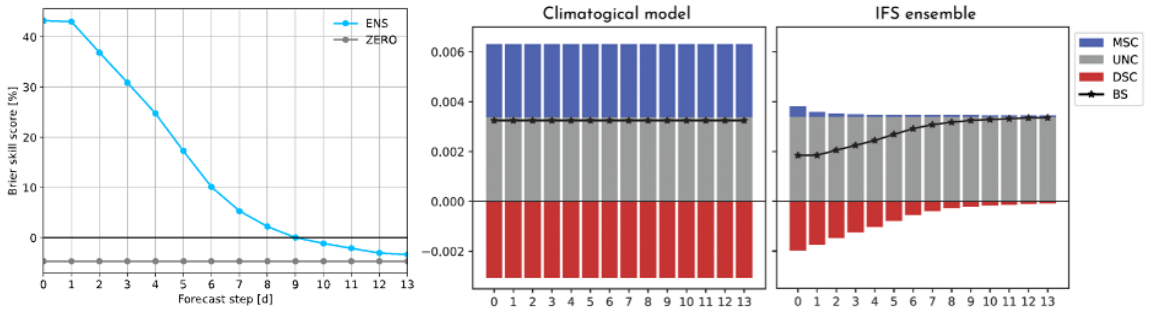


***Figure 3.2**:* (a) Brier skill score (BSS) (in %) of TC activity probability with respect to the climatological model as function of lead time. (b) BS decomposition into uncertainty (grey), miscalibration (blue), and discrimination (red) for the two benchmark models. Resulting BSs are displayed by the black asterisks.

3.1.4 Developed algorithms

The following three subsections present an overview of the ML algorithms that are developed as part of the baseline models, a convolutional neural network-based approach, and a hybrid model approach, respectively. For the first two categories, separate models are trained for each time lag (0, 1, …, 13 days), including in the features the values of the candidate drivers lagged by the selected number of days. In contrast, pre-trained lag=0 models are used for the hybrid model approach.

Baseline models

The first set of algorithms considered to perform this task are the classical methods designed for tabular data: *logistic regression* (Kleinbaum et al., 2002), *AdaBoost* (Freund and Schapire, 1997), and *extremely randomised trees* (Geurts et al., 2006). Subsequently, different FFNN (Schmidhuber 2015) architectures are considered. These approaches focus on different aspects and are designed to address different issues in ML. Indeed, these tabular approaches do not consider spatial and temporal patterns. On the contrary, they consider all samples to be independent and identically distributed, i.e., they assume that all samples are drawn from the same joint distribution.

The main advantage of these models is that they train a single model with all the data available for the region under analysis, with many samples and a reduced number of features, thereby mitigating the risk of overfitting. Their disadvantage is that they do not consider the spatial and temporal relationships among data, making them informed baselines for testing more advanced ML approaches.

Convolutional neural network-based approach

CNNs (LeCun et al., 1998) are ML methods designed to deal with image data with the aim of exploiting the spatial location of pixels in an image; this technique is promising for TC activity prediction because the spatial distribution of the meteorological features can be exploited to extract meaningful patterns. In this context, each feature is considered as a channel of an input image, and the target can be considered as a black-and-white image, where each pixel assumes a value between 0 and 1, representing the probability of TC occurrence**.**

The CNN architecture designed for this forecasting problem follows the structure of autoencoders (Baldi, 2012), with an encoder structure that extracts meaningful features in a latent space and a subsequent decoder part that reconstructs an image from the latent space, minimising the reconstruction error with respect to the target image. Given the relatively small number of training images (11,323), the number of layers and nodes is designed such that the parameter number did not exceed the order of magnitude of the number of samples.

A U-Net (Ronneberger et al., 2015) is also considered to compare the relatively simple structure of an autoencoder-based CNN with a more complex state-of-the-art CNN-based architecture specifically designed for image segmentation. An example of the U-Net architecture is shown in Figure 3.3.

These convolutional-based approaches are trained considering binary cross-entropy as loss function, allowing to tune the number of iterations and topology of the networks. Comparing the CNN approach with the U-Net approach in these terms, the U-Net shows a slight improvement, with the cost of a much larger number of parameters. Therefore, both architectures are considered similarly meaningful.

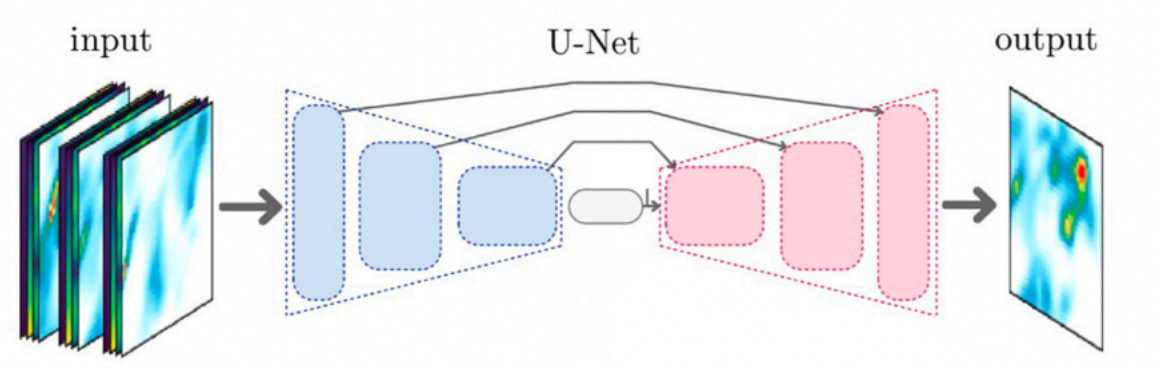


Figure 3.3*:* Example of U-Net, a state-of-the-art convolutional-based architecture considered for the task. Each feature is a channel of the input image and the output represents the spatial probability of occurrence of a TC. Credits for the image: (Serifi et al. 2021).

Hybrid model approach

For the best-performing purely data-driven ML models, a hybrid forecasting approach is tested, feeding the ML models with predictors extracted from NWP forecasts. The basic idea behind it is to get the best out of both approaches, i.e., combining the skill inherent in dynamical forecasts with the power of statistical or ML methods. This approach has already been demonstrated to enhance skill in forecasting weekly TC activity on sub-seasonal timescales (Maier-Gerber et al. 2021).

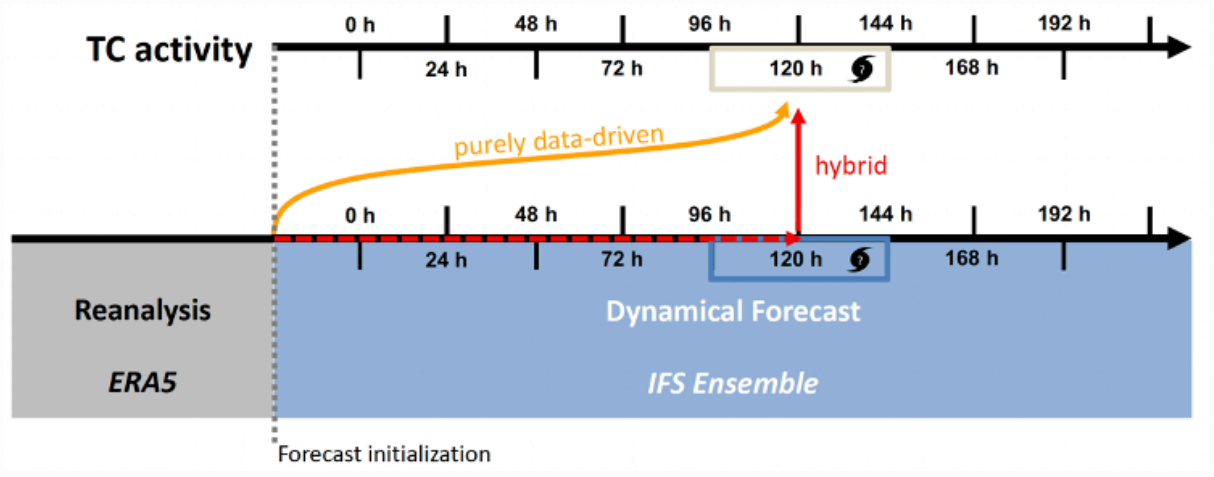


***Figure 3.4**:* Schematic illustrating the purely data-driven (orange arrow) and hybrid (combination of red arrows) modelling approach for the example of predicting TC activity at 120h lead time.

A fundamental issue that must be considered when implementing such a hybrid approach is the extent to which the lead time step should be modelled by the dynamical and ML model components, respectively. Assuming the goal is to predict TC activity with a lead time of, say, 120h, Figure 3.4 illustrates the two possible, and totally opposite, extreme cases. In the purely data-driven (non-hybrid) approach, a ML model pre-trained with lag=5d is applied to ERA5 data of the time when the forecast is made. In the most extreme form of the hybrid approach, predictors would be taken from the NWP forecast of the lead time to be forecast (i.e., at 120h) and fed into a pre-trained model trained for lag=0d. Between these extremes, it is of course possible to combine any lead time at which the predictors are taken from the NWP forecast with the remaining lead time that needs to be modelled by the ML model to cover the entire lead time considered. In a first step, however, we implement the hybrid model in the most extreme form to evaluate the largest contrast. ECMWF's ensemble control forecast served as the underlying NWP model component.

3.1.5 Results: AI-Enhanced forecasts and relevant drivers

As shown in Figure 3.5, the baseline models clearly perform worse than the dynamical model and are found to have the following descending order in BSS: *FFNN* performs best, followed by *extremely randomised trees*, *logistic regression*, and *AdaBoost*. Note that the BSS of the latter is so low that it is not shown for the sake of readability.

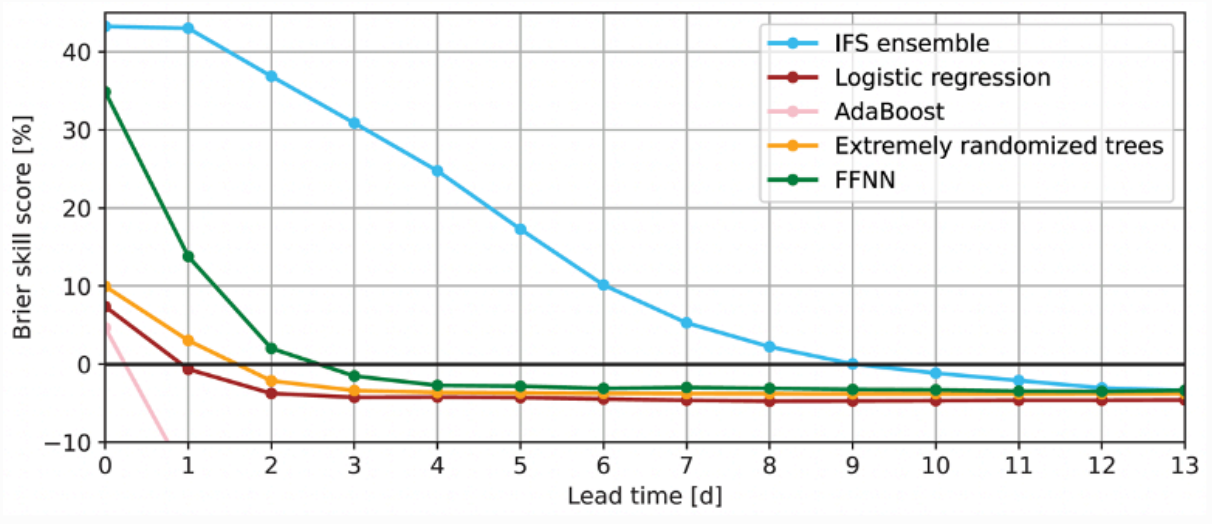


***Figure 3.5**: Same as in Figure 3.2a but including the baseline ML models.*

Figure 3.6 presents the BSS over lead time for the best-performing versions of the CNN-based ML methods plus the hybrid approach, comparing them with the dynamical ensemble benchmark and the FFNN baseline model. The performance of the LSTM model is similar to the FFNN, which, however, trailed behind the IFS ensemble by about 13 % and 8 % at day 0, respectively. The best CNN model performed best at day 0 but then lost skill quickly and dropped below the climatological skill after day 0, together with LSTM and FFNN. The overall best-performing purely data-driven approach is the U-Net, which also slightly exceeded the dynamical model skill, but then retained its skill a bit longer, crossing the zero-skill climatological line on day 5. Given its outstanding performance relative to the other ML models, the U-Net is the main candidate for testing and verifying the hybrid model approach. The hybrid version of the U-Net clearly had less skill than the original U-Net at lag 0 but started a little better than the FFNN. Exploiting the value of the NWP-based component, it is able to retain its skill much longer, thus extending the predictive skill towards longer lead times, eventually dropping below climatology between lead day 5-6.

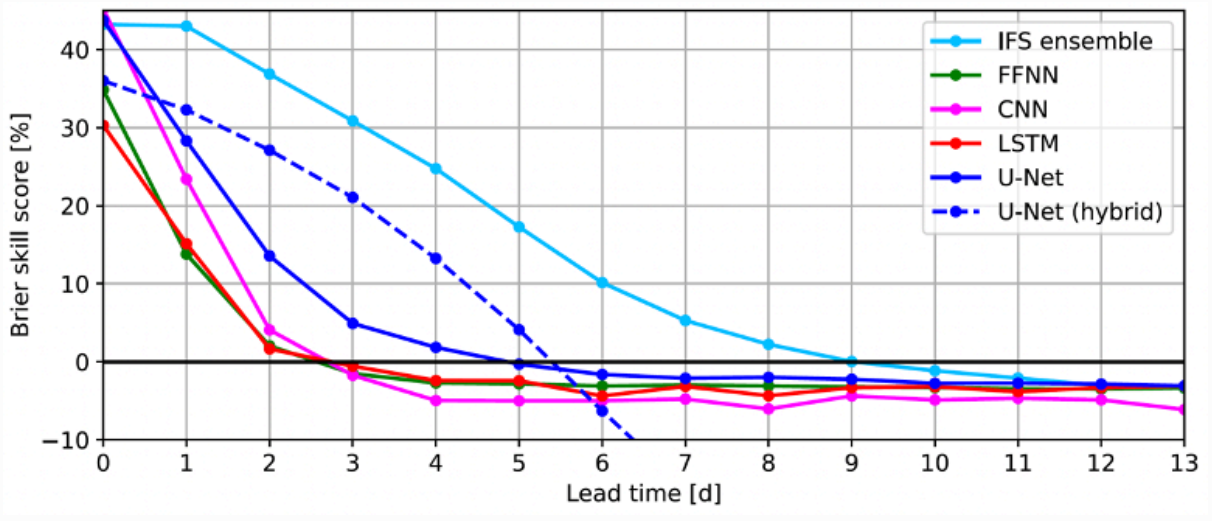


***Figure 3.6**:* Same as in Figure 3.2a but including the CNN-based ML models (solid lines) and the hybrid model approach (dashed line).

It should be noted that, in its current implementation, no bias correction has been applied to the control forecast, meaning that predictors are taken from the raw forecasts. Any bias statistics would have to be calculated over a period different to the one used for verification, i.e. for the time before April 2016. But because retrieving forecasts of this dataset are usually slow, it poses a strong practical limitation, and we refrained from downloading multiple years of data.

A decomposition of the BS for the two versions of the U-Net model indicates that forecasts if the original version are usually well calibrated, independent of lead time. In contrast, the hybrid version also starts well calibrated but then becomes increasingly miscalibrated, effectively degrading BS mainly beyond day 5. On a positive note, it can be seen that, apart from the miscalibration which could be corrected by post-processing, the hybrid U-Net model better discriminates between TC activity and non-activity for lead times longer than 0, corroborating the great potential of the hybrid approach (Figure 3.7).

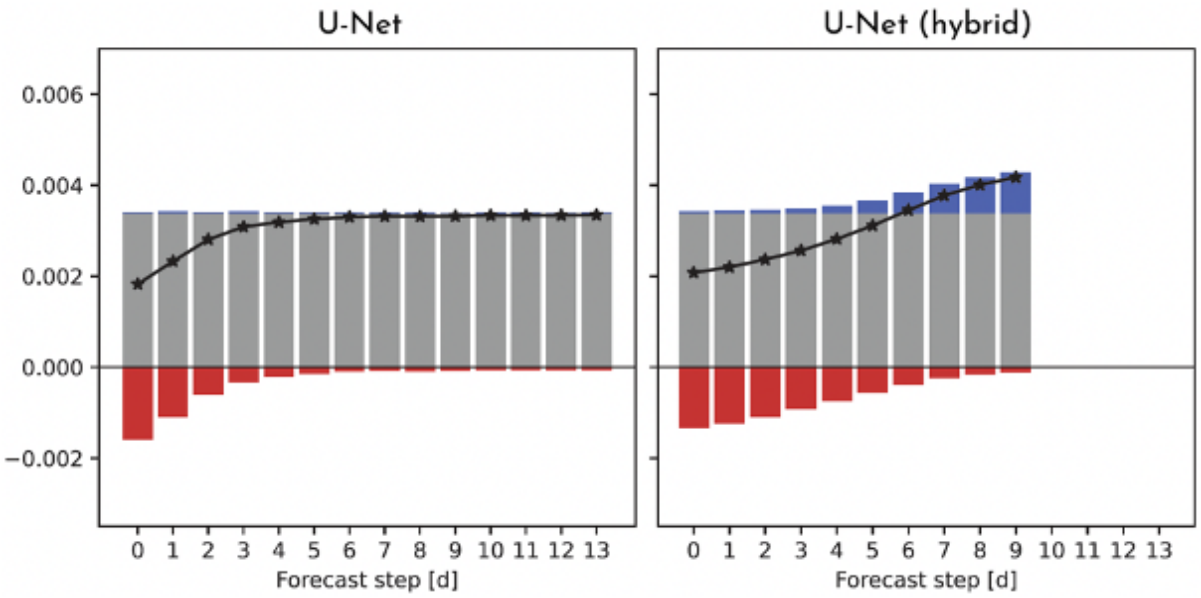


Figure 3.7: Same as in Figure 3.2b but for the original (left) and hybrid (right) versions of the best-performing U-Net.

A major part of the work carried out in this subtask of WP3 is dedicated to the development and improvement of model architectures and feature sets (see Deliverable D2.2 for the details of these tests). Therefore, we had frequent iterations between the identification and engineering of relevant features for TC activity prediction on the one hand and the quantitative assessment of various forecast aspects on the other. The main lessons learned from this process are summarised in Table 3.1 in a qualitative manner, indicating to what extent a change had been detrimental, neutral or beneficial.

Low values of outgoing longwave radiation (OLR) in the tropics hint at deep convection, which is why this variable is often used as a proxy. In an experiment, we test whether OLR could be replaced by the total column water vapour variable, as the latter, in contrast to the former, is an instantaneous (i.e. it refers to a specific point in time) parameter and hence simplifies preprocessing in any real-time application. As BSs are very similar (not shown), the change had a neutral effect but brought the aforementioned advantage. A neutral effect is also found, each in a separate experiment, for the inclusion of climatological probability, geographical information (latitude and longitude) and temporal information (year and day of year) as additional predictors, as well as for the addition of predictors from previous days.

A degradation in skill scores is found when experimenting with oversampling techniques to combat the imbalance in the target dataset and when training on global input fields. We also test to guide the learning of the ML model by the provision of condensed information about the large-scale flow and ocean state, represented by SST and geopotential-based climate indices. We tested it by including them in two ways, as separate channels together with the other input fields but also by ingesting them into the latent space after the encoder part of the model architecture. All these tests, however, degrade scores significantly. Another

experiment is to provide the input fields at a higher resolution (1°x1° instead of 2.5°x2.5°) but this also turned out to be unsuccessful as the higher resolved data did not provide enough new information to justify the considerable increase in the number of trainable parameters that resulted from the adapted architecture.

Clear improvements are achieved by expanding the predictor set using real-time observations (i.e., previous targets) and predictions for the previous day(s), highlighting the strong persistence component in this forecasting problem. Although operationally preferred, owing to the reduced data volume and lower pre-processing costs, the test to replace the daily averaged (thus, non-instantaneous) predictor data with only the 00-UTC values results in a non-desirable reduction in predictive skill and therefore is not being pursued further. This means that the representation of the intraday variability of the predictor data is important. To increase sample size and to allow for a better generalisation of the model, we added six other sub-basins, where TCs occur, to the SIO region, also resulting in slightly enhanced skill.

***Table 3.1**:* Qualitative summary of results of a number of the feature engineering and selection tests indicating to what extent a change has been beneficial (green plus symbols), neutral (grey circle) or detrimental (red minus symbols).

Replacing OLR by total column water vapour	o
Importance of climatological probability	o
Importance of global input fields	– –
Importance of features from previous day(s)	o
Importance of predictions from previous day(s)	++
Importance of intraday variability in predictor data	+++
Importance of geographical information	o
Importance of real-time observations	++
Importance of temporal information	o
Combining geographical and real-time information	+
Treatment of class imbalance	– – –
Increasing the number of areas for training	+
Including climate indices (e.g., NAO, PNA)	– –
Providing input data at higher resolution	– –

3.1.6 Summary and Outlook

In Subtask 3.1.2, ML models are developed for TC activity prediction on the medium range and compared against predictive skill of dynamical ensemble forecasts and climatological probability predictions. Various model architectures are trained, ranging from simple baseline models, such as Logistic Regression and AdaBoost, to more advanced CNN-based models, such as LSTM and U-Net. Trained on an extensive predictor pool, which combines well-known influencing factors from the literature, the U-Net model proves to be most useful for detecting and predicting TC activity, as in many other applications. This is certainly

due to the advantageous characteristics of its architecture, namely its ability to exploit spatial correlations in input fields through convolutions, and to compress and reconstruct data through the encoder-decoder architecture, together with the possibility to pass information through skip connections. The hybrid version of this ML model type further increases skill out to day five by extracting predictors from the ECMWF ensemble control forecast.

In conclusion, it can be stated that a broad set of features together with the testing of different model architectures, as well as their optimisation, and finally the realisation of a hybrid approach leads to successive improvements of the prediction skill in TC activity forecasting. We deliberately decide not to include predictions of the target variable from the IFS benchmark model to avoid turning it into a post-processing problem and also to be able to better analyse the importance of other influencing variables. Further improvements in skill are expected from the following work in progress. In the coming months, we would like to do more tests with the hybrid approach, for example, to use the full NWP ensemble, to remove biases from the NWP forecasts, or to optimise the split between the lead time covered by the NWP and the ML model predictions, respectively. In terms of predictor types, we would like to tap a source of predictability we do not consider explicitly yet, namely adding information from tropical waves, which are known to modulate TC activity. Apart from enhancing our own work, we plan to generate the target variable from ensemble predictions made by some of the large-scale deep learning models (e.g., ECMWF's AIFS model, Google DeepMind's GraphCast model).

The best-performing non-hybrid U-Net model has been deployed in close collaboration with WP8 as a WPS prototype on the Levante server at DKRZ. The prototype app is called Shearwater, a tribute to the bird species that combines two of the main predictors in its name and has been found to occasionally fly into the eye of TCs to rest in the calm winds.

3.2 Extratropical Transition

Here, we summarise the work conducted and the results found in Subtask 3.1.3 of the CLINT project. It concerns the topic of TCs undergoing extratropical transition (ET), a process that makes TCs affecting also the midlatitudes. The goal is to improve the prediction of ET through the use of AI by taking three different approaches to formulate and address the problem.

3.2.1 Overview

At the end of their life cycle, some TCs curve away from the tropics and start to interact with the waveguide in the mid-latitudes. Extratropical Transitions (ETs) can have a substantial impact in the mid-latitudes, both if the cyclones directly (Evans et al., 2017; Baker et al., 2021) hit a sub-tropical stage, soon after ET (e.g., Sandy 2012, Leslie 2018, Lorenzo 2019) or indirectly (Keller et al., 2019) as the ETs can lead to downstream development (e.g., after TC

Karl, 2016; Schäfler et al., 2018). Even if cyclones do not hit land, ocean waves can propagate over long distances and hit the coasts of Europe.

Whether or not a TC approaches the extratropics is primarily determined by steering flow. If a TC is close to a bifurcation point in the flow (Riemer and Jones, 2014), large track forecast uncertainties and track errors can occur. Therefore, it is critical to correctly predict bifurcations in the steering flow and TC track towards these points. Magnusson et al. (2014) discussed an example of such sensitivity for TC Sandy (2012) and Magnusson et al. (2019) for TC Joaquin (2015), where small changes in the subtropical ridge caused very large differences in the future tracks of these TCs.

A related uncertainty is phasing with the mid-latitude wave guide, where an upstream trough favours northward propagation into the extratropics. Correctly predicting the mid-latitude waveguide is crucial for capturing ETs. This sensitivity was highlighted by McNally et al. (2014), who found that satellite data over the northern Pacific influenced the predictions of landfall of TC Sandy.

While TCs that undergo ET may create substantial impacts downstream over Europe, the majority of TCs do not undergo ET. As was especially evident in 2020, several TCs could make landfall in the deep tropics or subtropics, spinning down quickly into a remnant low-pressure system. Other TCs weaken as they encounter high vertical wind shear or substantial low-humidity air, which may occur in the tropics, especially in the extratropics. As a TC moves into the extratropics, it encounters much colder waters, removing the supply of thermal energy and moisture from the ocean, which is necessary to maintain the TC.

In CLINT, we approach the ET problem in three different ways:

1. Two-dimensional fields of TC activity in the northern part of the Atlantic can be predicted based on a set of predictors (either two-dimensional fields or indices). The solution to this prediction problem is similar to that described in Chapter 1.

2. Given that TC genesis is observed, the likelihood of it reaching high latitudes can be determined based on a set of predictors.

3. Given that a TC reaches high latitudes, its impact in terms of weather extremes in Europe can be estimated (seasonal hindcast and climate projections).

In this section, we focus on (1) and (2), as (3) will be reported later in CLINT WP7. In this section, we focus on the Atlantic basin, but depending on the degree of generalisation, the ML methods are potentially transferable to other ocean basins in which TCs undergo ET (e.g., Northwest Pacific and Southern Indian Ocean).

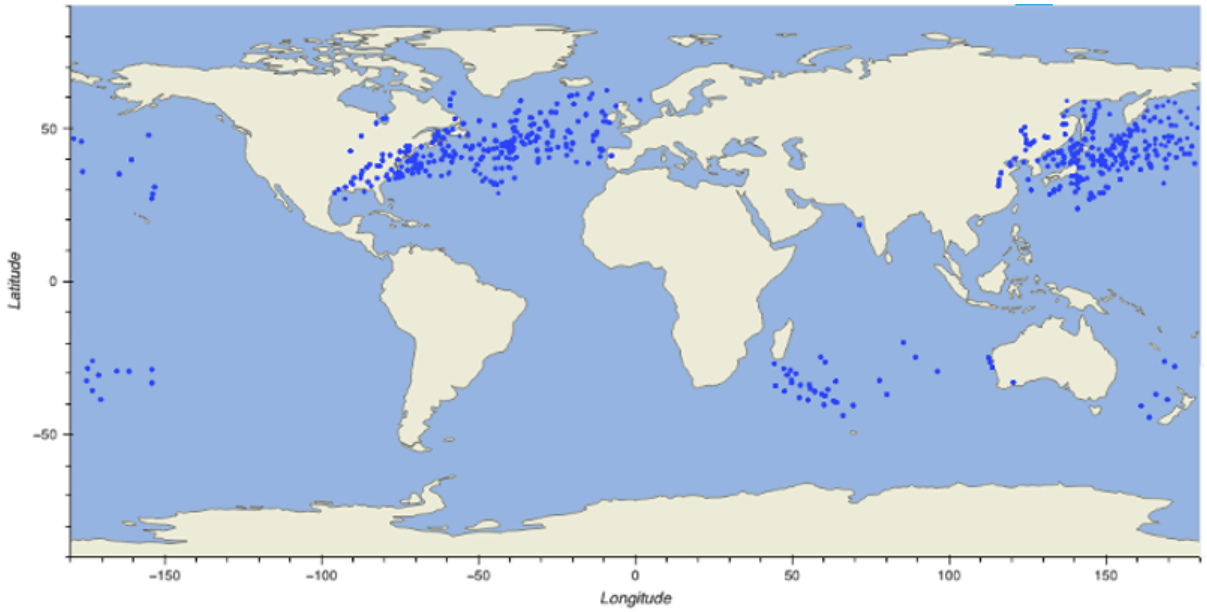


***Figure 3.8**:* IBTrACS positions cyclones in the extratropical stage for April 2016-December 2022.

3.2.2 Datasets, candidate drivers and target variable

In this subsection, we describe the preprocessing of target values and predictors for the problem formulation "Given a TC genesis, what is the risk for it to reach high latitudes".

This method required:

1. Definition of genesis instance.
2. Definition of criterion to count if the target region is reached.
3. Definition of predictors at the genesis time.

The data periods are defined as 1980-2015 for training and validation and 2016–October 2021 for the test period.

For the observation dataset for TCs, the alternatives are either based on estimations from the National Hurricanes Center (IBTrACS dataset) or from TCs tracked in reanalyses (e.g., ERA5; Magnusson et al., 2021). The advantage of IBTrACS is that the estimates are based on the best knowledge obtained from observations and human judgements. The disadvantage is the possible inconsistencies in time due to changes in practice. For ERA5, TCs are automatically tracked in atmospheric reanalysis ERA5 (Hersbach et al., 2020). The advantage here is the consistency of the method across time. One disadvantage is that the TC maximum wind is known to be underestimated by the reanalysis due to limited resolution and observation coverage. There could also be inconsistencies due to variations in the observation coverage during the reanalysis period.

In this report, we focus on the use of ERA5 as the observation dataset to keep the treatment of the ETs constant. Following common practice, we defined the TC genesis point as the first instance when the TC reached 17 m/s maximum wind speed at 10 m height. We define the target region to be north of 40N and between 98 W-0W, which agrees well with reported ET, as shown in Figure 3.8. If a TC at some point during its track passed inside that box, it counted as a true event (i.e., a TC that underwent ET). While the target variable is binary, all models produced a probability that indicated the likelihood that a given cyclone will undergo ET.

For ERA5, the total number of TCs and the number that reached the target region in the training and test periods are given in Table 3.2, together with the fraction. As can be seen from these numbers, the proposed train-test split preserves the fraction of ET cases in both subsets.

***Table 3.2**:* Total and train-test-split number statistics of all TCs and TCs reaching extratropical stage (in North Atlantic target region).

	Total	**Train**	**Test**
Total	481	390	88
Reaching target	182	147	34
Fraction	37.8 %	37.7 %	38.6 %

The predictors are based on the TC properties at the genesis (position, intensity, day of the year, etc.) and climate indices from the CLINT-TS dataset (see D3.1). Examples of climate indices include SST averages, such as the Nino3.4 index and SST in the Tropical Atlantic ("main development region for TC"), and Euro-Atlantic weather regimes based on 500hPa geopotential height. There is an option to add a temporal filter to the indices beforehand.

To benchmark the ML-based methods, we use two fundamentally distinct forecasting approaches. First, ECMWF ensemble forecasts (ENS) are used to compare the data-driven ML model with a physical model. Between March 2016 and July 2023, the ENS has a horizontal resolution of 18 km, but undergo several upgrades of the model and data assimilation. Based on automatic tracking (same as for ERA5 above), we examine the forecasts at the genesis time and count the number of ensemble members (50 in total) that featured a TC in the target box during the next 15 days. The fraction of the ensemble that fulfils the criteria determines the forecast probability. Second, the fraction of TCs undergoing ET in the training dataset (37.7 %) can be considered as a climatological forecast (CLIM), assuming the training sample represents the underlying distribution of the target variable and that there are no trends.

3.2.3 Skill and performance of existing forecasts

Because the target variable is binary and forecasts are probability values, for the ET forecasting problem we use the same verification tools employed for the short-term TC activity predictions. Therefore, we refer to the third paragraph of section 2.4.3.

A high ET fraction of 38.6% led to an equally high UNC of 0.237 (Figure 3.9), still close to BS=0.25, which is the value that a random forecasting approach would have resulted without any prior (e.g., climatological) knowledge. This demonstrates the large uncertainty associated with the forecasting problem. The CLIM model, being constant, has no discriminating ability; at the same time, it is by definition well calibrated, since its forecast probability is calculated from the underlying distribution of the target variable. In contrast, the ENS predictions exhibit considerable miscalibration, but their predictive ability to distinguish between ET and no-ET cases offset this by more than a factor of two, reducing the BS to 0.180.

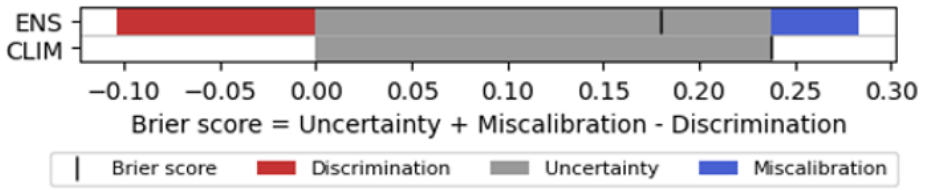


***Figure 3.9**:* Brier score (BS) decomposition into uncertainty (grey), miscalibration (blue), and discrimination (red) for the two benchmark models. Resulting BSs are denoted by the vertical black lines.

3.2.4 Developed Algorithms

Decision trees and random forests

A decision tree builds a sequential chain of conditions that would favour one of the outcomes. We used the “DecisionTreeClassifier” and “RandomForestClassifier” models from the scikit-learn package in Python. The only degree of freedom we explored is the depth of the tree (number of conditions). The choice of depth is a balance between stratifying the sample to get the most out of the training data and the risk of overfitting.

Logistic regression

For binary target variables, logistic regression models (Hastie et al., 2009) are a commonly chosen type of model that maps linear combinations of continuous predictor variables to a probability via a logit function. Regression coefficients are estimated by minimising a cost function maximum based on two terms: one corresponding to maximum likelihood estimation and the other applying an l2-regularisation, which keeps the coefficients of the predictors small and thus helps to prevent the model from overfitting.

For the logistic regression model, a forward sequential predictor selection scheme is applied to the predictor pool to determine a subset that maximises the predictive skill over the training dataset. In each step, the predictor is added from the remaining pool, for which a predefined score is optimised in a 5-fold cross-validation applied to the training data. We test several scores and finally chose the AIC over the frequently used negative log likelihood, since it penalises the model for including too many features and prevents overfitting. In a context of training data scarcity this allows a better generalisation.

3.2.5 Results: AI-Enhanced forecasts and relevant drivers

The results of the ROC analysis show that the LOG model is the data-driven approach with the best potential predictive ability, followed by random forests (Figure 3.10a). While CLIM by definition follows the diagonal, indicating no skill, the use of a single decision tree does not perform much better. A similar ranking is also obtained when miscalibration is considered (Figure 3.10b). All ML-based forecasts obtain a BS higher (i.e., worse) than the ENS, but lower (excluding the decision tree) than the CLIM. As revealed by BS decomposition, the LOG and random forests are better calibrated than the ENS but are much less able to discriminate between ET and no-ET. The fact that random forests are usually superior to decision trees (owing to their ability to reduce overfitting without massively increasing bias-related errors) can be seen by the highly reduced miscalibration and enhanced discrimination. However, the best BS among all data-driven models is achieved by the LOG model.

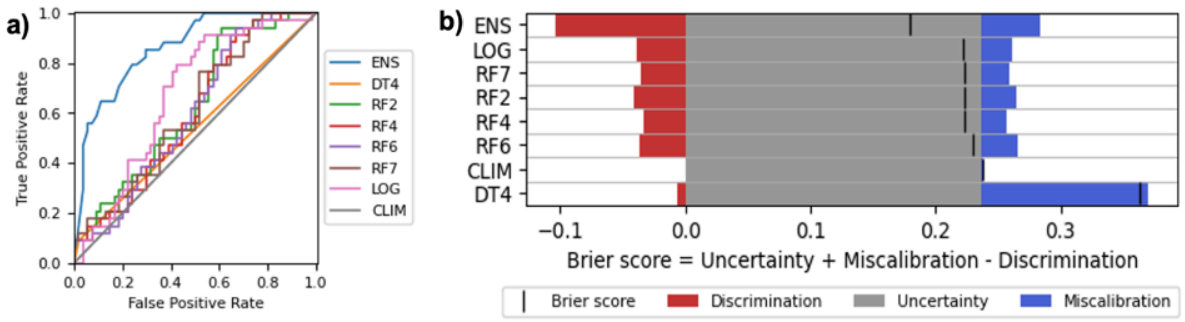


***Figure 3.10**:* (a) ROC curves for all models with AUC scores in the legend. (b) As in Figure 3.9, but including the results for the ML models sorted by BS.

From the statistics of the predictor selection process (Figure 3.11), conclusions can be drawn regarding the optimal number of predictors needed. This should be large enough to provide the model with the necessary predictive signals, but also small enough not to unnecessarily increase multicollinearity between predictors. Our choice to use the AIC for scoring results in only the latitude and longitude positions of TC genesis being included in the optimal subset (red curve). Employing the negative log loss, optimization would have been reached including the radius of maximum wind and standard deviations of anomalies of Nino3.4 and NAO indices (black curve). However, the small improvements gained with their addition are

a sign of overfitting, which the negative log loss is prone to. Using the BIC would lead to a decrease in the number of selected features, so that the optimum would already be reached with the latitude of the genesis predictor. Given that the dynamical model still clearly outperformed the ML-based models, despite the larger miscalibration, and the generally low number of predictors being selected, the final predictor pool seems to still lack relevant predictors.

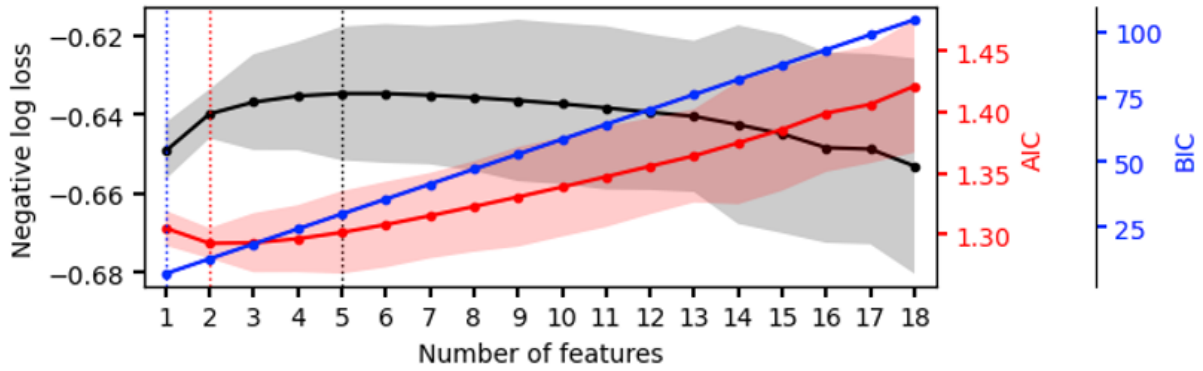


***Figure 3.11**:* Results of the sequential predictor selection applied to the logistic regression. Mean (line) and standard deviation (shading) of the negative log loss, the AIC, and the BIC as a function of the number of features. The dotted vertical line marks the optimal number of features identified for the corresponding score.

3.2.6 Summary and Outlook

In the exploration of ML to predict whether a TC would reach high latitudes based on the properties at genesis time, we find it difficult to improve the ECMWF ensemble forecast, despite a positive frequency bias in the ensemble. The strongest influence is found to be related to the latitude and longitude of the genesis, which is reasonable particularly if a cyclone already forms at high latitudes. We find that mid-latitude flow and SST indices at the genesis time had a small influence on the chances of the TC reaching the target region. So far, the mid-latitude climate indices considered are based on principal component analysis, which yields variance-maximising but rigid flow patterns. Since the prediction of ET is often subject to subtle local deviations from these large-scale patterns (e.g., phasing with trough and bifurcation points), we will develop and test more tailored indices.

However, there are many degrees of freedom to explore this prediction problem, such as the choice of the model and model settings, index selection, and index smoothing. As a next step, best practices and method applications coming out from work on the other extreme events reported in this deliverable will be solicited to be tested on the ET prediction for TCs.

Another future direction is to use the prediction from a dynamical forecast as an input predictor, which can either be a probability from the ENS or a binary result from a deterministic forecast (e.g., ERA5-based). However, this limits the sample size, which is already low.

4 HEATWAVES AND WARM NIGHTS

As a result of the wide range of impacts of heatwaves (HWs), from excess human mortality to agricultural losses and abrupt changes in energy demand (Thomas et al. 2020; García-Martínez et al. 2021; Zuo et al., 2015), there are a growing number of indicators for extreme heat events, such as warm nights (WNs - the equivalent for night-time temperatures) and health-related indices (Fischer et al., 2013; Perkins-Kirkpatrick et al. 2015; Davis et al., 2016). Detecting extreme temperatures and identifying their drivers are crucial to the development of prevention plans and mitigation strategies that can minimise the risks associated with all types of heat extremes (e.g. Lowe et al., 2016).

The skill of forecast systems, from short-term to seasonal, in detecting HWs and warm nights (WNs) has already been tested (e.g. Prodhomme et al., 2022; Torralba et al., 2024). Early warnings provided by the current generation of operational seasonal forecast systems remain inhibited by poor representation of European summertime conditions, such as the representation of jet stream flows and persistence of weather patterns such as blocking (Domeisen et al., 2023). As a consequence of limited reliability of dynamical models, efforts in recent years have turned to exploiting the power of Machine Learning methods to extract information on HW/WN drivers from observations/reanalysis. Such methods attempt to reduce the dimensionality, and therefore the computational expense, of the forecasting problem by using area-averaged time series (e.g. Zhang et al., 2022) or modes of variability as predictors (e.g. Kämäräinen et al., 2019).

The work done in Task 3.2 has used traditional and novel ML methods to define heat extremes, select definitions and indicators with most relevance to impacts, identify variables and indices which determine HW occurrence and propensity, and to create data-driven and hybrid approaches to seasonal forecasts. We explore modifications of HW indices to take into account human sensitivities, such as humidity and ability to acclimatise (Section 4.1); an ML-based detection of most impactful HW indices on agricultural crop yield (Section 4.2). Then, we describe the applications of the feature selection framework developed earlier in the project (Section 4.3). This flexible framework, depending on its set-up, can be used to identify predictors of HW indices (Section 4.4.1), and make forecasts (Sections 4.4.2 and 4.4.3). In Section 4.4.2, we present a purely-data driven approach to forecast temperature extremes based on climate-model training. In Section 4.4.3, a hybrid version of the framework, which incorporates existing dynamical forecasts, is described. Finally, we include a brief discussion on the differences between day and night heatwaves in Section 4.5.

4.1 Indices and Datasets

The characterization of extreme events is usually performed by employing specific indices which describe the severity, magnitude and duration of the events. A comprehensive index should allow a comparison of the event to other episodes or to typical conditions. As evidenced in the literature and Deliverable 3.1, a wide range of indices exist, each with their pros and cons.

All indices here are built on the commonly used percentile-based method of defining HWs (e.g Russo et al., 2015). A HW is typically defined as a temperature exceedance over the daily 90th-percentile, with many studies also imposing a minimum duration of 3 days. All indices can be applied to any daily temperature time series; for example, Tmax is used for the most-commonly used definition of heatwave, Tmin to define nighttime heatwaves, and average apparent (humidity-dependent) temperature at night used to define "warm nights" (D3.1). We refer to all heat extremes generally as HW, until Section 4.5 when we provide discussion on differences between temperatures from different parts of the daily cycle. Beyond daily indicators of HW occurrence, we employ the Heat Wave Magnitude Index (HWMI) and the number of days above the 90th percentile in a given season (NDQ90).

Indicators can describe various time scales, from precise daily information (e.g. HW occurrence) to seasonally aggregated information (e.g. number of HW days in a month). In addition, the Excess Heat Factor (EHF) is used to consider a HW metric that is related to human health impacts (Nairn and Fawcett, 2014). It combines two indices describing two different temporal characteristics: the significant deviation from the long-term mean and the short-term anomaly.

The ERA5 reanalysis is used as the ground-truth against which HW indices in forecasts are validated. Its hourly resolution permits the study of both day and night extremes with a spatial resolution of approximately 30 km. This reanalysis provides a large set of variables that are used as potential drivers for the detection of EE (see D3.1 and D3.2).

The "past2k" simulation is a reconstruction of the atmosphere and ocean climate over the past two millennia. It has been performed with the MPI-ESM1.2-LR model, using ECHAM6.3 as its atmospheric component (2^{o} horizontal resolution with 47 vertical levels) and the MPIOM1.63 as its ocean component (1.5, reaching 30-40km in sub-polar North Atlantic, with 40 vertical levels). A detailed description of the MPI-ESM model and the past2k simulation can be found in Jungclaus et al. (2014 & 2017). The original reconstruction of years 0-1850 has since been complemented by an extension period from 1851-2014. The model is forced by reconstructions of greenhouse gases in the atmosphere, land-cover, volcanic aerosols, solar forcings (with artificial 11 year cycle) and monthly average ozone concentrations (Jungclaus et al., 2014 & 2017).

The current capabilities of the seasonal prediction systems to predict HWs have been explored in the multi-system framework provided by the Copernicus Climate Change Service (C3S) initiative. The ability of the C3S seasonal forecast systems to detect those EE at seasonal timescales will be used as a benchmark to quantify the potential added value of the ML methods developed in Task 3.2. In this deliverable, comparisons are made to the ECMWF (European Centre for Medium-Range Weather Forecasts) SEAS5 system, as it provides a longer continuous reforecast period until 2022. Previously (D3.1 & 3.2), other C3S seasonal forecasts produced by different institutions were employed, namely DWD (Deutscher Wetterdienst), Météo-France and CMCC (Centro Euro-Mediterraneo sui Cambiamenti Climatici. The main specifications of these prediction systems are listed in D3.1, but all the systems provide 6-hourly fields spanning six months into the future, with a

spatial resolution of 1° and global coverage. The number of ensemble members (i.e. the different realisations used to sample the seasonal forecast uncertainty) varies among the different seasonal forecast systems.

4.2 ML-based identification of HW indicators for agriculture

In this section, the impacts of temperature extremes, including both HWs and WNs, on agricultural productivity in the Lake Como Climate Change Hotspot are studied to support the planning of adaptation strategies in view of mid-to-long-term climate change projections (in WP7). Specifically, we first simulate historical crop yields using a detailed, process-based model of the agricultural districts. Then, we use correlation analysis and the Patient Rule Induction Method to identify the most relevant drivers of crop failure among different indices describing the occurrence and intensity of HW and drought events. This method, as well as an analysis of projected change of indices and crop yield in CMIP6 models, will be published in Giuliani et al (in preparation).

4.2.1 Impact Model

The impact model simulating the dynamic processes in the irrigation districts served by the Lake Como releases is the IdrAgra model. The model is composed of three distinct modules devoted to specific tasks:

1. a distributed-parameter water balance module that simulates water sources, conveyance, distribution, and soil–crop water balance, including the application of irrigation (Facchi et al., 2004);
2. a crop phenology module that computes the sequence of growth stages as a function of the temperature according to the Heat Units theory (Neitsch et al., 2011); and
3. a crop yield module that estimates the optimal and actual yields, accounting for the effects of stresses due to insufficient water supply that may have occurred during the agricultural season (Steduto et al., 2009).

The water balance module partitions the irrigation district with a regular mesh of cells with a side length of 250 m (i.e., each cell covers an area of 6.25 hectares), which allows for the representation of the space variability of crops, soil types, meteorological inputs, and irrigation distribution. The study area consists of 32,820 grid cells, for a total cultivated area that amounts to 205,125 hectares.

We used meteorological data from the ERA5 hourly reanalysis extracted for the box 46.5° North, 10.9° East, 44.5° South, 8.65° West. Specifically, the impact model described in the previous section requires the following inputs:

- Daily minimum and maximum temperature (Tmin and Tmax).
- Total precipitation.
- Daily minimum and maximum relative humidity.
- Daily average wind speed (derived from the V and U components).

- Daily total solar radiation.

Daily time series of observed levels, releases, and net inflows of Lake Como are available from 1946 (the start of the lake regulation after the dam construction) to 2022. The release data are here used to simulate the irrigation supply to the considered agricultural.

To capture the occurrence of HWs and WNs, we consider seasonal indices from April to September which correspond to the agricultural season of the main crops cultivated in the Po valley. These include the HWMI and NDQ90, alongside the daily heatwave occurrence and intensity series, and the series of occurrences of temperature above the 90th percentile. The reference period used to compute the indices is 1981- 2010.

The Standardized Streamflow Index (SSI) is also considered to verify the possible influences of drought conditions on crop production. The index is computed from the monthly data of Lake Como inflows cumulated over a six-month period.

In total, we considered 51 indices, which fall into the following categories:

- HWMI calculated with maximum (tmax) and minimum temperature (tmin);
- NDQ90 (tmax and tmin);
- Number of HW occurrences over the agricultural season (April to September) (tmax and tmin);
- Sum of HW intensity over the season (tmax and tmin);
- Number of HW occurrences in the individual months over the season (tmax and tmin);
- Sum of HW intensity in the individual months over the season (tmax and tmin);
- NDQ90 in the individual months over the season (tmax and tmin);
- Number of drought months each year;
- Average SSI during the year;
- Average drought intensity each year;
- September SSI, aggregated over 3, 6, 9 and 12 months.

4.2.2 Patient Rule Induction Method

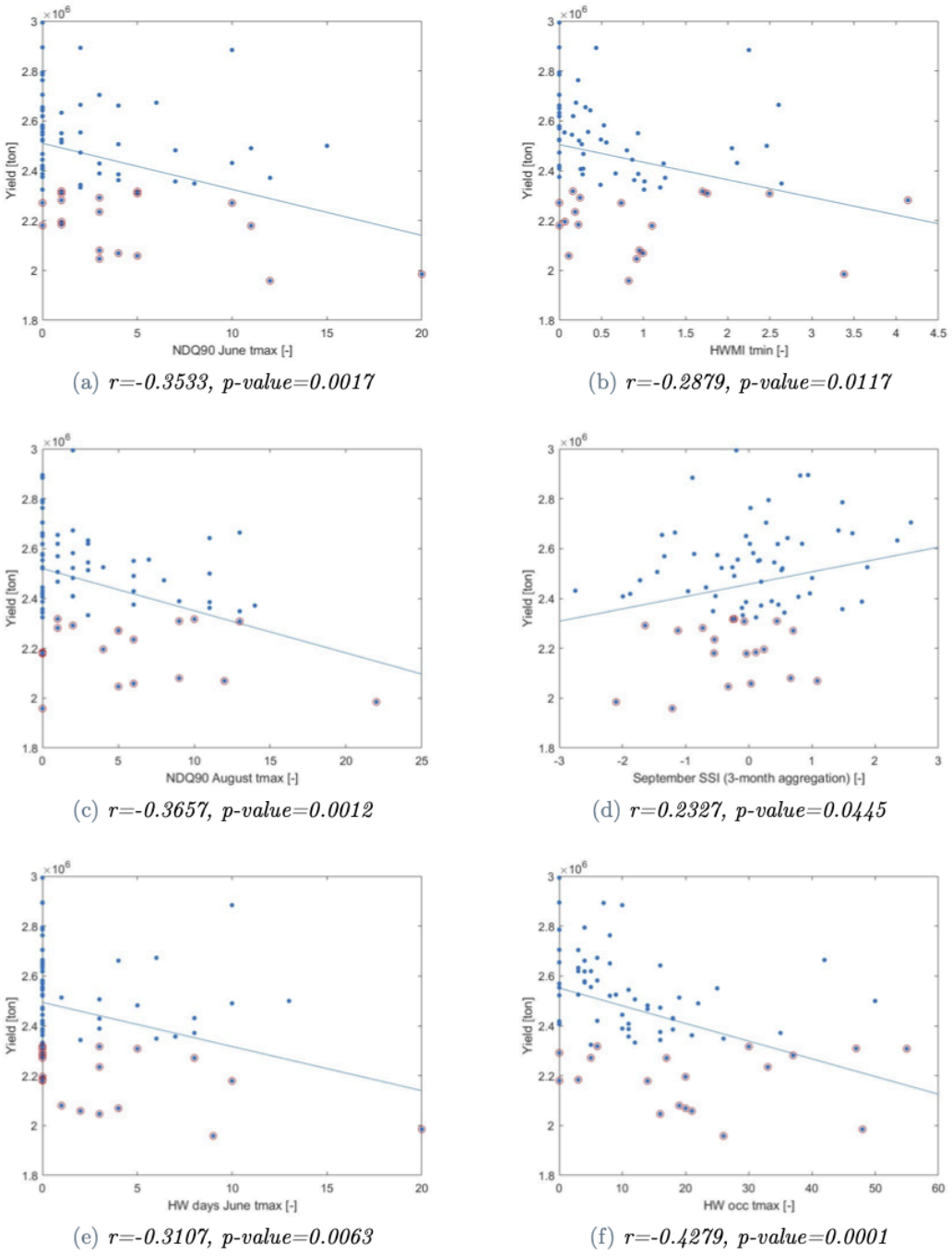
The Patient Rule Induction Method (PRIM) is used as an aid in the analysis of the relationship between crop production and heatwave-drought indices to complement the findings obtained from a simple correlation analysis between HW indices and simulated crop production. PRIM is a statistical clustering method originally introduced by Friedman and Fisher (1999) that is often used in scenario discovery analysis (Kasprzyk et al., 2013; Giuliani et al., 2022). It belongs to a group of algorithms called "bump-hunting" algorithms, which are used to find regions, called scenario boxes, in the input variable space that are associated with the highest or lowest mean value for the outcome (Nannings et al., 2008). Boxes correspond to a simple square in a two-dimensional input variable space and to a hypercube in a multi-dimensional space. This is unlike regression models, which seek to model the whole population by optimizing a likelihood function. In this work, the input

variable space is composed of the HW and drought indices, whereas the outcome is the yearly yield. More specifically, what is of interest is the identification of the indices associated with the low yields, which are labeled as crop failures. Therefore, a threshold has to be set on the outcome variable so that the algorithm can distinguish between failures and non-failures when building the scenario boxes.
The scenario boxes are constructed by optimizing several competing metrics, namely coverage - how many failure scenarios are captured within a box - and density - how many of the captured scenarios in each box belong to the failure set. Ideally, a scenario box should have both a high coverage and density. This guarantees that the inputs used to build a box are able to explain the highest number of failure points possible and that the noise generated by uninteresting points is minimum, which happens when the density is high.

4.2.3 Results

To understand the link between yearly yield and extreme temperatures, and potentially droughts, a correlation analysis is conducted and 27 out of the 51 correlations evaluated proved to be statistically significant. As an example, the scatter plots in Figure 4.1 show the correlations for the six most relevant indices (as determined by the factor mapping analysis described below). As can be deduced from these scatter plots, there are obvious trends but, in all instances, it is difficult to partition the variable space in a way that fully separates the failure points from all the others. For this reason, we built on this preliminary correlation analysis and ran a formal factor mapping using the PRIM method described in the previous section. The aim of factor mapping is not to perfectly classify the data but to facilitate the interpretation of the results and the identification of the most relevant drivers of the crop failures. The candidate drivers considered in this analysis are the 27 statistically significant indices, while the output variable is the crop yield labelled as failure/safe outcome using the 25th percentile threshold.

Figure 4.2 shows the PRIM results as a Pareto front of scenario boxes navigating the trade-off between coverage (x-axis) and density (y-axis). Ideally, a good scenario box would be positioned in the upper right corner, where both metrics have their highest values. This case, however, does not exist because of the conflict between the two metrics. Moreover, each point in the figure is assigned a colour based on the number of inputs used to build the corresponding scenario box, ranging from zero inputs (dark blue) to eight variables (yellow). The downside of considering too many dimensions is that the coverage values decrease drastically, meaning that the drivers used to build the box are not good at explaining a high enough percentage of all the failure yields.



(a) r=-0.3533, p-value=0.0017 (b) r=-0.2879, p-value=0.0117

(c) r=-0.3657, p-value=0.0012 (d) r=0.2327, p-value=0.0445

(e) r=-0.3107, p-value=0.0063 (f) r=-0.4279, p-value=0.0001

Figure 4.1: *Scatter plot of the most relevant heatwave/drought indices and crop yield. Each point in the figure represents a value of simulated yield (y-axis) associated with the corresponding value of the heatwave/drought index (x-axis). The blue line corresponds to the least-squares regression line, while the red markers identify crop failures (i.e., yield values below the 25th percentile of the historical series). The values of Pearson linear correlation coefficient and corresponding p-values are reported below each panel. The correlation values are statistically significant if the p-value is smaller than 0.05.*

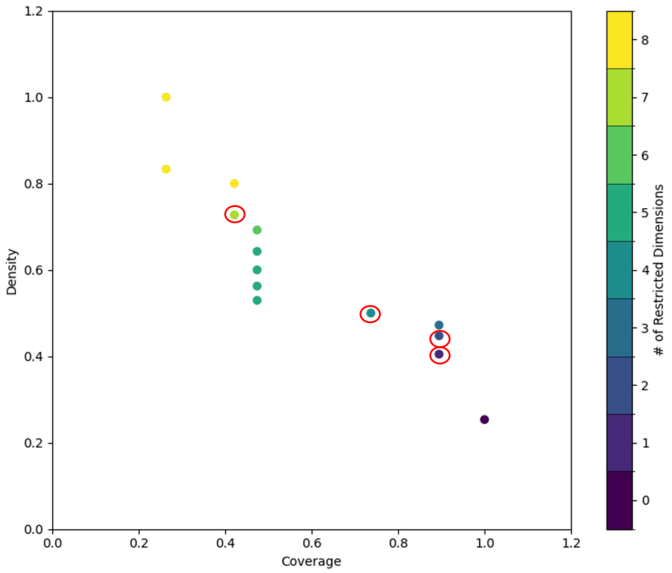


***Figure 4.2**: PRIM results in terms of scenario boxes navigating the trade-off between coverage (x-axis) and density (y-axis) of crop failure with respect to heatwaves/drought indices. Circled points are the scenario boxes analyzed in detail.*

From these results, we select four scenario boxes relying on an increasing number of inputs to be analyzed in more detail. The first two solutions are chosen to isolate the most relevant drivers of crop failure, supporting the definition of scenario boxes with high coverage (but low density); the 4-dimensional solution is chosen because it is characterised by a density equal to 0.5 (but coverage equal to 0.74); finally, the 6- dimension box is chosen because it further increases the density while simultaneously reducing the coverage only slightly below 0.5. The crop failure drivers selected by PRIM in the different solutions are the following:

- 1 dimension (89.5% coverage, 40.5% density): NDQ90 in June calculated with maximum temperature (NDQ90_occ_june_tmax),
- 2 dimensions (89.5% coverage, 44.7% density): NDQ90_occ_june_tmax and HWMI calculated with minimum temperature (HWMI_tmin),
- 4 dimensions (73.7% coverage, 50.0% density): NDQ90_occ_june_tmax, HWMI_tmin, NDQ90 in Au- gust calculated with maximum temperature (NDQ90_occ_aug_tmax) and the SSI aggregated over 3 months and calculated in September (SSI_sept_3_months),
- 6 dimensions (47.4% coverage, 69.2% density): NDQ90_occ_june_tmax, HWMI_tmin, NDQ90_occ_aug_tmax, SSI_sept_3_months, number of heatwave days in June calculated with maximum temperature (HW_days_june_tmax) and number of yearly heatwave days calculated with maximum temperature (HW_occ_tmax).

The PRIM results show that the most important drivers selected as responsible of crop failures in the Adda River basin are the NDQ90 in June (tmax) and the HWMI (tmin). The former, by definition, also includes days that are not necessarily part of a HW event,

meaning that extreme temperatures in general are detrimental to crop yield. The latter is particularly interesting because it shows the significant role that nighttime temperature extremes and WNs play in the agricultural sector. A drought index also appears among the HW indices, which means that water stress contributes to crop failure but not as much as temperature stress since it is selected only starting from the 4-dimensional solution.

These findings have also been used in Task T7.4 to investigate the projected evolution of these critical indices under climate change, in order to support local end users (farmers) in better understanding future risks and explore the opportunity to replace some of the crops currently cultivated in the area, primarily maize, in favour of heat-tolerant varieties, such as soy or cereals, in order to ensure more reliable productions in the coming years (for details, see Deliverable D7.2).

4.3 HW precursors: ML-defined weather regimes

Here, we assess atmospheric drivers related to HW indicators for the case study of Sweden over the past decades. The analysis is based on ERA5 reanalysis data for the period of 1940 to 2022 using the pattern analysis method of weather regimes (WRs). The selected HW indicators are the Excess Heat Factor (EHF) indicator which can be used for human health impacts and the NDQ90_occ, which represents the 90th percentile of maximum temperature. This study focuses on Stockholm city, where 985 HW events were detected from 1940 to 2022, May to August.

EHF is a measure of HW intensity related to human health impacts and is a product of two indices describing two different temporal characteristics: (i) the significant deviation from the long-term mean (Excess Heat Index significance - EHIsig) and (ii) the short-term anomaly (Excess Heat Index acclimatisation - EHIaccl). In EHIsig, the daily mean temperature averaged over three days (DMT, daily mean temperature as the average between the daily maximum and minimum temperature) is compared against the 95th percentile of DMT over a climatological reference period (here 1981-2010). The EHIaccl index is a measure of how hot a 3-day period of DMT is with respect to the previous 30 days. It follows the idea that human bodies are in general able to acclimatise to their local climate but have difficulties in adapting to sudden temperature rise (Nairn and Fawcett, 2014). The two indices are multiplied (EHF = EHIsig × max (1, EHIaccl)), resulting in a quadratic measure of HW intensity with a HW when EHF is positive. In order to describe HW across different regions compared to its local climatology, a normalisation is applied of each daily EHF intensity value divided by the 85th percentile of the climatology. This results in the EHF-severity index with different severity thresholds: Low- Heatwave intensity: daily EHFsev > 0 and <1; Severe Heatwave: daily EHFsev ≥ 1 and <3, and Extreme Heatwave: daily EHFsev ≥ 3 (Nairn and Fawcett, 2014).

Weather regimes (WRs) are classified based on the concept of fuzzy sets (Zadeh, 1965, Bárdossy et al., 2002), using imprecise statements to describe the climate system. The anomalies of daily mean 500hPa geopotential height (z500) from the reanalysis data (i.e., ERA5, ~ 0.5°), serve as a predictor to derive temperature-induced WRs. The daily anomalies

at each grid have been computed as daily deviations from the long-term climatology (here 1981-2010) over the Euro-Atlantic region (24.75°-74.75°N, 135°W-45°E). Daily mean temperature from 34 observation stations distributed in the whole Sweden serve as predictand to reflect the variability of local climate. They help optimise predefined fuzzy rules describing individual WR. Each of the classified WRs describes a recurrent and persistent atmospheric state.

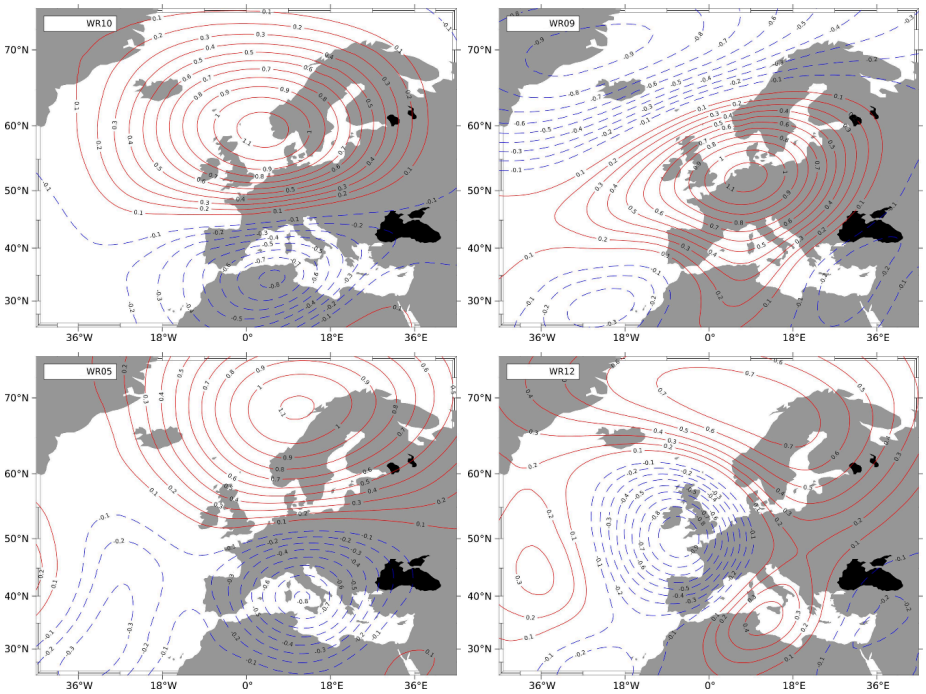


***Figure 4.3**: Composite maps of WR10, WR09, WR05 and WR12, where red isolines indicate higher-than-average pressure anomalies and blue isolines indicate lower-than-average pressure anomalies.*

A set of twelve temperature-induced weather regimes are classified (examples shown in Figure 4.3). Type 10 (WR10) is found highly related to the detected HW events, which explains nearly 40% of HW events detected from May to August during the period of 1940 to 2022 (Fig. 4.4 top), in particular, 47% of detected HW events in August. The overwhelming high-pressure system situated over the North Sea and the southern Scandinavian likely caused the warmer-than-average temperatures over northern and central Europe. The pressure distribution of Type 5 (WR05) is found similar to that of Type 10 (WR10), but its positive anomalies move further northward and causes warm weather over northern Europe. Type 9 (WR09) is featured with positive height anomalies prevailing over northern Germany. It favours anticyclonic dry and warm weather across the whole of Europe except Northern Scandinavia and Iceland. The weather regime of Type 12 (WR12) is shown with dominant deep anomalies covering the North Atlantic Ocean, while weaker positive anomalies over the European continent from the Mediterranean region to the Norwegian Sea, bringing warm air from northern Africa and the Mediterranean basin to the European

continent. In comparison to Type 10 (WR10), the linkage of Type 5 (WR05), Type 9 (WR09) and Type 12 (WR12) to the detected HW events are relatively weaker.

Also, it is found that if one among Type 5 (WP05), Type 9 (WR09), Type 10 (WR10) and Type 12 (WR12) occurs, on the same day a HW event is detected. A statistical test has been used to study and approve the significance of occurrence of a given WR to the detected HW events (see Figure 4.4), where the occurrence of these types of WRs on the day when the HW events (presented as Observed in Figure 4.4 left) occurred are much higher in comparison to their long-term average occurrence during the reference period (presented as Expected in Figure 4.4 left). Similar results are found for the EHF (Fig. 4.4 right).

Figure 4.5 presents the annual number of HW days provided by different HW indicators for a selected grid of ERA5 reanalysis data for Stockholm over the period 1940 to 2022. NDQ90_occ (occurrence of T > 90th percentile) shows a clear upward trend in the annual number of HW occurrences, higher than for HW_occ which has a requirement of 3-day persistence. The EHF severity index greater than 1, presenting severe and extreme HWs, shows a slight increasing evolution within the last decades with a strong HW summer in 2018, which is prominent in all three indicators presented here.

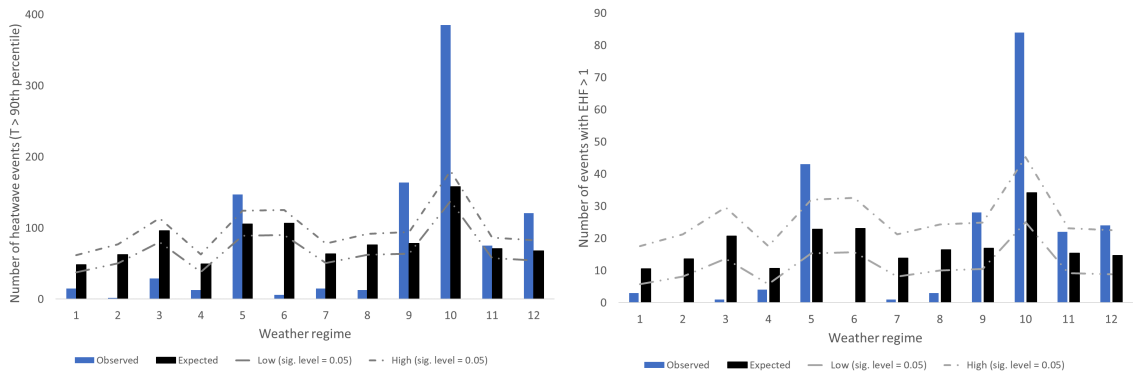


***Figure 4.4**: Histogram of extreme events (May to August, 1940-2022), defined using standard heatwave definition (left) and the EHF (right), with given WRs and corresponding tests of statistical significance (sig.level = 0.05) .*

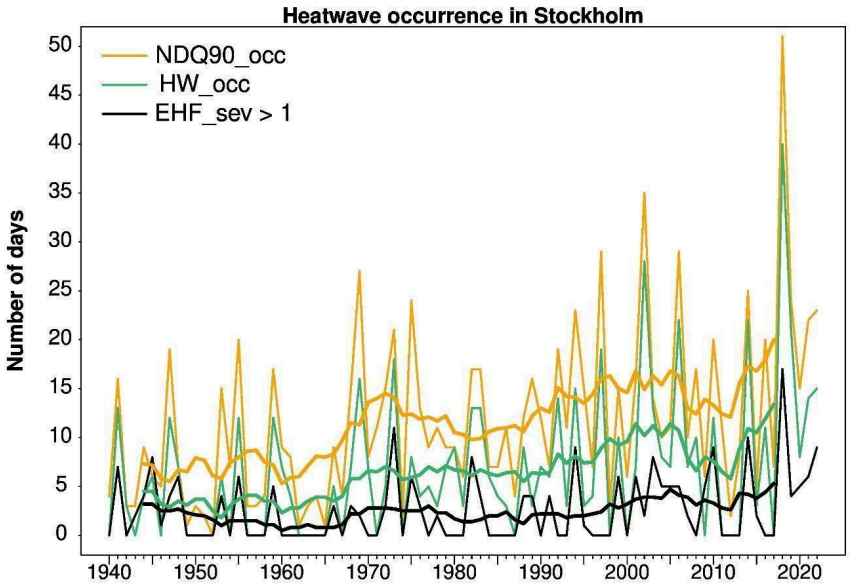


***Figure 4.5**: Annual number of heatwave days (May to August, 1940-2022) based on NDQ90, HW_occ and EHF severity > 1 (severe and extreme heatwaves) over the period 1940-2022 during the summer season (May to August) selected for a grid of ERA5 representing Stockholm. The bold lines represent the 10-year moving average.*

4.4 Optimisation-Based Feature Selection Framework: Driver Detection & Forecasting

Here, we describe a spatio-temporal feature selection framework developed for HW direct detection, and its applications to forecasting. The framework is composed of two steps: a dimensionality reduction of global variables, followed by a feature selection that identifies the optimal combination of drivers and lag times. The chosen variables are evidenced to potentially play roles in European HW occurrence, and represent atmospheric circulation (z500; Dong et al., 2018; Kornhuber et al., 2020; Kautz et al., 2022), ocean-atmospheric interactions (SST), precipitation (Stefanon et al., 2012), soil moisture (Materia et al., 2021), sea ice (Coumou et al., 2018) and more. Global clustering of specific variables (e.g. sea ice, outgoing longwave radiation) allows the potential identification of teleconnections. A k-means clustering is applied to each variable to extract 5 clusters per domain (Appendix Figures A1-A4); the daily area-averaged means are taken as predictors. In the second step, an evolutionary-type algorithm is employed to select the clusters of the different variables which provide the optimal detection skill for the target time series. The framework allows us to quantify the relative importance of each variable and cluster and, crucially, to identify the time lag from short-term to seasonal time scales. The framework will be described in Pérez-Aracil et al (in preparation).

In CLINT, we apply this framework in three varied ways in order to produce:

1. A feature selection of regional HW cluster predictors.
2. A hybrid forecasting approach.
3. A purely data-driven forecasting system.

4.4.1 Identification of Regional HW Cluster Predictors

The first application uses the framework in a detection mode. Potential predictor data from lag times up to 0 days are included, hence this method is more a "recreation" and not a forecast. The aim is to understand how well represented HW occurrences are with reduced dimensionality of their predictors in the framework. In initial experiments on local target data over the Lake Como region, the framework could recreate the HW record from 2011-2022 with a high accuracy (F1-scores of roughly 0.7; D2.2), thus demonstrating strong potential to be applied to other problems.

Here, we apply the framework to the time series of regional HW cluster occurrences. Clusters of temperature exceedance over the 90th percentile are calculated for the wider Europe domain using the Simulated Annealing and Diversified Randomisation (SANDRA) method with ERA5 data (as described in D3.2); as in the HW definition, a minimum duration of 3 days can be added as a post-processing step. Clusters represent common geographic occurrence patterns of extreme temperatures, and each displays varying extents and typical intensities (Fig. 4.6). For example, cluster 10 resembles the 2003 event over central-western Europe, while the 2010 blocking-related event over Russia falls into clusters 7 or 9. Daily time series of each cluster's occurrence is used as the target data for the feature selection. The training/validation period is 1951-2014, while the test period is 2015-2022. The imbalance of the datasets is shown by the contribution to the total variability of each cluster (Fig. 4.6); cluster 1 corresponds to no HW occurrence and thus is the most common. Cluster characteristics will be described by Hansen et al, (in preparation).

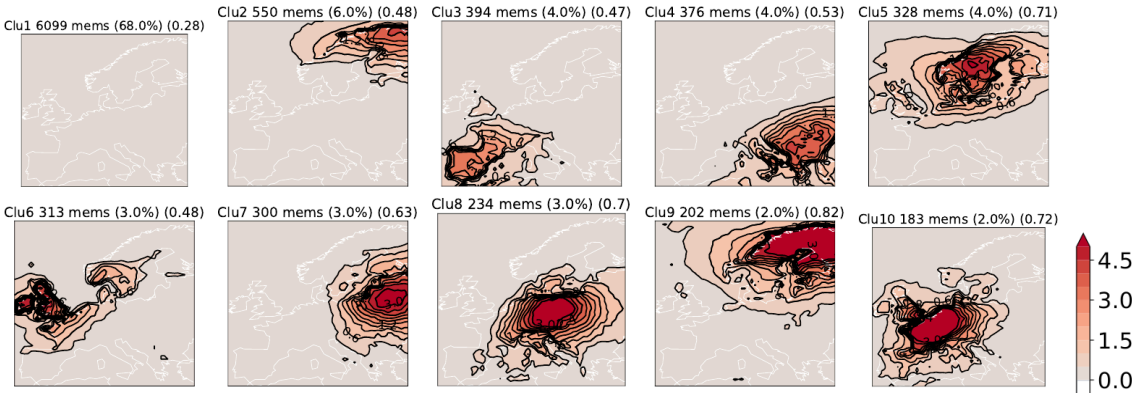


***Figure 4.6**: HW clusters over the European domain, coloured by their average intensity (contours correspond to 0.3 °C intervals). Clu1 corresponds to no HW. Variability (in parentheses) explained by each cluster is also a measure of dataset imbalance.*

The output of the feature selection is a large number of solutions, each using a different combination of predictors and lags. Studying the most frequently used clusters in the top 10% of solutions (ranked by the training F1-score) is a means of quantifying the optimal selected features. For cluster 6 (British Isles; Fig. 4.7), for example, soil moisture in Europe (sm1Eur 3 & 5) and regional z500 (z500Eur4) appear in nearly all the best solutions. Clusters, such as mean sea level pressure over Europe (mslEur2), are common but seemingly not strictly needed for the recreation of the target. Otherwise, the vast majority of clusters "chosen" (shown in Fig 4.7) represent noise; in D2.2, sensitivity analysis is shown to be useful as an extra step to filter out noisy contributions. The key selected features have short lag, nonetheless extending back to 20-30 days, implying that tendency, not just values from the previous day, are important. In the case of cluster 6, the relevance of sm1Eur5 extends from 45-75 days.

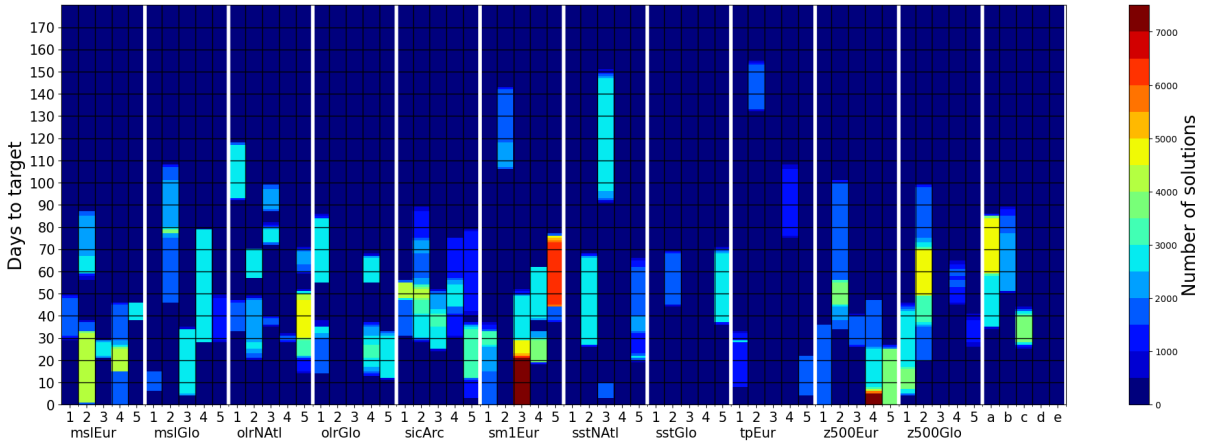


***Figure 4.7**: Example feature selection for Clu6 (British Isles). Maps of each cluster can be found in the appendix. Letters correspond to the following climate indices and dummy variables: a - ENSO, b - NAO, c - IOD, d - atmospheric CO2 concentration, e - day of year.*

***Table 4.1**: Quality of recreation of HW cluster occurrence (cross-validation F1-score) with optimal solution (0 lowest, 1 highest).*

F1-score - training/test				
Clu1 - No HW	**Clu2 - Scand_N**	**Clu3 - Iberia**	**Clu4 - Eur-SE**	**Clu5 - Scand-S**
-	0.50/0.32	0.47/0.35	0.50/0.32	0.5/0.38
Clu6 - British Isles	**Clu7 - Eur-E**	**Clu8 - Central**	**Clu9 - Scand**	**Clu10 - Eur-W**
0.5/0.36	0.43/0.32	0.28/0.22	0.44/0.32	0.39/0.27

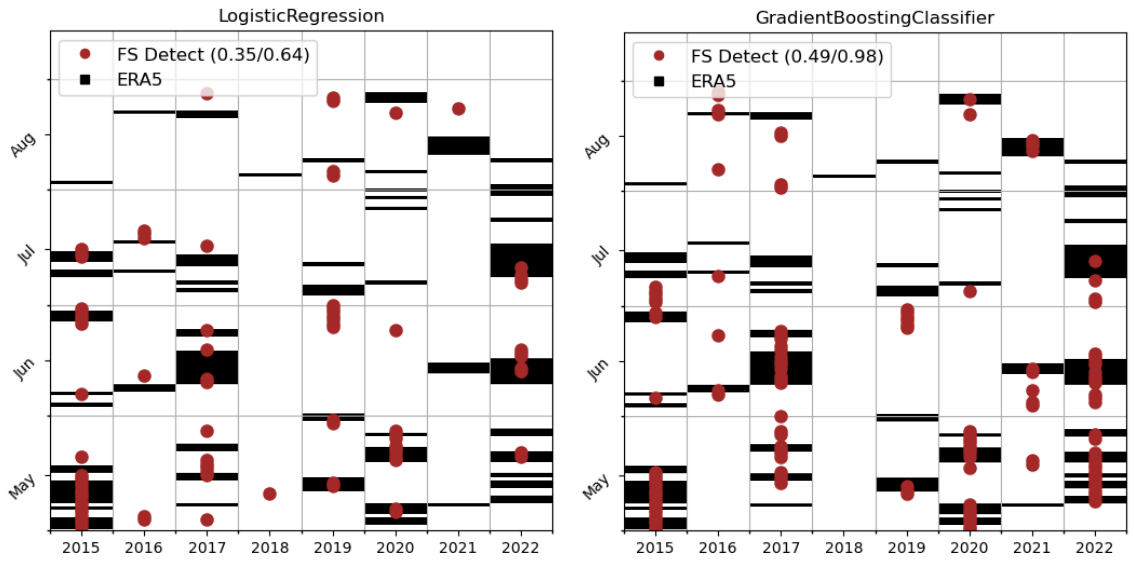


***Figure 4.8**: Recreation of HW cluster 6 indices from 2015-2022 (test period) from optimal features input into different models (Logistic Regression and Gradient Boosting Classifier). Values in the legend correspond to F1-score (left) and correlation of total summer days each year (right).*

The optimisation of predictors represents the best combination related to the potential predictors, which in this case has undergone an intense dimensionality reduction. Consequently, the reconstructed targets may be inaccurate. For example, for cluster 6 the F1-score of validation is 0.47. When applied to the test period, this drops to 0.35 (Fig. 4.8). A model that selects random values returns a score of 0.22, so the optimal solutions nonetheless provide some added value. The variability of detection skill strongly varies between clusters (Table 4.1). Certain clusters may be rarer and thus translate to a more imbalanced target (e.g. cluster 10 in Table 4.1), rendering detection a more difficult task. For others, their driving phenomena may be more masked by the dimensionality reduction. The cluster datasets are generally very imbalanced (i.e. typically less than 5% EE occurrence; Table 4.1), especially compared to local-scale HW indices (up to 8%; as tested in Deliverable 3.2).

The framework has been further modified to improve re-creation skills. First, more complex models are able to better leverage the same input information than, for example, the logistic regression used in optimisation. For example, inputting the optimal solutions into a Gradient Boosting Classifier increases the F1-score to 0.49. In both cases, the summer statistics (NDQ90) agree significantly with ERA5 (see correlations in Fig. 4.8). Moreover, here we purposefully remove clusters of European temperature in order to focus on less-obvious predictors of HWs. Previously (e.g. D3.2), we found local T2M clusters are commonly selected as a key feature on 0-20 days lag times, and that accuracy is higher. Including T2M clusters for the optimisation of cluster 6 returns F1-score values for training and test at 0.65 and 0.49, respectively.

Identification of predictors warrants further study into the physical mechanisms detected by the feature selection. For example, while previous studies have identified a role of soil moisture in European summer temperatures, here we identify potential region-dependent roles on different timescales. Alternatively, the framework can be used to quantify the relative importance of known drivers. As we will show in the following section, the predictors may also be exploited for forecasting purposes. These results can also be used to boost computational efficiency; rather than repeat the FS for each local point over a domain, clustering of events, a significantly more economical process than the optimisation framework, can be performed first. When interpreting selected features, however, care must be used. First, we are selecting from a predefined pool of potential predictors, and potential drivers may be missing. Secondly, identification of clusters which contribute to skill does not automatically lead to identification of key processes.

4.4.2 Hybrid Seasonal Forecasting

Dynamical seasonal forecasting systems have been validated for their skill in predicting seasonal propensity of HW occurrence (Prodhomme et al., 2023; Torralba et al., 2024). While in large areas of Europe skill has been shown to be significant, other areas remain difficult to predict (i.e. central to northern parts of the domain). In principle, dynamical forecasting systems should be more reliable when predicting larger-scale patterns and averages (for example the cluster averages shown here), than for predicting local scale HW indicators. In section 4.4.1, large-scale area averages of potential predictors (dimensionality-reduction), when input into certain ML models, were used to recreate the HW record to a high accuracy. Here, we attempt to leverage the skill of forecasts of cluster averages by inputting forecasts of drivers, as identified by a feature selection process similar to Section 4.4.1.

First, we modify the feature selection process described in Section 4.41; here, temperature clusters are included as potential predictors and local HW occurrence is the target data. The feature selection (optimisation) is applied to each point in a sub-domain covering central Europe, despite its size, displays a homogeneity of skill (Prodhomme et al., 2023). Each point, therefore, has specific optimal solutions for the detection of local HW occurrence. The optimisation is performed over a training period of 1951-2004 and a test period of 2005-2022. The next step transforms this approach into a hybrid forecast system; ERA5 data is used as predictor input up to the start time of the forecast (e.g. May 1st), and following dates are covered by dynamical forecast ensemble members (e.g. Figure 4.9). All predictor data are anomalies calculated with respect to dataset-specific climatologies (i.e. 1981-2010), so biases between datasets are removed. The ML forecasts are trained in 1951-2004, with 2005-2022 used as a testing period (as for the optimisation); this choice represents a compromise between allowing enough time for training and leaving a long enough period for validation. The number of predicted summer HW days is calculated for each ensemble member, and then averaged.

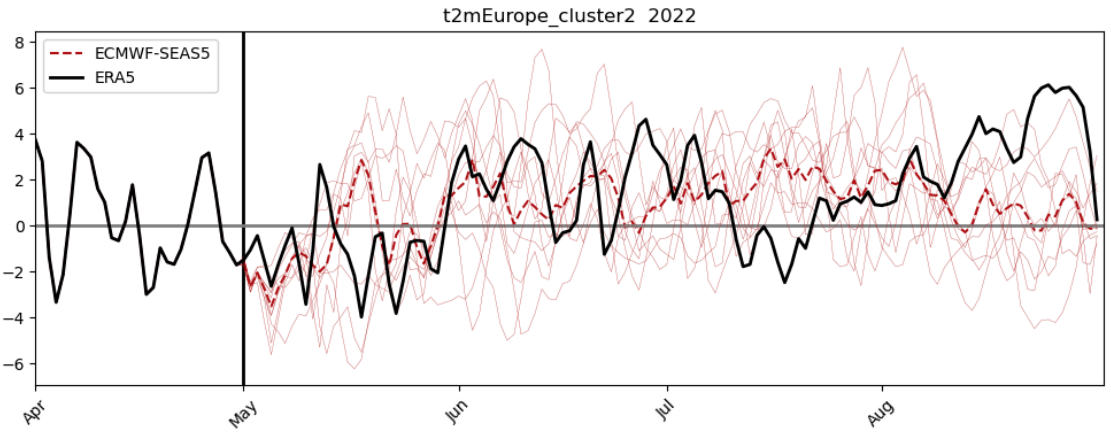


***Figure 4.9**: Time series of t2m temperature anomaly in Europe cluster 2 in ERA5 (black), ECMWF-SEAS5 ensemble members (red) and ensemble median (red dashed). Units in °C. In the hybrid framework, predictor data from May onwards (black vertical line) is taken from dynamical system (e.g. ECMWF-SEAS5) forecasts.*

The aim here, then, is to test whether forecasts of reduced-dimensionality drivers can provide a more accurate prediction of HW indices than the dynamical system directly. There are three possible outcomes of this analysis: (1) though HW occurrence is not directly well-predicted, the links between large-scale clusters and local HW dynamics captured by the FS framework allow a recreation via prediction of predictors in the dynamical system (e.g. AI-enhancement); (2) no enhancement is achieved because even the large-scale dynamics, shown to determine HW occurrence, are poorly predicted by the dynamical system or (3) no enhancement is achieved because the replication skill of HW indices using the optimal features is low. This method can serve as a (potential) AI-enhancement or an extra form of validation specific to the target data (i.e. extreme events).

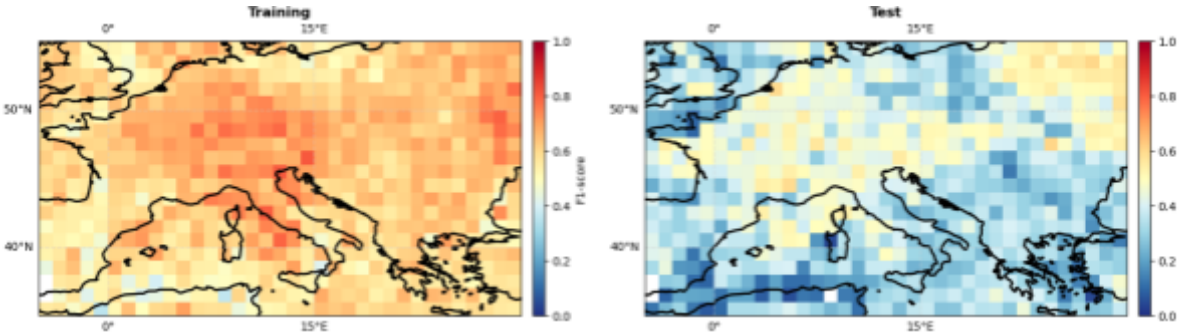


***Figure 4.10**: F1-score for HW occurrence over training period (1951-2004) cross-validation and test period (2004-2022).*

First, we assess how the ability to recreate daily HW indices within the framework varies across a reduced European domain (Fig 4.10, left). Overall, the training period validation is fairly accurate, with an average F1-score of 0.52 and maximum values of 0.7 are reached around 10E, 50N. HWs over the southern Mediterranean and North African coast are relatively less well replicated than those over central Europe. This analysis provides a first indication of where the selected predictors, once input into ML models, provide detection

skill. There is a drop in recreation skill when selected predictors are applied to the test period (Fig. 4.10, right). This in turn leads to a poor representation of interannual variability of seasonal indices; accurate detection by the framework is found only in a limited zone across western-to-eastern central Europe (Fig. 4.11, centre). The framework, in its current set-up, performs insufficiently across much of Europe; this is indicative of an overfitting to the training data, and possibly an overly imbalanced target dataset or insufficient information in the potential predictors. We remind here that the potential predictors are fewer than used in the previous section. The "detect" mode, using only ER5 data, provides an upper bound of what is capable with the hybrid model.

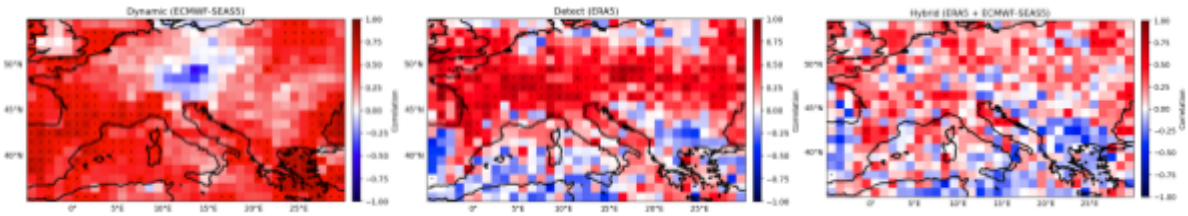


***Figure 4.11**: Correlation skill score of NDQ90 (May-July) in the period 2004-2022 between detection/forecast systems and ERA5. Detect uses only ERA5 predictors, while Hybrid replaces predictor data after May 1st with ECMWF-SEAS5 forecasts. Black stippling indicates statistical significance. For Dynamic and Hybrid, forecasts are initialised in May.*

The dynamical system (ECMWF-SEAS5) displays significant skill in NDQ90 forecasts across France and north-western Mediterranean, and a zone of poor skill (weak negative correlation) is present around 15E, 48N (Fig. 4.11, left). The hybrid system, on the other hand (Fig. 4.11, right), displays no significant skill in the correlation score of the ensemble mean. It could be expected from the analysis of the feature selection approach (Fig. 4.10), that the central European zone displays skill but, even here, no significant skill is found. Overall, the results demonstrate the current lack of capability for (AI-)enhancement of forecasts of selected HW predictors. On the contrary, the data-driven approach (to be introduced in the following section), displays similar skill patterns to the dynamical system. Further discussions of added-value are included in the next section, once the data-driven approach is introduced.

The difference between Detect and Hybrid is a function of the forecast skill of HW predictors and the skill of the ML models linking predictors to local HW occurrence. More work is needed to understand and quantify the added-value of AI-enhancement. In particular, a validation of HW predictors in the dynamical system is needed.

4.4.3 Data-Driven Seasonal Forecasting

Alongside the hybrid approach, a data-driven forecasting system is developed, again within the optimisation-based feature selection framework. In the hybrid approach, predictor data from ERA5 is replaced with dynamical forecasts. Here, the framework is switched from a detection approach (as in Section 4.4.1) to an inherently forecast-based approach by

ensuring that potential predictors are restricted to certain lag times. In this way the system resembles a dynamical forecast system, which receives climate information only on and before the initialisation date and not after. The cut-off time for potential predictors determines the effective "initialisation" time; for example, using predictor data prior to May 1st to target summer HWs is equivalent to a May initialisation of the dynamical system. Here the framework is altered to target the number of seasonal HW days (e.g. May-June-July, Fig 4.12). We optimise the normalised RMSE (N-RMSE, normalised by interannual variability) of the predicted number of HW days.

Given that ERA5 data provides an insufficient amount of training data (one value per year for 72 years), here we employ a multi-centennial paleo-simulation developed at MPI-ESM (hereon in "past2k"). The past2k model is a reconstruction of the period 0-1850, using realistic land-use change and atmospheric forcings reconstructed from polar ice cores (Jungclaus et al., 2014 & 2017). The model climate is assumed to be stable enough for HW drivers to be consistent throughout the model period, as demonstrated by the lack of sensitivity of HW indicators to the climatology period chosen (Fig 4.12). Here, the period 1821-1850 (green) is used. Predictors are calculated by applying the cluster masks of ERA5 (D3.2) to past2k. Further details on past2k are found in Deliverable 3.2 and (Jungclaus et al., 2014 & 2017).

The development of a working data-driven seasonal forecasting system can also provide answers to several scientific questions. First, can the reduced dimensionality approach work for monthly/seasonal aggregated indicators as it did for daily HW occurrence (Section 4.4.2)? Second, how transferable is model-world ML training to the "real-world" (in our case, represented by ERA5)? In other words, are the model-world drivers sufficiently similar to those in the real world? Lastly, and of great practical importance, we will determine how the data-driven approach compares to the dynamical and hybrid systems.

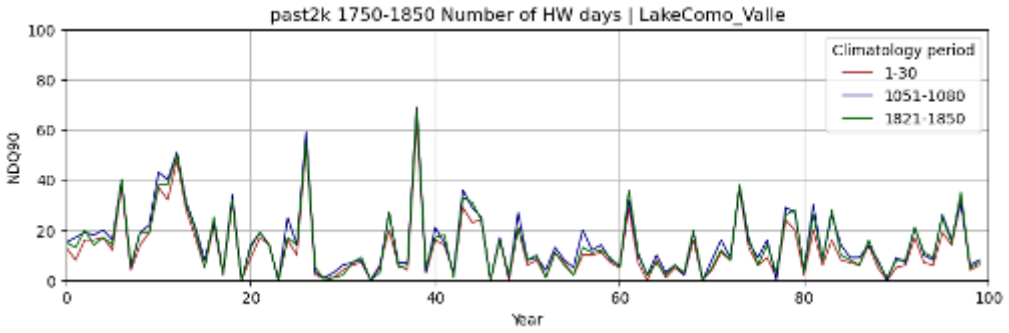


***Figure 4.12**: Sub-sample of past2k target data (number of May-June-July HW days) over the period 1750-1850, defined relative to diverse climatology periods.*

The first step is to determine the optimal HW predictors in past2k with the optimisation feature-selection framework. We perform FS across the European domain, using a training period of 0-1600 and a test period of 1601-1850. The optimisation of N-RMSE of the number

of summer HW days for an example grid point (Fig. 4.13) is representative of the wider domain; solutions converge to values below 1, albeit with some overfitting to the training data, before roughly 15000 evaluations. N-RMSE below 1 indicates error magnitudes less than the interannual variability, and as a rule-of-thumb corresponds to significant correlation values. Training and test scores for the optimal solutions across the European domain show that the framework provides skilful predictions of model-world HW indices (Fig. 4.14). Skill is highest around the Mediterranean and North Atlantic regions, while a relatively low-skill corridor stretches from Scandinavia to central Russia.

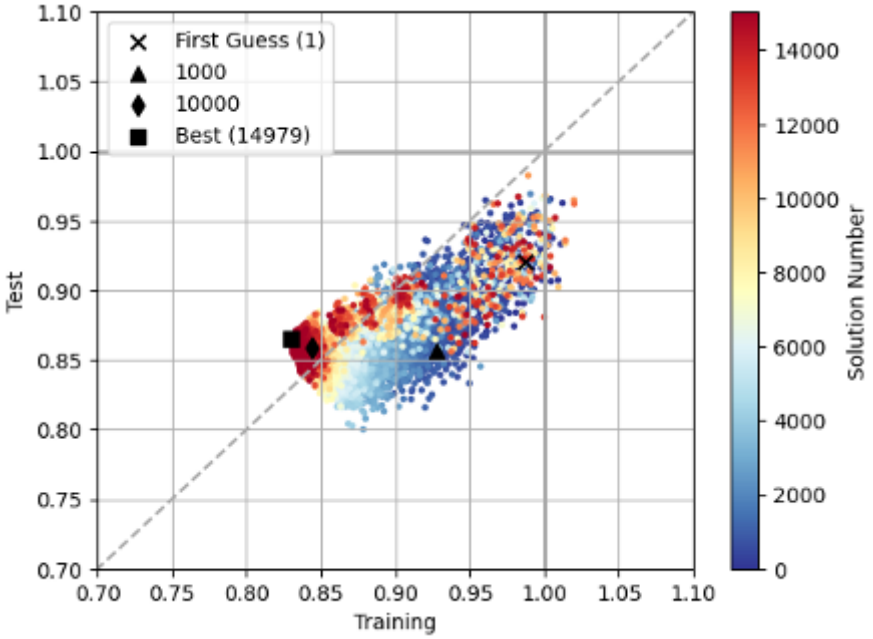


***Figure 4.13**: Example optimisation and feature selection for grid cell (East Mediterranean Sea).*

Collecting the optimal predictors from across Europe provides an overview of the model-world HW drivers at a regional level (Fig. 4.14). The most commonly picked predictors across the domain are the European soil moisture, temperature and z500 clusters. Predictors which represent more distant precursors include Arctic sea ice, and north Atlantic SST. Interestingly, the most frequently picked lag times for predictors falls around six weeks prior to initialisation (i.e. mid-March; Fig. 4.16). The key temporal lag, however, depends on the variables. Temperature and z500 clusters are selected more in the run-up to initialisation and decay gradually with lag, while soil moisture and sea ice selection peaks between 7-8 weeks prior to initialisation.

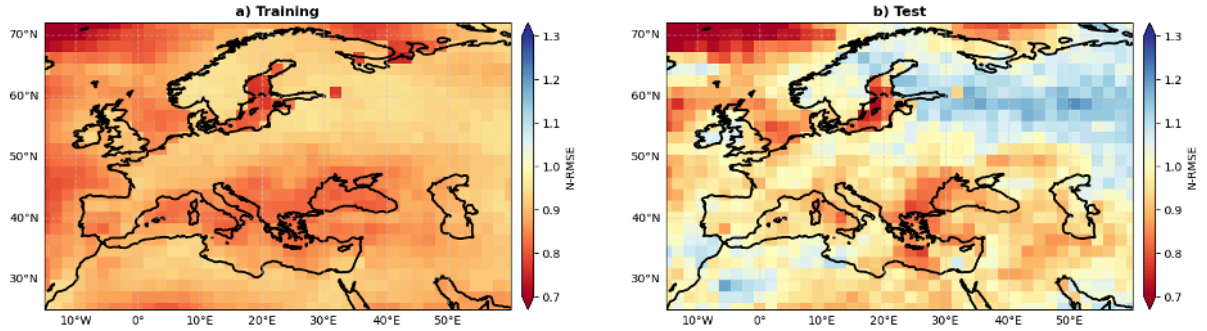


***Figure 4.14**: NRMSE of optimised solutions across the European domain for recreation of past2k HW indicators. Left: training period cross-validation 0-1600. Right: test period 1600-1850.*

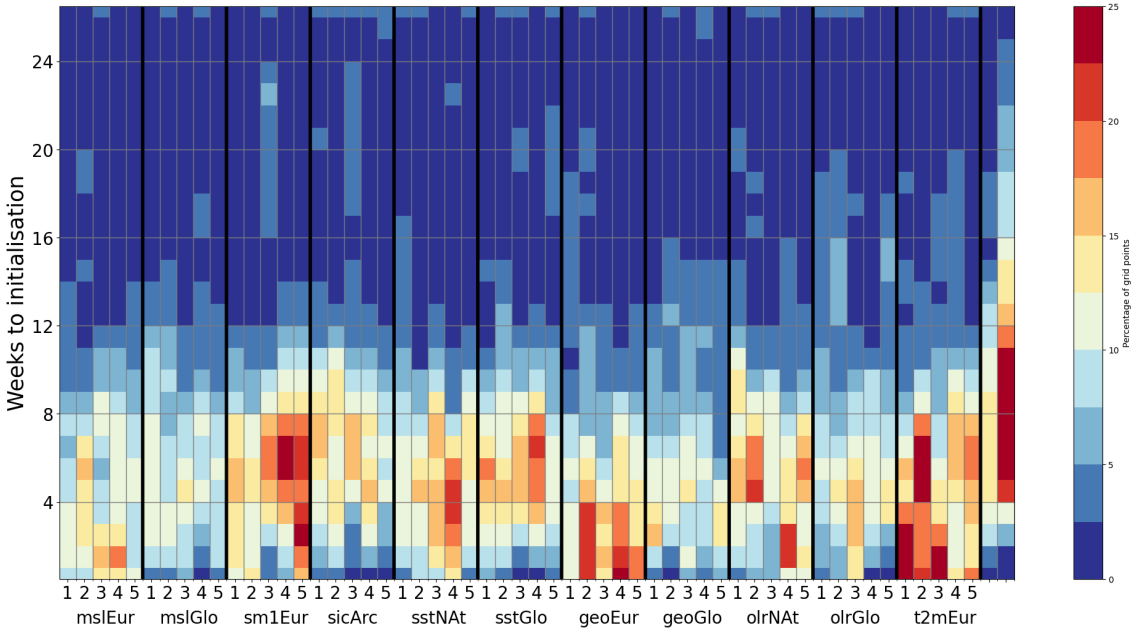


***Figure 4.15**: Identification of selected predictors for the whole European domain. Percentage of grid points which use cluster and lag in optimal solution. Weeks from initialisation (May 1st).*

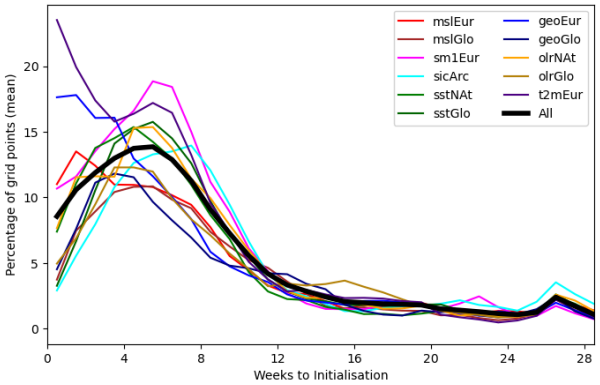


***Figure 4.16**: Percentage of grid points which select features based on lag time.*

Using the selected optimised predictors, the data-driven forecast system is adjusted to train on the entire past2k simulation period (0-1850) and then tested on ERA5. Data-driven reforecasts of the 1993-2016 period, using ERA5 as a benchmark, highlight significant correlations over central Europe and the Mediterranean Sea, with pockets of skill elsewhere (e.g. in the Northern Atlantic). Patterns of skill in model-world test period (Fig. 4.14) align with those of the data-driven forecasts (Fig. 4.17), in particular the insignificant skill over northern central Europe and northwestern Africa and the better skill over central Europe and the Mediterranean Sea, indicating the transfer of learning from the model- to the real world.

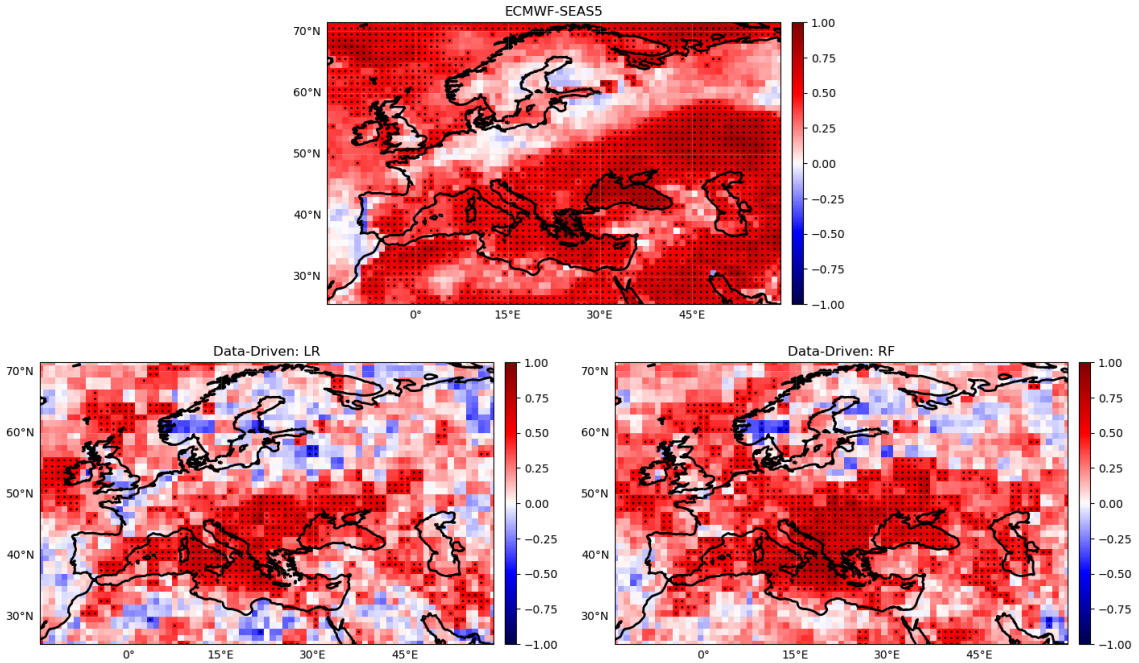


***Figure 4.17**: Anomaly correlation of skill scores over Europe for the period 1993-2016 for the dynamical system ECMWF-SEAS and data-driven approach implemented in two ML models (LR - Logistic Regression; RF - Random Forest). Black stippling indicates statistical significance.*

In the dynamical system, the area of significant skill extends much more widely than in the data-driven approach. Where the data-driven forecast displays skill, the dynamical does also, indicating that in its current state the former provides added-value not in an increase of skill but only through the reduced computational resources required. By inputting the predictors into a Random Forest model, we see an increase in the area of significant skill over the Linear Regression model; thus, more complex models are able to better leverage the same predictor information. Overall, this demonstrates that model-world training using early predictor information (i.e. at 4-8 weeks prior to the target season) can generate skilful forecasts in certain regions. By changing the validation period to 2004-2002 and comparing to the hybrid approach of Section 4.4.2 (A4.2), we find that significant skill is present across the same region but with different spatial patterns. Thus, it is premature to state which approach, data-driven or hybrid, adds more or any value over the dynamical system; instead, both have demonstrated the ability to make seasonal forecasts of summer heat extremes. The data-driven seasonal forecast system will be described in McAdam et al. (in preparation).

4.5 Night-time extremes

The work of Task 3.2 also covers the extension of heatwave analysis to complementary indices of heat extremes (described in D3.1 and Chapter 4.1). In particular, a focus has been placed on night-time extremes. Given the extensive development work on feature selection and new forecast systems performed in this task, we have not repeated all analysis for night-time extremes. However, all techniques developed are applicable to night-time extremes. Moreover, work on nighttime extremes has been covered in previous Deliverables. Here, we summarise the results and provide a perspective on differences in predictability and forecast skill.

As presented in D3.2 and in Torralba et al. (2024), records of heat extremes on the European scale depend on the time of day used to determine the temperature. Even seasonal statistics on nighttime heat waves (using Tmin) differ from those of daytime heatwaves (using Tmax). Warm Nights, defined with apparent temperature based on humidity, were also studied for their impact on both physiological productivity (i.e. comfort) and agricultural yield. D3.2 presented differences in the occurrence of day and night-time heatwaves, by means of a clustering analysis (e.g. Fig. 4.6 and A4.2). Meanwhile, non-negligible differences were found in seasonal forecast skill between day and night extremes (Torralba et al., 2024; Figure A4.3). Night-time and day-time extremes are, therefore, influenced by similar but not identical drivers, and their predictability is not equivalent. Future studies and operational systems should strive to differentiate between the two.

4.6 Summary and Future Steps

In Task 3.2, a range of ML-based techniques have been used to tackle ongoing problems in the field of heatwaves: identification of drivers, identification of key indicators, development of data-driven forecasting systems to compete with existing dynamical systems, and attempted enhancement of dynamical systems.

Regarding drivers, work has been done to achieve a reduction of dimensionality of key atmospheric variables and explore how well ML can recreate HWs (within the feature selection framework), from local to regional scales. This has implications for understanding drivers and forecasting HWs on a large scale, particularly for achieving greater computational efficiency. Other analyses have used ML to define particular phenomena (weather regimes), which are then studied for their role in HW occurrence.

A purely data-driven approach has been developed using training data from a long-term paleo-simulation. When applied to the "real" world (i.e. ERA5 data), large parts of Europe are well predicted, and in some areas this approach beats the ECMWF-SEAS5 dynamical system. These results imply that learning on model-worlds can be transferred. There is added value in terms of computational efficiency as well; optimisation performed on individual grid cell requires 4-5 hours of CPU time (total time found by multiplying by domain size) and once complete, each forecast requires only minutes to be performed. The next steps will explore whether the data-driven system skill can be boosted with multi-model training; 1850 samples are perhaps too few, and zones of low skill have been identified even in the recreation of model world.

A hybrid approach has also been developed, using forecasts of regional-scale HW predictors selected by the feature selection framework. The AI-enhancement does not display improved skill over the purely dynamical system in forecasting seasonal HW indicators. Further tests may include extending the pool of potential predictors. The added-value of the data-driven and hybrid methods, in terms of both skill and computational cost, will continue to be studied in WP9.

5 DETECTION OF EXTREME DROUGHTS

5.1 Introduction

Drought is a natural phenomenon mostly related to the reduction in the amount of precipitation received over an extended period of time, such as a season or a year, and is also influenced by other variables such as temperature, wind, humidity (Mishra and Singh, 2010). In contrast to aridity, drought is not permanent, although prolonged droughts may propagate through the full hydrological cycle, resulting in significant long-term economic, social, and environmental costs (Spinoni et al., 2016). Despite drought being a hot topic extensively studied in the literature, there is still no unanimity on its definition. Depending on the time horizon considered and the hydro-climatic variable used in drought characterization, droughts are generally classified into three categories (Pedro-Monzonís et al., 2015), namely meteorological, agricultural, and hydrological, respectively associated with anomalies in precipitation, soil moisture, and streamflow. Just as there is no single definition of drought, there is no single index that accounts for all types of droughts. Capturing the evolution of drought dynamics and associated impacts across different temporal and spatial scales still remains a critical challenge.

Here, we provide a final update on the activities of Task 3.3, in which we developed two methods for advancing the detection of extreme droughts supported by ML algorithms. The first method (described in Section 5.2) represents an improvement of the FRIDA methodology presented in Deliverable D3.2 for the identification of impact-based drought indices over the pan-European domain. The method aims to identify the relevant drivers of observed drought impacts (e.g., vegetation stress) from a pool of candidate hydro-meteorological predictors. The selected predictors are then combined into an index representing a surrogate of the drought impacts in the considered area. The second method (illustrated in Section 5.3) contributes a novel framework combining a synthetic generator of drought events and factor mapping methods for supporting the identification of the most critical drought features – i.e. intensity, duration, frequency – that produce the most severe impacts. This method is demonstrated using the Lake Como Climate Change Hotspot.

Beside these results focused on drought detection, other methodologies and experiments have been conducted in collaboration with WP2, WP6 and WP7 to produce also AI-enhanced drought forecasts:

- Deliverable D2.2 describes a novel method for meteorological drought forecasting, a streamflow forecasting approach based on Long Short-Term Memory models, and ML-based post-processing of hydrological model predictions
- Deliverable D6.2 illustrates and discusses the application of the developed methods for hydrological services at the pan European scale
- Deliverable D7.2 illustrates and discusses the application of the developed methods for drought forecasting in the Rijnland, Duoro, and Zambezi hotspots

5.2 AI-enhanced impact-based drought detection via multi-task learning

This chapter describes an AI-enhanced methodology that builds on the FRIDA methodology presented in Deliverable D3.2 to address two key challenges:

1. the presence of spatially correlated drivers, but spatially heterogeneous impacts, which can benefit from Multi-Task Learning methods;
2. the computational complexity of FRIDA due to its wrapper feature extraction process, which can be mitigated using a filter-based method.

The proposed method, which is described in detail in Deliverable D2.3, is demonstrated using the same pan-European case study presented in Deliverable D3.2.

5.2.1 Case study and data

This section presents the data used in our study of drought monitoring across the European continent. We focus on the FAPAR Anomaly as a proxy for drought impacts, leveraging satellite-derived data to assess vegetation health. Additionally, we detail the hydrological predictors derived from the European Hydrological Predictions for the Environment model (E-HYPE) and HydroGFD2.0 reanalysis data, which constitute the inputs for our models. The integration of these datasets allows us to explore both regional and global perspectives on drought dynamics.

5.2.1.1 FAPAR anomaly

FAPAR measures the fraction of solar radiation within the 400–700 nm spectral range that vegetation absorbs for photosynthesis, represented as a unitless ratio from 0 to 1. A value equal to 1 indicates that 100% of the incoming radiation is being absorbed by the vegetation canopy, suggesting that the vegetation is dense and highly efficient in absorbing all available light for photosynthesis and growth processes. It is a critical indicator for understanding the water, energy, and carbon balance within vegetation and plays an important role in various models including those for ecosystems, climate, and crop yield estimation (Qin et al., 2018).

Recognized as one of the 50 Essential Climate Variables (ECVs) by the Global Climate Observing System (GCOS), FAPAR is crucial for global climate monitoring and supports initiatives by organizations like the United Nations Framework Convention on Climate Change (UNFCCC) and the Intergovernmental Panel on Climate Change (IPCC). Satellite-based observations of FAPAR offer advantages over in-situ measurements due to their global coverage, long-term data availability, various spatial resolutions, and consistent data records (Peng et al., 2019).

The FAPAR Anomaly is monitored by the Copernicus Global Drought Observatory (GDO), calculated from MODIS FAPAR products averaged over ten-day periods and spatially represented at a 0.1-degree resolution from an initial 500m resolution. In our study, we recalculated the FAPAR Anomaly at monthly intervals to align with meteorological indices

such as SPI and SPEI, which are computed over monthly time scales. Furthermore, our model utilises predictors from the E-HYPE model (see next Section), which provides data at a sub-basin level across Europe, and the 0.1 degree pixels of the GDO FAPAR Anomaly often overlap with these sub-basins.

To derive the FAPAR Anomaly for each sub-basin, we utilise the raw satellite images of FAPAR (MOD15A2H, Collection 6) provided by NASA spanning from January 2001 to December 2018. Following the methodology outlined by GDO, we process the data as follows: we first download the FAPAR images along with metadata rasters. These enable us to mask out pixels affected by adverse conditions, such as cloud cover or high solar zenith angles. As a result, we retain only good quality pixels. Next, to derive monthly FAPAR values, we interpolate the original 8-day FAPAR images using a weighted average approach. This interpolation method assigns weights proportional to the number of days each pixel contributes to within the month. GDO also performs a temporal smoothing with an exponential filter (α=0.5), but since we are dealing with broader (monthly) FAPAR averages, we skip this step. We then spatially aggregate these monthly FAPAR values to align with sub-basins defined by the E-HYPE model. This aggregation involves calculating the average FAPAR values from the 500m resolution MODIS pixels falling within each sub-basin.

The original good quality 500m pixel measurements are not always present everywhere in the area under analysis. As a consequence, in order to maintain the integrity and significance of our derived FAPAR data, we implement specific quality heuristics: we exclude observations where the original MODIS pixels cover less than 25% of the sub-basin area. This criterion ensures that our dataset retains sufficient spatial coverage for meaningful analysis. Additionally, sub-basins are excluded if MODIS pixels never cover more than 50% of their area. This step aims at avoiding the characterization of sub-basins predominantly composed of non-vegetative surfaces, such as urban areas or lakes, with vegetation indices. Following these steps and quality checks, our study focuses on analysing FAPAR data for 34,066 sub-basins that meet those criteria.

The FAPAR anomalies are calculated by comparing the derived 1-month FAPAR values with a consistent baseline of FAPAR statistics, covering the period from 2001 to 2018. For each 1-month period, starting from January 2001, the FAPAR anomalies Y_t are computed as $Y_t = (X_t - X_m)/\sigma$ where X_t is the FAPAR for the 1-month period t of the current year, X_m is the long-term average FAPAR and σ is the standard deviation (both calculated for the same 1-month period t using the available time series). Due to missing satellite data from the MODIS sensor in June 2001, we exclude FAPAR Anomaly values from 2001 and consider in our analysis data from 2002. Furthermore, we only consider FAPAR Anomaly values from April to September, resulting in a maximum of 102 observations per sub-basin. The focus on the summer months is due to two main reasons: first, high-quality satellite data from the MODIS sensor are not available in winter months for high-altitude regions due to low light reflectance (caused by a large solar zenith angle); second, the vegetative activity during

winter is low in absolute value, but varies in the anomaly, which could negatively impact the calculations.

Since we are considering supervised models with one observation per month, we consider only sub-basins with sufficient observations. The maximum number of missing (NA) values allowable for retaining a sub-basin is set equal to 4, as this value corresponds to the elbow of the distribution of the number of sub-basins over the number of NA observations. This results in 30,007 valid sub-basins, each with 98 to 102 observations.

5.2.1.2 Hydroclimatic predictors

In this study, we consider some of the most widely used drought indices among the set of candidate predictors. These indices represent various components of the hydrological cycle, such as precipitation, soil moisture, and river flow, each associated with a specific type of drought. They reflect statistical anomalies relative to the long-term climatology at a given location and time, providing a measure of the probabilistic severity of a drought event. By fitting the long-term record of the considered variable to a probability distribution, drought is identified when observed values significantly and persistently fall below normal conditions (Mishra and Singh, 2010).

The eight indices considered in this work are:

- Standardised Precipitation Index at 1-month scale (SPI-1) and 3-month scale (SPI-3);
- Standardised Precipitation and Evapotranspiration Index at 1-month scale (SPEI-1) and 3-month scale (SPEI-3);
- Soil Moisture Anomaly Index at 1-month scale (SMA-1), 3-month scale (SMA-3) and 6-month scale (SMA-6);
- Standardised Streamflow Index at 6-month scale (SSI-6).

The accumulation time for each index is chosen based on the characteristics of the drought and its potential impacts (for further details, see Deliverable D3.1). SPI and SPEI, which are indicators of meteorological droughts, are computed for short accumulation periods (1 and 3 months) to indicate immediate impacts. SSI-6 is directly linked to streamflow and refers to long-lasting hydrological droughts (6 months). SMA represents agricultural droughts, typically a medium-term process, and is calculated over 1, 3, and 6 months of accumulation periods.

In addition to these eight indices, we also consider the raw hydrological variables from which these indices are derived. Although these variables exhibit strong seasonality, they can be advantageous in complex models. Specifically, the monthly observations of each raw variable for each sub-basin are obtained from the European Hydrological Predictions for the Environment model (E-HYPE; Lindstrom et al., 2010), combined with the HydroGFD2.0 reanalysis data over the period 1993-2018 (for further details, see Deliverable D3.2).

5.2.2 Multi-task learning drought detection

In this section, we present the workflow and the methods adopted for the development of models capable of detecting drought impacts on vegetation at the European scale using multi-task learning. Specifically, we examine two applications:

1. grouping nearby sub-basins into broader regions to develop local models;
2. developing a single global model that detects drought impacts across the entire European continent.

The two multi-task learning approaches have distinct motivations and potential outcomes. By employing local models, we aim to identify combined indices that better capture the vegetation state of each region. Given the importance of interpretability, especially in detecting environmental hazards, regions derived from the clustering algorithms are represented by the average observations of their respective sub-basins. This means that the FAPAR Anomaly of a region for a specific month is computed as the average of the FAPAR Anomalies of the sub-basins within that region for the month, and similarly for the predictors. Therefore, each region has around 102 observations, which should be less noisy than the original observations of each sub-basin. Due to the relatively small number of observations, we expect the local models to be simple and explainable. Conversely, the single global model might be more complex due to its broader coverage across Europe, encompassing a wider data distribution. In this setting, we aim to determine whether it is possible to improve the accuracy of local models by leveraging all available information, potentially at the expense of interpretability.

Therefore, this study sequentially explores the following frameworks:

- one local model for each region;
- a single global model fitted on observations from all the sub-basins.

These two approaches are contrasted against a baseline represented by one local model for each sub-basin.

The implementation of the two frameworks includes three key processes: model selection, clustering, and feature selection as summarized below.

Local models:

1. Model selection: we train and cross-validate one model for each sub-basin. Given the limited number of observations (at most 102), we choose between linear regression and support vector regression (SVR). SVR is robust against overfitting compared to other non-linear models. It is evaluated using a grid search over its hyperparameters (C, epsilon and gamma). Linear regression uses the year and the indices as features, while SVR also considers raw hydrological variables. Raw variables exhibit strong seasonality, making them unsuitable for linear regression aimed at interpretability. Their deseasonalised version results in the indices already used as predictors. However, non-linear models might leverage seasonal patterns in raw data to enhance predictive performance.

2. Clustering: we perform hierarchical bottom-up clustering and hierarchical Non-Linear Correlated Target-Feature Aggregation (NonLinCTFA), using both centroid and average linkage methods. To avoid the computational cost of clustering until a single cluster is reached, we periodically evaluate the performance of models trained on each cluster in cross-validation: the Mean Absolute Error (MAE) is calculated for each sub-basin, with the corresponding predicted FAPAR Anomaly of the region in which it is contained; we stop hierarchical clustering when the minimum average MAE across all sub-basins is found. Hierarchical NonLinCTFA, instead, automatically stops when it is no longer beneficial to aggregate; we evaluate it in cross-validation at completion. We select the clustering method to proceed with, based on overall performance and the nature of the resulting regions.
3. Feature selection: using the regionalisation from the previous step and the selected model, we test two feature selection methods: a forward wrapper and a CMI-based filter. Each method generates a feature ranking specific to each region. The impact of varying the number of selected features is globally assessed with cross-validation.

Global models:

Model selection: models are trained using all observations from all sub-basins. We consider two primary models: linear regression and a neural network. For neural network training, early stopping considers the validation error over a randomly selected fraction (20%) of the training set. The best-performing neural network architectures was identified according with the accuracy in cross-validation. The year is included as a feature, and raw hydrological variables are tested as inputs for the neural network. Additionally, we investigate the efficacy of incorporating geographical coordinates (longitude, latitude, and altitude). This inclusion effectively allows for region-specific combinations of features, while at the same time facilitating information sharing among sub-basins, leveraging their spatial relationships. Moreover, it aims to address potential missing information that other features may not capture.

The detailed description of the methods used in these frameworks is provided in Deliverable D2.3.

5.2.3 Numerical Results

5.2.3.1 Local models

Clustering sub-basins are aggregated into wider regions using traditional hierarchical clustering methods and hierarchical NonLinCTFA, with centroid and average linkages. The average MAE across all sub-basins obtained by these clustering methods, after applying a linear model to each cluster fitted on the aggregated time series, is shown in Figure 5.1. Traditional hierarchical clustering requires exploring an aggregation threshold, which specifies the condition necessary for merging two clusters. A lower threshold indicates a propensity to aggregate more. In our study, the optimal hierarchical clustering is found at

thresholds of 0.7 and 0.4 for centroid and average linkage, respectively. Conversely, hierarchical NonLinCTFA automatically discovers an estimated optimal solution.

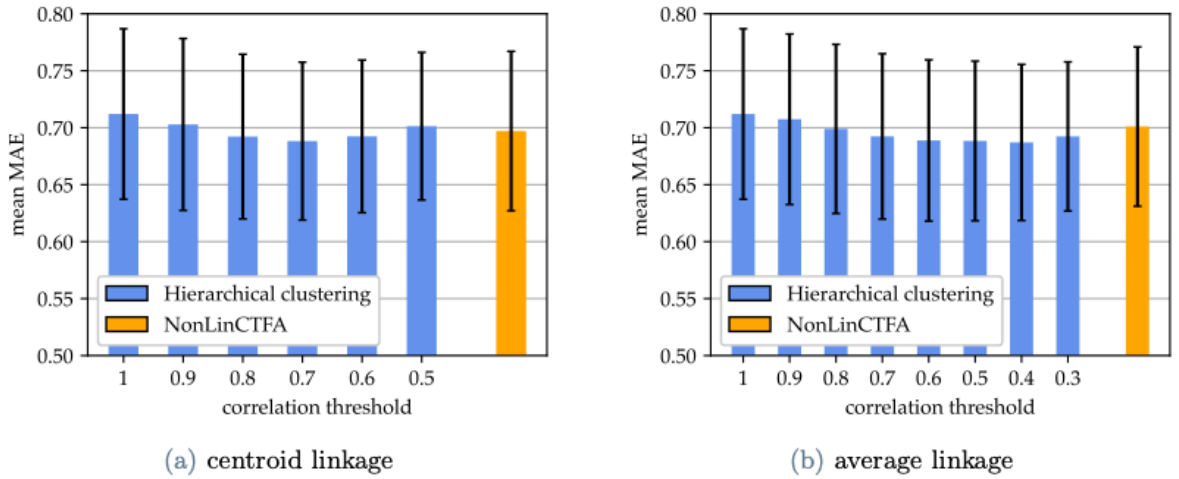


***Figure 5.1**: mean MAE across all sub-basins, obtained from utilising centroid (a) and average linkages (b), respectively. Both linkages are implemented within hierarchical clustering and hierarchical NonLinCTFA algorithms.*

The optimal hierarchical clusterings and the hierarchical NonLinCTFA clusterings computed during the first cross-validation split are displayed in Figure 5.2. The figure also shows the distribution of MAEs. Although the improvements in error at different clustering thresholds may seem modest, while the standard deviation is relatively high, the plots clearly show that the standard deviation remains stable and slightly decreases. This stability, coupled with a visual inspection of error on the map, suggests that the performance improvements from aggregating sub-basins are consistent and reliable.

However, these results show that adding the NonLinCTFA procedure to traditional hierarchical clusterings does not improve the performance. This is probably motivated by the limited number of samples in the available data. This limits the algorithm ability to learn the data distribution adequately and prevents an accurate estimation of the generalisation error.

To determine the most important drought predictors for a given region, we consider the hierarchical clustering method with average linkage and threshold of 0.4, since it demonstrates slightly better performance than the other computed clusterings and it creates relatively compact clusters.

We compare two feature selection techniques, a forward wrapper method and a CMI-based filter. Both methods independently select features for each region across different cross-validation splits. The time feature, primarily used to detrend the time series, is not included in the feature selection process, as it is instead added as a predefined feature in both algorithms. This ensures that feature selection accounts for its effect when ranking subsequent features. Both feature selection algorithms proceed by iteratively selecting the

next most important feature until all features are chosen. Subsequently, we reconstruct the FAPAR Anomaly considering only the year and the sequentially selected features for each cross-validation fold. Figure 5.3 illustrates how the mean MAE and mean correlation, computed across sub-basins, vary as more features are selected using both the filter and wrapper methods.

The forward wrapper consistently achieves higher accuracy compared to the filter method. This is expected since wrapper methods evaluate many combinations of features, optimising feature selection directly for the model's performance. In contrast, the CMI filter is able to capture non-linear relationships, potentially identifying features that cannot be fully utilized by a linear model. However, this is likely not affecting its performance. Indeed, when training the SVR models, we did not find significant improvements from considering non-linear relationships. Instead, a difficulty likely lies in the estimation of CMI, which may be challenging due to the limited data in each region's time series. Although globally the minimum error is reached when all features are utilized, some regions might still require fewer predictors. Since each cross-validation split has calculated its own clustering from the training folds, we determine, for each cluster in each clustering, the optimal number of features by identifying the minimum mean MAE on the validation fold for that cross-validation split. The mean MAE is calculated by averaging the MAEs from the sub-basins contained in the cluster.

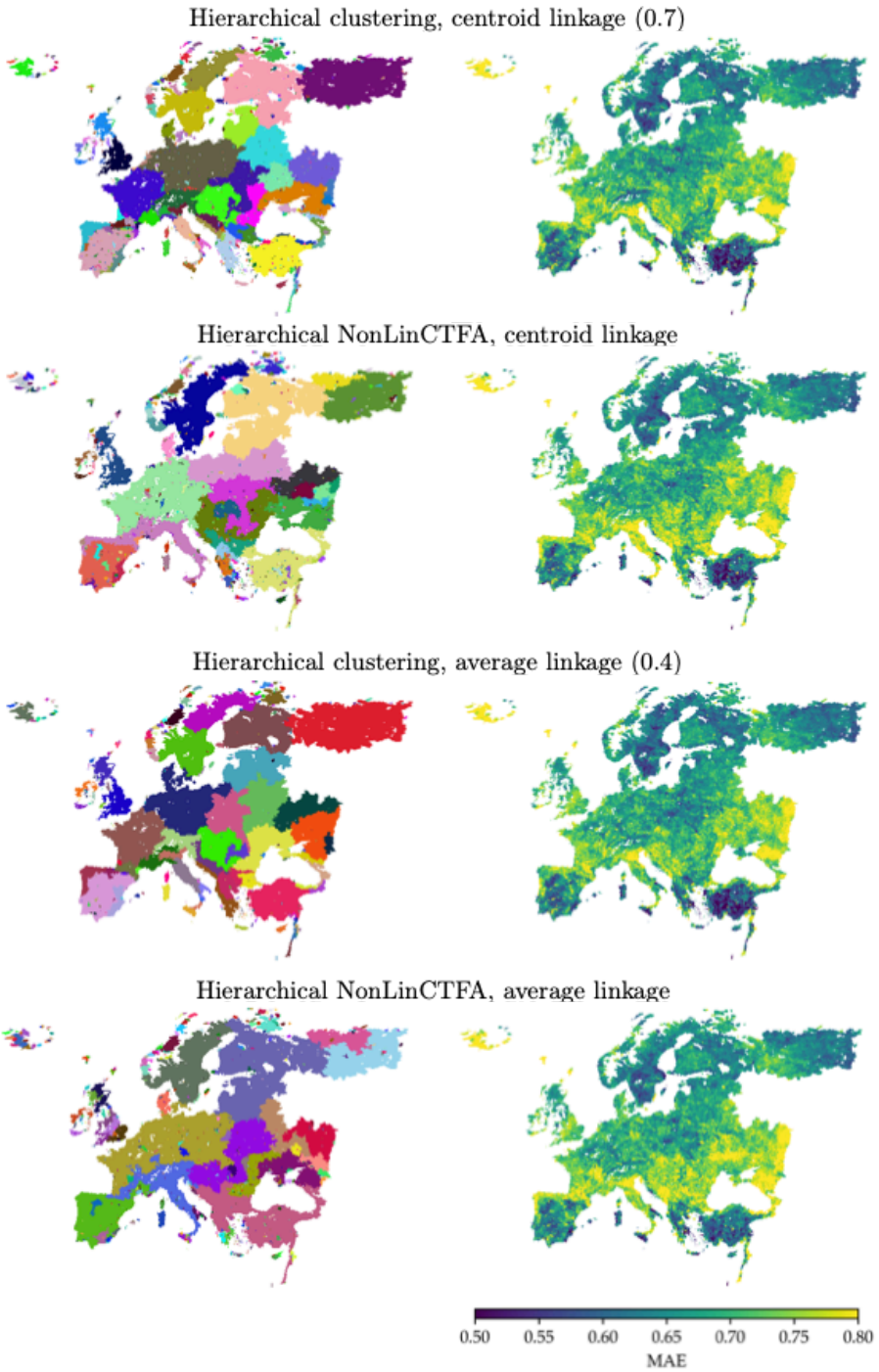


***Figure 5.2**: Results from local models. The hierarchical clusterings with the lowest mean MAE and the NonLinCTFA clusterings, considering centroid and average linkages, with their respective MAEs on the right. The clusterings visualized on the left are the ones generate in the first split of cross-validation, while the MAEs consider the entire reconstructed FAPAR Anomaly time series.*

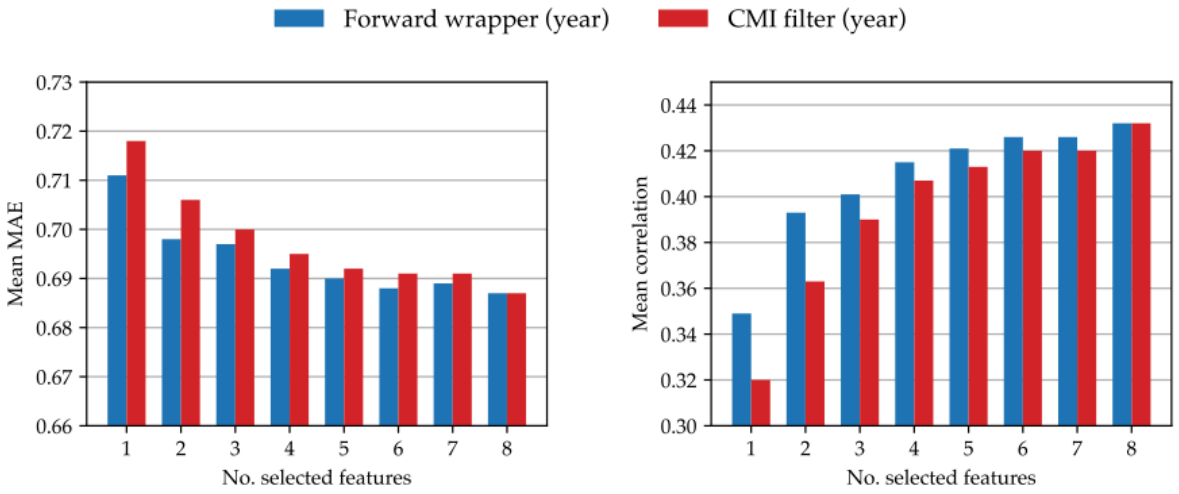


Figure 5.3: *Impact of CMI filter and nested forward wrapper as feature selection methods on model performance metrics with increasing numbers of selected features. The linear regression models are trained on data aggregated by hierarchical clustering with average linkage and a threshold of 0.4.*

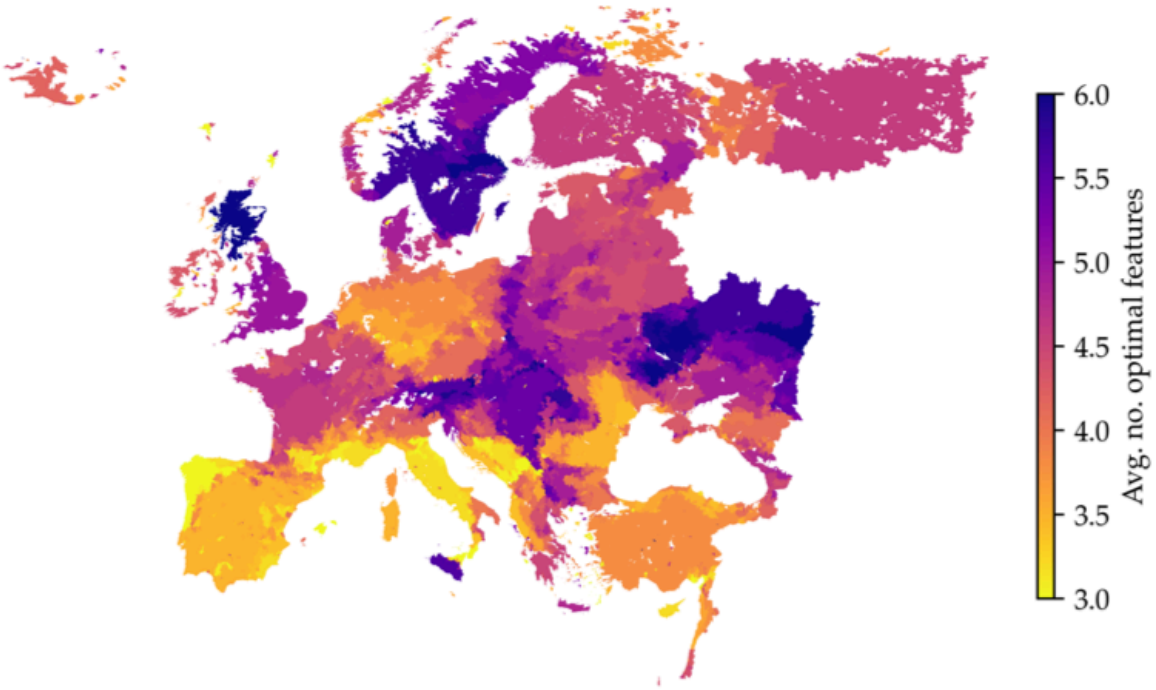


Figure 5.4*: Estimated number of optimal features in each region, obtained by averaging the number of optimal features for each cluster from the 17 clusterings (one for each cross-validation split).*

Figure 5.4 illustrates the optimal number of features considering all the 17 clusterings derived from cross-validation splits. The sub-basin's colour reflects the average number of optimal features among the 17 clusters that contain it, highlighting a relatively lower number of features required to detect drought impacts on southern European regions. At this point, it becomes interesting to understand which features are considered the most important and where. Since in each fold the algorithms might have selected different features for the same region, this can be a valuable way to assess the robustness of the dataset and algorithms.

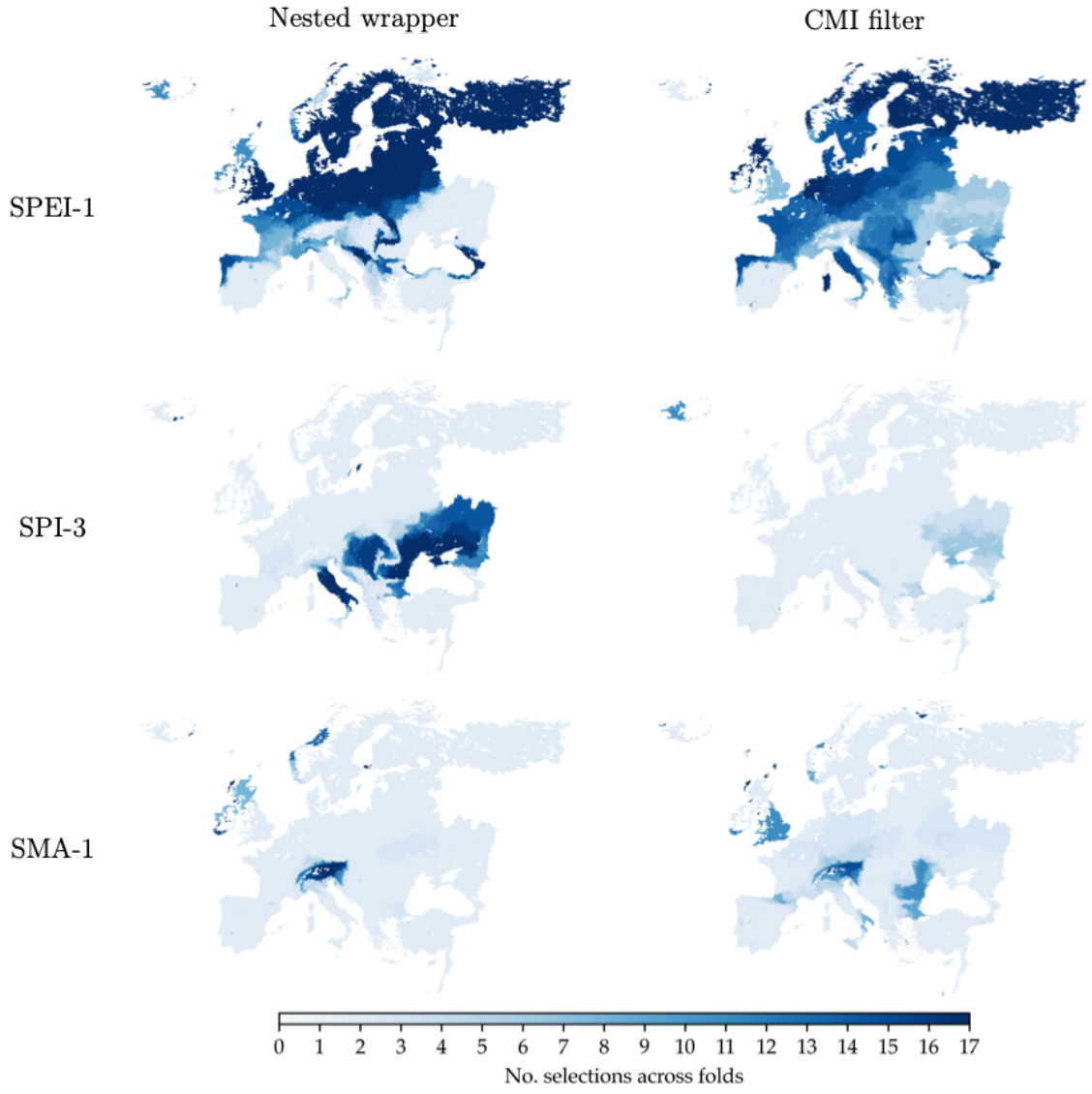


***Figure 5.5**: Regions where SPEI-1, SPI-3, and SMA-1 are selected by the nested forward wrapper and the CMI filter. Each map indicates where and how many times, during cross-validation, the feature has been selected.*

Finally, Figure 5.5 shows, for both forward wrapper and CMI filter, the regions where specific features such as SPEI-1, SPI-1, and SMA-1 are selected as the most important (selected first) across all clusterings generated by cross-validation. Observing the results of the wrapper method, these three features are consistently chosen in specific regions as the first option across multiple folds, indicating robustness in feature selection. The CMI filter method agrees with the wrapper only on certain regions, highlighting differences in the computation of the two methods.

5.2.3.2 Global models

The global model is again trained in 17-fold cross-validation, but this time each training set consists of observations from all sub-basins – approximately 3 million. In particular, we tested two types of models, a linear regression and a neural network with a single layer of 16 neurons. Table 5.1 presents the mean MAE and mean correlation from the predicted FAPAR Anomaly time series, by utilising the two models with different sets of predictors:

1. year and indices;
2. year, indices, and variables;
3. year, indices, variables, and coordinates.

The results demonstrate that linear regression performance remains consistent across different feature sets and it is worse than that of local models. On the other hand, the neural network significantly improves accuracy when incorporating raw hydrological variables, which inherently exhibit strong seasonality. On average, its predictions surpass those obtained with local models, meaning that in this setting it is beneficial to assume that all observations come from the same joint distribution. The inclusion of geographical coordinates has minimal impact. Either they are not significant, or in complex models like neural networks interactions between existing features may already encapsulate spatial dependencies implicitly.

***Table 5.1**: Performance of different global models in reproducing the FAPAR Anomaly time series of all sub-basins using alternative sets of input features, namely year and standardized drought indices only, considering also the original variables, and considering also the coordinates of the different sub-basins.*

Model	Features	Mean MAE	Std. MAE	Mean correlation	Std. correlation
Linear Regression	year+indices year+indices+variables year+indices+variables+coordinates	0.724 0.723 0.724	0.049 0.048 0.049	0.339 0.344 0.339	0.121 0.121 0.121
Neural Network	year+indices year+indices+variables year+indices+variables+coordinates	0.724 0.660 0.657	0.047 0.057 0.058	0.343 0.497 0.501	0.113 0.109 0.111

Given the results of the local models showing how the inclusion of an aggregation phase through hierarchical clustering with average linkage yields an improvement in the results, together with a reduction of computational costs, we finally explore the potential for a single model trained on the observations of all sub-basins aggregated through hierarchical clustering. Specifically, we use the hierarchical clustering with correlation threshold of 0.7, considering year, indices, variables, and coordinates as predictors which resulted to attain the best performance in terms of MAE.

Figure 5.6 compares the global model working on individual sub-basin with the one trained on aggregated clusters: a slightly improved MAE and a reduced variance across the estimates of the different sub-basins is obtained. This shows that the neural network is exploiting the heterogeneity of the data from different sub-basins to model the state of vegetation at the European scale, and it is convenient to average similar data, reducing both collinearity and computational costs.

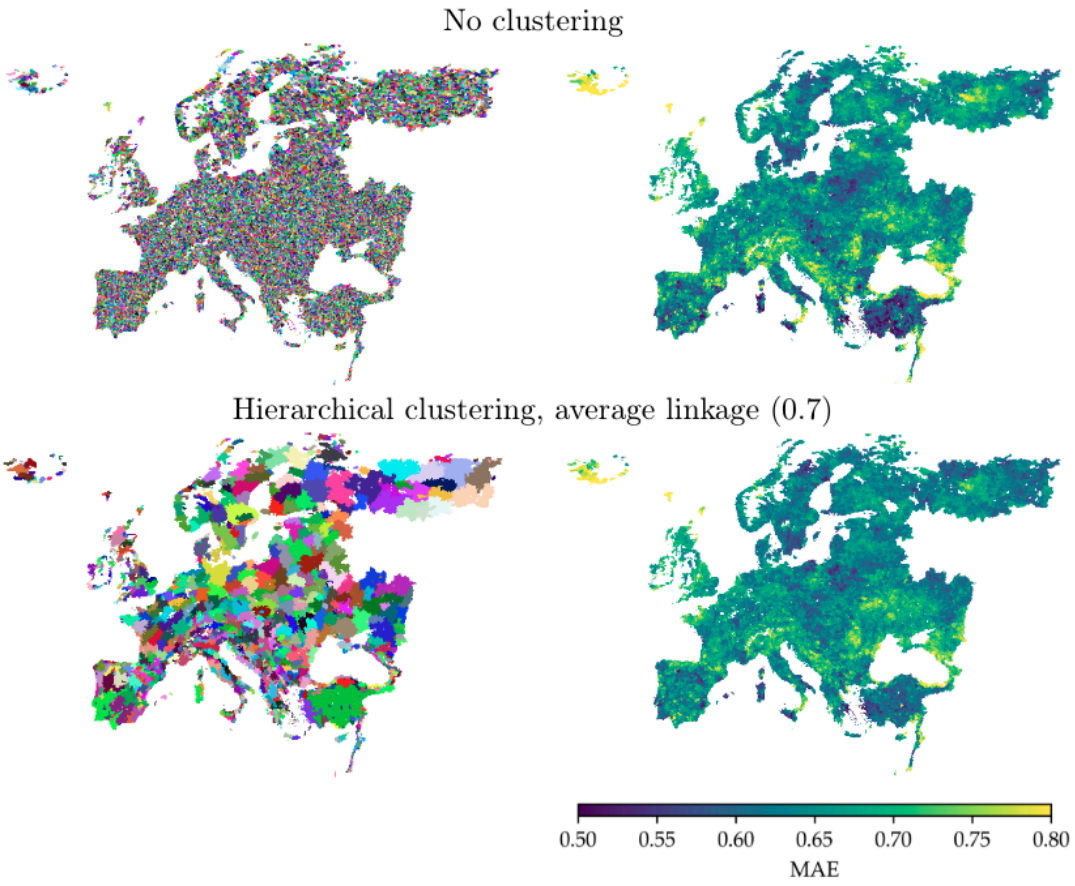


***Figure 5.6**: Comparison of global models considering individual sub-basins as baseline. The maps on the left show the granularity of the input features, while the maps on the right the corresponding model's accuracy in terms of MAE.*

The superiority of the global modeling approach is confirmed looking at Figure 5.7, which reports the results on a specific area in northern Italy. In this region, local models show difficulties in the prediction of the target variable. In particular, some clusters are well-performing, while others (such as the cluster isolating the Po Valley region) present substantially higher errors. The global approach, on the other hand, clearly exhibits an improved accuracy that is likely motivated by the enlarged dataset used for training a single global model.

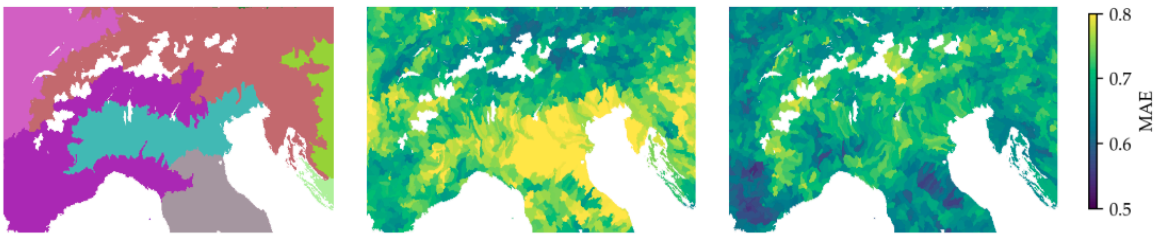


***Figure 5.7**: Comparison of local and global models on a specific area in northern Italy. The left map shows the clusters obtained in the local case, with the associated models' MAE visualized in the middle panel. The right panel shows instead the MAE of the global model trained on clustered features.*

To conclude the analysis of the global models, forward wrapper and CMI filter feature selection allow to drive some conclusions on the relevance of the considered features. In particular, Figure 5.8 shows the accuracy of different global models that use from 1 to the full set of 18 input features, as selected by CMI and wrapper feature selection. The same results with the addition of coordinates (i.e., latitude, longitude, altitude) are also reported in the figure, showing a significant performance with the inclusion of the coordinates also with a very limited number of features. Additionally, considering six features and the coordinates, the performance is close to the best value, showing that the majority of information provided by the inputs is already exploited. Finally, from 8 features on, there is no improvement with the addition of coordinates, proving that the neural network model is exploiting the larger number of features to identify the location of the samples, without the need to explicitly encode the coordinates in the inputs.

Finally, Figure 5.9 reports the most selected features in the global models considering the wrapper feature selection approach with 6 features. In this case, the SPEI-1 is still considered as the most selected features, as in the local case, while the SPI-1 is not anymore among the most relevant ones, being replaced by more long term-based features such as its three-months counterpart (SPI-3), and the three-months SPEI, together with its associated raw feature.

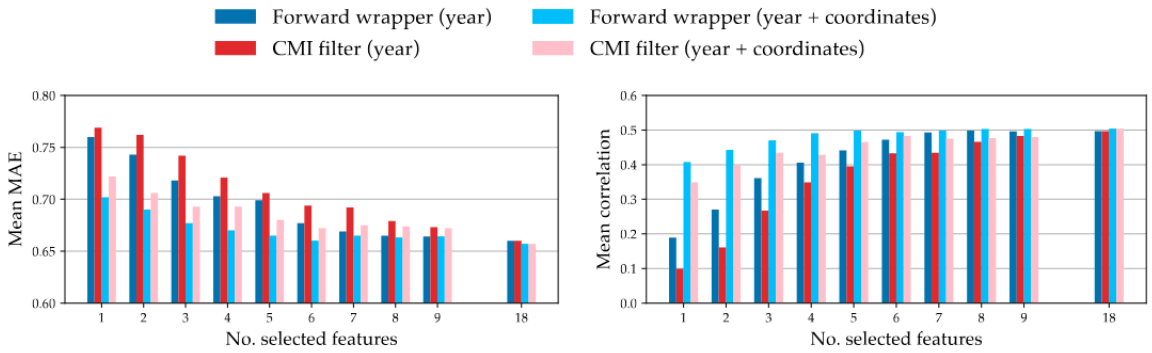


***Figure 5.8**: Accuracy of global models combined with CMI and wrapper feature selection methods in terms of average MAE (left panel) and correlation (right panel).*

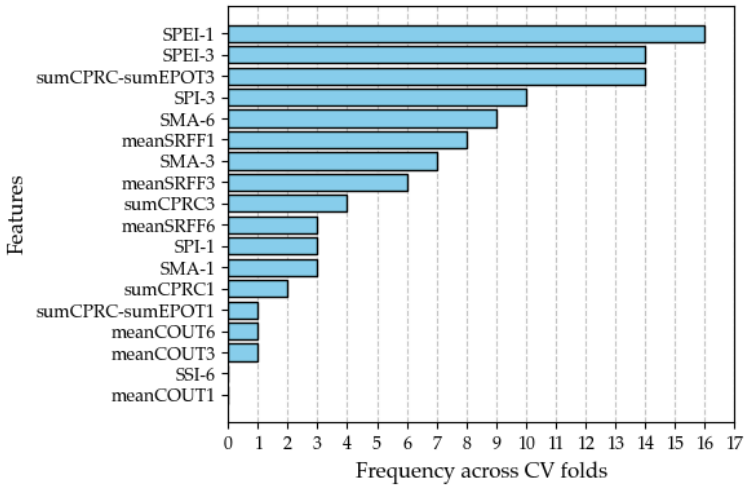


***Figure 5.9**: Most selected features for the global model, considering wrapper feature selection with 6 selected features in 17-fold cross-validation.*

5.3 Identification of critical drought features

This chapter introduces an AI-enhanced methodology to identify the most critical features of drought events, namely intensity, frequency, duration, that produce severe drought impacts. The proposed method is demonstrated using the Lake Como Climate Change Hotspot as a case study.

5.3.1 Impact model

Lake Como is an Italian subalpine lake with an operative storage capacity of 247 Mm^3 regulated since 1946 by a regional public authority, Consorzio dell'Adda, that operates the dam located at Olginate (South-East branch of the lake). Its hydrological basin is in the Italian Alps, close to the border with Switzerland, and corresponds to the upstream part of the

Adda River Basin with a catchment area of 4500 km^2. The water resources drained by the Lake's basin provide water for irrigation to a wide cultivated area (1320 km^2) and for energy production to 16 hydropower plants (13% of national hydropower). Most hydropower plants are in the northern upstream part of the basin, but some run-of-the-river plants are also operated downstream of the lake. The operation of the lake dams is also fundamental to controlling flooding along the lake shores, particularly in the city of Como. Additional interests are represented by tourism (especially after the development of the last two decades), navigation, fishing, and ecosystem conservation.

To simulate the dynamics of the Lake Como system, a simulation model of the reservoir operations at the daily time step is adopted here. Its main component is the water balance equation:

$$s_{t+1} = s_t + i_{t+1} - r_{t+1} \quad (5.1)$$

being s_t and s_{t+1} the lake storage at time t and $t+1$, i_{t+1} the net inflow (which already includes losses such as those due to evaporation and infiltration) into the lake between time t and $t+1$, and r_{t+1} the release in the same time interval. The actual release r_{t+1} is modeled through a stochastic and nonlinear relationship of the release decision u_t. The releases from the Olginate dam are indeed constrained by the minimum (N^{min}) and maximum (N^{max}) release functions, which respectively define the minimum and maximum outflow from the lake for each possible level. These functions are mathematically defined as follows:

$$N^{min} = \{0 \qquad if\ h_t < h^{lb}\ q_t^e \quad (5.2)$$

$$N^{max} = \{0 \qquad if\ h_t < h^{lb}\ 153^{4} \quad (5.3)$$

being h_t the lake level on day t at 8 am and q_t^e the minimum environmental flow. $h^{lb} = -0.4\ m$ and $h^{ub} = 1.1\ m$ defines the lower and upper bound of the operating range, respectively. The legislation requires completely opening the dam gates above h^{ub}, and closing them below h^{lb}. Between h^{lb} and h^{ub}, the lake operator can decide the amount of water to be released, provided that it does not exceed the range defined by the minimum (N^{min}) and maximum (N^{max}) release curves:

$$r_{t+1} = min(N^{max}, max(N^{min}, u_t))\Delta t \quad (5.4)$$

The decision on the release u_t is the output of the so-called operating policy, a function that takes in input the lake level h_t and information on the time of the year:

$$u_t = p_\theta\left(h_t, sin(2\pi t/365), cos(2\pi t/365)\right) \quad (5.5)$$

θ is a vector of parameters that define the shape of the policy, which is specifically selected to be highly flexible. In this case, we use a network of radial basis functions (Giuliani et al., 2016) specifically identified to reproduce the historical operations.
Using this formulation, the system can be simulated starting from an initial level to generate the level and release trajectories for a time horizon of length H. From them, the numerical values of the downstream deficit can be computed through this equation:

$$J^{def} = \frac{1}{H}\sum_{t=1}^{H}\left(\left(w_t - \left(r_{t+1} - q_t^e\right)\right)^n, 0\right) \quad (5.6)$$

where w_t is the downstream water demand that represents both the needs of the agricultural districts and of the hydropower plants. The exponent n is set to 2 during the irrigation season (from April 1st to October 10th); to 1 for the rest of the year. In this way, deficits (especially the largest) are weighted more during the irrigation season, when many mild shortages are preferred to a few severe ones.
Inflow scenarios given in input to the impact model are analysed using the Standardized Streamflow Index (SSI; Vicente Serrano et al., 2011), which adopts the same procedure of the Standardized Precipitation Index (SPI) but is used to consider streamflow instead of precipitation anomalies.
The analysis of the SSI allows to univocally identify a drought. It begins when the SSI is lower than -1 for at least two consecutive months and ends when the SSI becomes positive. Applying this definition to the historical data (1946-2021) of the inflow into Lake Como we spot 18 drought events (Figure 5.10). Other than the drought frequency (number of droughts in the considered period), Zaniolo et al (2023) defined two other key properties: persistence (the total duration of a dry spell) and intensity (the mean SSI value during a drought). The 18 drought events shown in Figure 1 have an average duration of 10.67 months and intensity equal to -1.24.

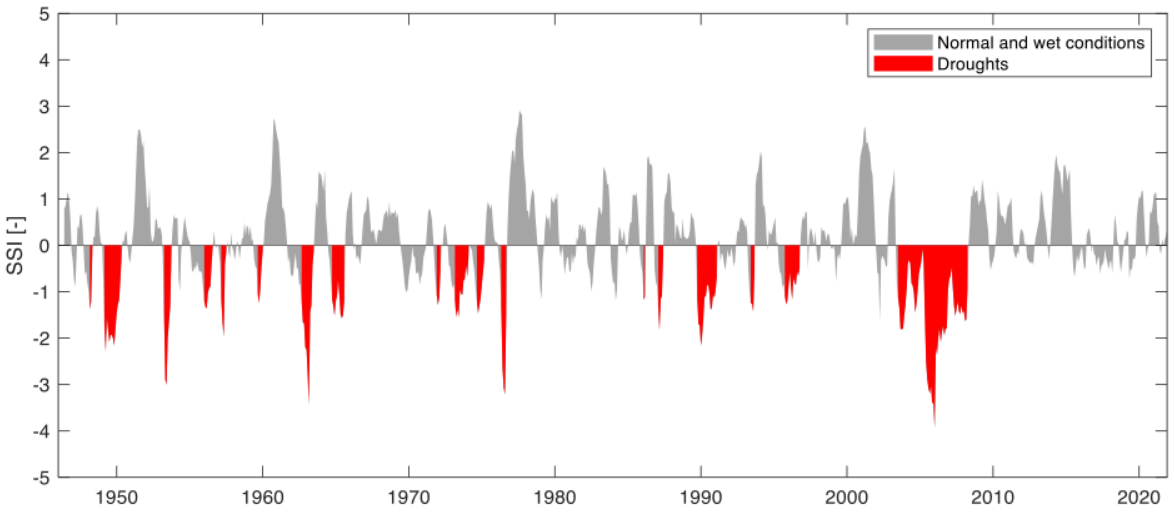


***Figure 5.10**: SSI drought index computed for inflow into Lake Como in the historical period 1946-2021. The hydrological droughts are highlighted in red.*

5.3.2 AI method

The proposed methodology to identify the most critical features of drought events is based on two sequential steps: the first consists of generating a comprehensive set of synthetic scenarios, spanning different drought conditions that could potentially occur due to the future hydro-meteorological regimes influenced by climate change; the second analyses the performance indicator obtained by simulating the system with the synthetic drought scenarios in order to define which are the critical drought features leading to a failure.

5.3.2.1 Synthetic generation of drought scenarios

We perturb the historical inflow time series following the FIND (Frequency, INtensity, and Duration) algorithm presented in Zaniolo et al (2023). It generates an arbitrary-length scenario with controlled statistical features, matching user-specified values of frequency, intensity, and duration. FIND is composed of the following steps:

1. Parameter and time series initialization: The user defines the Simulated-Annealing (SA) (Kirkpatrick et al 1983) parameters and selects the target frequency, intensity, and duration of droughts. The initial (parent) time series is generated by randomly extracting monthly values from historically calibrated monthly streamflow distributions.
2. Swapped time series generation: a new (swapped) time series is generated by replacing a randomly selected portion of predefined length from the parent time series.
3. Objective value calculation: the aggregate objective value is calculated for both the parent and swapped time series as a weighted sum of 5 single objective values (deviation from the target frequency, intensity, duration,

monthly autocorrelation over 12 months, and the 25th, 50th, and 75th percentiles during non-drought periods).

4. Time series selection: a time series is selected between the parent and swapped to become the new parent time series for the next iteration. According to SA selection principles, if the swapped time series has a lower (better) objective value, the swapped time series becomes the new parent. If the parent time series has a lower objective, the algorithm can occasionally select non-improving swaps following the SA principles.
5. Iterate until termination: the time series selected during the previous step becomes the new parent time series. The algorithm proceeds by iterating through steps b–e until one of the two terminating criteria is met, namely the parent time series aggregated objective is lower than a tolerance, or the maximum number of function evaluations is reached.

The described procedure is repeated for $N = n{\cdot}m$ scenarios: n combinations of frequency, intensity, and duration are extracted using Latin Hypercube Sampling (LHS); for each of them, replicas are generated (the features can be associated with different streamflow time series). The variability ranges of LHS must be defined to explore relevant combinations of frequency, intensity, and duration.

5.3.2.2 Scenario discovery

Once the synthetic generation of scenarios is completed, the scenario-discovery phase is conducted through these steps:

1. Impact model simulation: For each scenario, a simulation of the impact model is performed in order to compute values of the performance indicator.
2. Failure threshold Definition: Each scenario must be labelled as a failure (or non-failure) depending on its performance. This requires defining a failure threshold that is defined based on prior knowledge about the water system under examination and/or a statistical analysis of the performance indicator values over the historical period.
3. Supervised learning task framing: The data must be organized into input-output pairs. For each scenario, each representing a sample, the inputs are the duration, intensity, and frequency of the drought event. The target output is a boolean variable (failure/non-failure).
4. Decision tree classifier identification: A Classification And Regression Tree (CART) is fitted by solving the classification task described in step c. Mathematically speaking, the decision tree maps the failure/non-failure from the space of the performance indicator (or indicators if there are more than one) into the 3D space of the drought features. The decision tree structure must be limited (not too deep or too wide) to ensure a high interpretability. The inspection of the tree structure allows us to define an importance ranking of the drought features. The splitting points provide the thresholds separating failure/non failure for each feature.

5.3.2.3 Numerical results

As described in the previous section, the first step of the procedure requires sampling combinations of drought persistence, intensity, and frequency by LHS. The ranges have been defined starting from the historical features:

1. Persistence between 7.47 months (70% of the historical average duration) and 63.80 months (110% of the historical longer drought event);
2. Intensity between -2.59 (110% of the most intense event) and -0.87 (70% of the historical average intensity);
3. Frequency between 3.13 and 6.77 (60% and 130% of the historical frequency, respectively). Note that historical frequency has been rescaled over the 22-year horizon used for simulation, instead of the original 76-year horizon shown in Figure 5.10.

The LSH sampling ensures uniform coverage of the 3D feature space (Figure 5.11a), also including very critical drought with extreme persistence, intensity, and duration simultaneously.

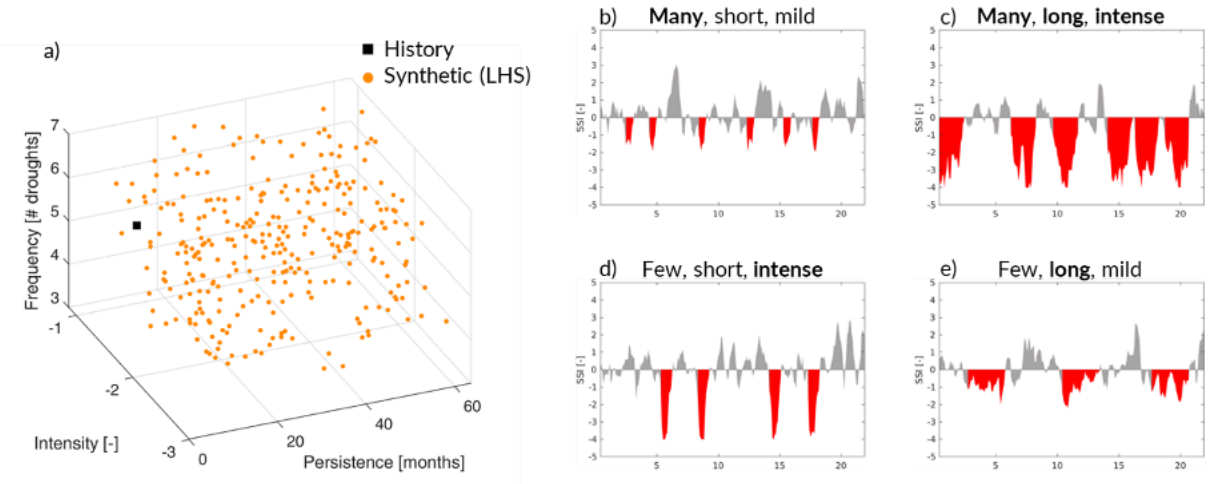


***Figure 5.11**: Three-dimensional space (persistence, intensity, frequency) with three hundred LHS samples (orange circles), with a black dot representative of the historical period (a). Examples of SSI time series for some extreme cases (b-e).*

For each combination, we run the FIND algorithm $m = 10$ times, each one starting from a different random seed thus producing different time series of the inflow and SSI. Four examples of SSI time series relative to extreme drought features can be seen in Figure 5.11b-e. In this way, we produced $N = 3000$ synthetic scenarios that have been then fed into the impact model to compute the downstream deficit (i.e., the considered performance indicator). The failure threshold has been set to 2933, corresponding to the average deficit of the three most critical years of the considered historical horizon 2000-2021, obtaining 2362 non-failure scenarios and 638 failure scenarios.

The decision tree trained to solve the supervised classification task is shown in Figure 5.12. Its structure suggests that intensity is the most critical driver of water supply failure in the Lake Como system, followed by persistence and, as last, frequency. Scenarios with average intensity >-1.8 are not usually strong enough to generate a system failure. Among those with intensity ≤ -1.8, an average persistence > 23.8 months and a frequency > 2.5 events over 22 years are necessary to generate a system failure.

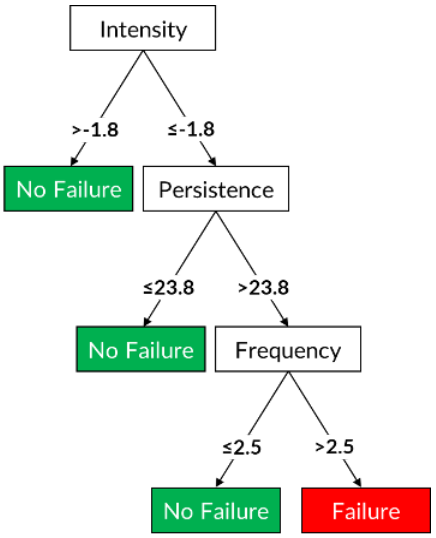


Figure 5.12. *Decision tree classifier structure.*

Finally, an assessment of the decision tree accuracy is fundamental to determine the significance of the importance ranking and the thresholds reported above. The analysis of the structure of inaccurate decision trees may lead to meaningless insights. In the case at hand, the decision tree has an overall accuracy of 0.89, perfectly balanced between the two possible output classes (Table 5.2), ensuring the significance of the analysis presented above.

Table 5.2: *Confusion matrix assessing the accuracy of the decision tree classifier.*

Actual		No Failure	Failure
	No Failure	0.89	0.11
	Failure	0.11	0.89

Decision Tree Output

5.4 Conclusions

In Task 3.3, a range of ML-based techniques have been used to address the challenge of designing impact-based drought indices via feature extraction. Moreover, given the complex relationship between drought characteristics and impacts, we developed a novel method for supporting the identification of the most critical drought features – i.e. intensity, duration, frequency – that produce the most severe impacts.

Regarding the definition of impact-based drought indices, the proposed Multi-Task drought detection method addressed the two limitations of the FRIDA method identified in Deliverable D3.2 and advances its upscaling at the pan European scale. The results reported in Section 5.2 show that the local approach is able to design a simple, fully interpretable index for each sub-basin based on linear aggregations and linear models, which is fast to train and easy to apply. Conversely, the global modelling approach is able to capitalize on the larger training dataset encompassing all sub-basins and yields the highest model accuracy in reproducing the drought impacts on vegetation. However, this second approach introduces high nonlinearity in the index reconstruction, which becomes ultimately difficult to interpret.

Regarding the methodology for identifying critical drought features, it effectively combines synthetic scenario generation with impact model analysis. These scenarios (generated using the FIND algorithm by perturbing historical inflow data based on frequency, intensity, and duration) are used in a supervised classification task that indicate drought intensity is the most influential factor driving water supply failures in the Lake Como case study. This method is fully generalizable and able to provide insights for managing future drought risks in other systems with different drought-related impacts.

The next steps will explore the added value of using these AI-enhanced, impact-based drought indices for better understanding similarities and differences in multisector impacts in the Lake Como climate change hotspot, in order to advance existing monitoring practices that are used for triggering drought management actions (these results will be reported in Deliverable D7.3). Moreover, we will also explore future drought risks at the pan European scale by producing projections of the AI-enhanced indices to support climate change adaptation strategies (these results will be reported in Deliverable D6.3).

6 COMPOUND EVENTS AND CONCURRENT EXTREMES

6.1 Introduction

While many type of extreme events have been associated with significant socio-economic impacts (e.g., Zampieri *et al.*, 2017; Zscheischler and Fischer, 2020; Hao *et al.*, 2022), the combination of such events, in space and/or time, can further amplify these impacts in a non-linear manner. This is because many systems possess resilience against single extreme events but cannot cope with multiple stressors (Leonard *et al.*, 2014; Hao *et al.*, 2018; Zscheischler *et al.*, 2018; Zscheischler and Fischer, 2020; Zscheischler *et al.*, 2020; Zhang *et al.*, 2021; Hao *et al.*, 2022). Multiple types of extreme events that are dependent in space and time are often considered concurrent extreme events (Toreti *et al.*, 2019b). However, the combination of multiple climate drivers can cause significant impacts without any of them being individually extreme. For instance, some ecosystems are directly adapted to the co-variability of temperature and precipitation, such that a bivariate anomaly can have a large impact without either variable being extreme in a univariate sense (Mahony and Cannon, 2018). This has motivated the study of compound events, which are often defined as multiple climate events and/or hazards that contribute to societal or environmental risk (Zscheischler *et al.*, 2018). Given that many socio-economic sectors are affected by weather and climate conditions, adequate risk assessment relies on understanding the multivariate nature of these events (Raymond *et al.*, 2020).

Within CLINT we aim to characterise compound events at European scale, thereby focussing on the socio-economic sectors addressed in WP6. Our focus is on:

a) Relatively wet and warm late winters followed by dry and warm spring, with severe impacts on agriculture.
b) Dry winters followed by hot summers, since accumulating pressure on the agriculture and the energy sector with direct impacts on the hydropower capacities during increased demand period.
c) Wet and warm spring, with impacts on water management, increased flood risk due to precipitation excess and early melting season.

The first event is motivated by so-called *false spring events*: these occur when humid and warm conditions manifest in winter triggering an unusual early crop growth, which is then harmed by a frost event or an intense drought in the following spring resulting agricultural damages (Ault *et al.*, 2013; Allstadt *et al.*, 2015; Chamberlain *et al.*, 2019). Due to climate change and the resulting increase in winter temperature, such events are expected to occur more frequently in the future (Ault *et al.*, 2013). The second type of compound event pertains to water shortages in winter, which are then exacerbated by prolonged drought and heat in summer. This situation places significant stress on summer crops, such as grain maize, and challenges water management systems, as irrigation may fail due to water

shortages. Furthermore, the energy sector can be adversely affected, particularly in the context of hydropower generation. The final compound event concerns the alpine region and is primarily characterised by flooding resulting from snowmelt due to warm temperatures, compounded by additional rainfall. This leads to an excessive surplus of water in the valleys surrounding the mountains.

While the compound events focus on the European scale with impacts on socio-economic sectors, the task of concurrent extremes considers the interconnectivities of heatwaves and droughts on a global scale. The combination of heatwaves and droughts are often of interest, as they are strongly interrelated by their physical nature and have been shown to cause significant impact in many socio-economic sectors (e.g., Zampieri *et al.*, 2017; Toreti *et al.*, 2019a). When investigating these dependencies, it is essential to obtain a long dataset for the computation of dependence statistics. However, the nature of climate change complicates this approach, as it affects both heatwaves and droughts, making it challenging to distinguish the dependency signal from the warming signal. Therefore, methods that account for the non-stationarity of these two phenomena need to be investigated.

Chapter 6 is structured as follows: Section 6.2 discusses the data and methods, which have been more intensively discussed in D2.3; Section 6.3 covers the analysis on compound events; and 6.4 addresses concurrent extremes. These findings are then summarised in 6.5.

6.2 Data and methods

For all meteorological variables, the ERA5 reanalysis (Hersbach et al., 2020) is utilised, except for Sea Surface Temperature (SST), which is derived from the OISST dataset (Huang et al., 2021). One of our experiments benchmarks the Standardised Precipitation and Evapotranspiration Index (SPEI; Vicente-Serrano et al., 2010) against a new nonparametric version. For this analysis, we use the CRU TS 4.07 dataset (Harris et al., 2020), as it has been employed in previous evaluations of the SPEI (Beguería et al., 2014). Quality-controlled data from the AGRI4CAST Portal (https://agri4cast.jrc.ec.europa.eu/DataPortal/Index.aspx) is used as dataset for the impact on the food sector, while data from the European Network of Transmission System Operators for Electricity (ENTSO-E; https://www.entsoe.eu/) is sourced for the impact on energy sector. To consider impacts in the water sector, data from the high-resolution pan-European hydrological analysis (HERA; Tilloy et al., 2024) is employed, with temperature and total precipitation data obtained from the EMO-5 meteorological dataset (Thiemig et al., 2022). Lastly, soil data is sourced from the Land Data Assimilation System (LDAS; https://ldas.gsfc.nasa.gov/gldas/soils). A detailed description of the datasets can be found in D3.1 and D3.2.

All the methods used for this Deliverable are reported in detail in D2.3, and we give below only a very short description.

Soft-Dynamic Time Warping (SDTW)

Soft-Dynamic Time Warping (SDTW; Cuturi and Blondel, 2017) offers a flexible approach for clustering multivariate time series by capturing temporal evolution and non-stationarities, both of which are commonly found in climate data due to climate change.

Sure Independence Screening (SIS)

Sure Independence Screening (SIS; e.g., Fan *et al.,* 2020) is used to filter out irrelevant variables from large datasets, in combination with the Reflection via Data Splitting (ReDs; Guo et al., 2023) method. The SIS approach ensures that features with predictive power are retained with a probability tending to one, while ReDs controls the rate of false discoveries or the false identification of inactive features with a user-given probability.

Kernel Regularized Generalized Canonical Correlation Analysis (KRGCCA) and Preimages

Kernel Regularized Generalized Canonical Correlation Analysis (KRGCCA; Tenenhaus et al., 2015) is applied to extract dominant components from multiple high-dimensional climate variables, thereby effectively processing spatial dependencies while maximizing the non-linear relationships between those components. Similar to PCA, KRGCCA generates time series for each input variable, which are associated with spatial patterns. However, unless a linear kernel is applied, the spatial patterns are typically constructed in higher-dimensional spaces, making them difficult to visualize in the usual way. To overcome this, preimages (as described by Honeine and Richard, 2011) are introduced, providing approximate spatial patterns of KRGCCA within the same dimensions as the original feature space.

Imbalanced Random Forests

We use imbalanced random forests based on the q*-classifier (O'Brien and Ishwaran, 2019) to model imbalanced datasets. This method develops an optimized decision rule for classifying imbalanced data, focusing on the likelihood of observing the minority class. Following the recommendation of O'Brien and Ishwaran, 2019, we tune the forest using the G-mean, which combines the sensitivity (True positive rate; TPR) and specificity (True negative rate; TNR) via

$$G - mean = \sqrt{sensitivity \ * \ specificty}$$
(6.1)

Where TPR is defined as TPR=TP/(TP+FN) and TNR=TN/(TN+FP), with TP denoting true positives, TN true negatives, FP false positives and FN false negatives. At the same time, we employ the Accuracy metric, defined as (TP+TN)/(TP+TN+FP+FN)

AI-Enhanced Inhomogeneous J-Function

The inhomogeneous marked J-function (Cronie and van Lieshout, 2016) is employed to assess whether extreme climate events display clustering, inhibition, or independence, while also accounting for changes in their frequencies due to climate change. To briefly summarize, we refer to D2.3 for further details: suppose we are investigating whether extreme events E_1 in a country C_1 at time points T_1 are influenced by extreme events E_2 in another country C_2 observed at time points T_2. The marked inhomogeneous J-function

$J_{C_2,E_2->C_1,E_1}(\Delta T)$ evaluates whether, for time points separated by a distance ΔT >0, the events cluster ($J_{C_2,E_2->C_1,E_1}(\Delta T)$<1), are independent ($J_{C_2,E_2->C_1,E_1}(\Delta T)$=1), or inhibit each other ($J_{C_2,E_2->C_1,E_1}(\Delta T)$>1). We enhance this method by automating the final interpretation process using AI, allowing it to be applied to large data sets (see D2.3 for more details).

Non-Parametric SPEI and Climate Indices
A non-parametric kernel-based estimator is used to overcome the limitations of traditional distributions, such as the log-logistic, in the context of water balance and climate indices like the Standardized Precipitation and Evapotranspiration Index (SPEI; Vicente-Serrano et al., 2010). This estimator is unbounded, data-adaptive, and capable of managing a broader range of climate variability and extremes.

Quantile Regression using I-Spline Neural Networks and Accumulated Local Effects (ALE)
Quantile Regression using I-spline neural networks (QUINN; Xu and Reich, 2023) is employed to estimate the conditional distribution of desired climate variables. These models utilize I-splines and Bayesian neural networks to quantify uncertainties, and the impact of covariates on different parts of the distribution can be analysed through Accumulated Local Effect (ALE) plots (Apley and Zhu, 2020). ALE plots help visualize how predictions change when input features vary, aiding in the assessment of key drivers in climate data analysis.

6.3 Compound Events

6.3.1 Relatively wet and warm late winters followed by dry and warm springs

This compound event focuses on agricultural impacts, specifically on winter soft wheat in France, the largest producer of this crop in Europe. The underlying mechanism involves an anomalously wet and warm winter that triggers early crop growth, which is subsequently harmed by hazardous events, like frost or drought. Such phenomena are commonly referred to as *false-spring events*, and they are expected to increase in frequency due to warming of winter temperatures (Ault *et al.,* 2013). We define the late winter period as January and February, and spring as March, April, and May and for each of these months we obtain the climate variables, which we describe below. To identify which period is crucial for predicting crop failures, we perform a large-scale analysis, aimed at identifying the key months and climate patterns. Based on these identified patterns, we establish a definition of the compound events using machine learning techniques.

6.3.1.1 Identification of large-scale patterns

Adaptation measures, as the improvements in agricultural practices, primarily affect long-term crop measurements and those signals need to be removed to isolate the impact of climate-related events on agriculture. It is common to assume that the signal related to

adaptation is primarily reflected in the long-term trend (e.g., Ceglar et *al.,* 2016; Zampieri *et al.*, 2017). In order to obtain the so-called yield anomalies, we first log-transform the yield data to address heteroskedasticity (Lobell, 2013; Ceglar *et al.*, 2017) and then we apply non-linear detrending. This last action is performed using local polynomial smoothing with an automatic bandwidth selection method (Feng *et al.*, 2020), which accounts for serial correlation. With the aim of modelling the impact of wet and warm winters two indices are considered: the Nonparametric Standardized Precipitation and Evapotranspiration Index (NPSPEI), which captures wet and dry conditions and the Standardized Active Temperature Sum (SATS), which is a temperature-related metric for crop growth (see D3.2). Following the methodologies explained in D3.2 and D4.1, we include large-scale variables to ensure physical consistency:

1. Relative humidity at 700 hPA, an indicator of drought and precipitation.
2. Temperature at 850 hPA to capture additional heat conditions.
3. The 500 hPA geopotential height, which represents large-scale atmospheric circulation, including blocking highs.
4. SSTs, which influences seasonal predictability and has a strong effect on heatwaves and droughts (Domeisen *et al.*, 2022).

The large-scale patterns linked to yield failures are analyzed using the non-linear dimension reduction method KRGCCA. This method simultaneously reduces the dimensionality of multiple variables, generating time series or components for each, while maximizing their dependencies. The user can specify which relationships to maximize through a hypothesis matrix, which we choose such that KRGCCA is able to identify relevant climate patterns linked to yield anomalies (Appendix A6.1). The corresponding climate and yield patterns are then captured by the components. We note that the dependencies between the climate variables themselves are also indirectly considered (see D2.2). To obtain the spatial patterns associated with these components (similar to PCA; see D2.3), we compute preimages. The first three components (time series) along with their spatial patterns for the yield anomalies are presented in Figure 6.1.

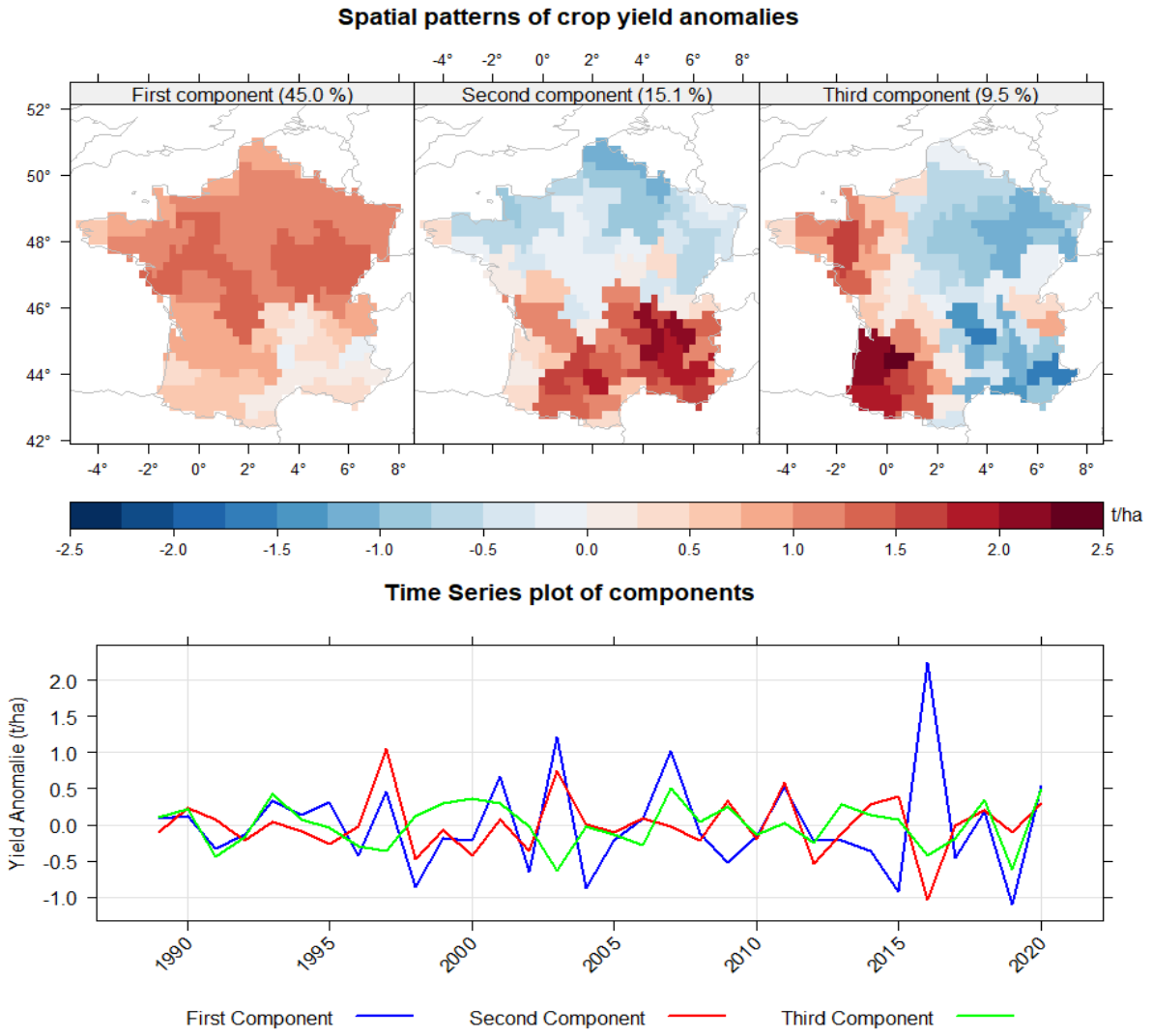


Figure 6.1: *Retained spatial pattern of soft winter wheat anomalies (top panel) and corresponding time series (bottom panel).*

In Figure 6.1 crop yield anomalies are negatively oriented, implying that positive values (red) of the spatial patterns correspond to crop failure and negative ones (blue) represent surpluses of crop yield. In years with positive values, the corresponding spatial patterns (see Figure 6.1) are "active", with larger values indicating an increased intensity of the pattern. For years with negative values, the pattern is active in a negative phase and can be visualized by multiplying by -1 the values displayed in the top panel of Figure 1. The first pattern reveals a signal related to crop failure mainly throughout the northern and central part of France. The corresponding time series in the bottom panel of Figure 6.1 shows that this pattern dominates the other two components by accounting for approximately 45.0% of the variance and is linked to significant crop failures, including those in 2003, 2007, and the severe failure in 2016 (Ben-Ari *et al.*, 2018). The second component is dominant in the

Mediterranean region, where the signal from the first component is weaker. In contrast, the third component predominantly reflects failures in the southeastern region. Crop failures related to these three components do not always occur simultaneously, as it can be seen for 2003 (failures signalled by first and second component), and 2016 (only first component).

In the next steps, we inspect which climate variables play the most important role in explaining the patterns of yield anomalies determined above. Our analysis will primarily focus on modelling the first component, since it explains the majority of the variability. KRGCCA has identified patterns for NPSPEI-1 and SATS-1, which are linked to the yield patterns in Figure 6.1 with corresponding time series (similar to PCA analysis; see D2.3 for a detailed description). We can use them to model the yield components, so that we can identify which climate variables (and corresponding months) have the most influence on the observed yield anomalies. In order to do so, we employ a vine copula-based quantile regression (Kraus and Czado, 2017), with which we can model the non-linear relationships and capture the conditional distribution function (CDF) of the first component of the yield anomalies (Figure 6.1). This allows us to focus on the tails that correspond to significant crop yield failures. In order to avoid overfitting, Kraus and Czado (2017) suggest the use of information criteria, which balances model complexity, variable selection and predictive performance. Using the Bayesian Information Criterion (BIC; Schwarz, 1978) the model selects three variables: NPSPEI-1 in April, SATS-1 in February and SATS-1 in May. With those variables the CDF is found to be standard uniformly distributed (Anderson-Darling Test p-value: 0.90) indicating that the Probability Integral Transformation (PIT) works well and the CDF is well modelled.

Among these patterns, the ones that play a pivotal role in predicting significant crop yield failures are found using variable importances based on ALE plots in conjunction with the vine copula regression. We recall this method allows to focus on higher conditional percentiles. However, the relatively small sample size (N=32) prevents from using very high percentiles for this analysis. By consequence we focus on the 70th conditional percentile as an indicator of high impact. We construct the preimages (i.e., the approximative spatial patterns; see D2.3) in February, April, and May for the selected variables. Additionally, we include January, which is found to be quite similar to February. The variable importance criteria and the reconstructed preimages are presented in Figure 6.2.

The most influential variables for predicting significant crop yield impacts are NPSPEI-1 in April followed by SATS-1 anomalies in May and February. The corresponding preimage, generated for the positive phase of the first component, reveals wet and warm conditions in January and February which could initiate early crop growth. The subsequent dry and warm conditions in April can cause significant damage, and when followed by warm conditions in May, the impact could be further intensified. To better understand the combined role of these hazards, we examine the marginal and interaction effects of NPSPEI-1 in April and SATS-1 in May, as shown in Figure 6.3.

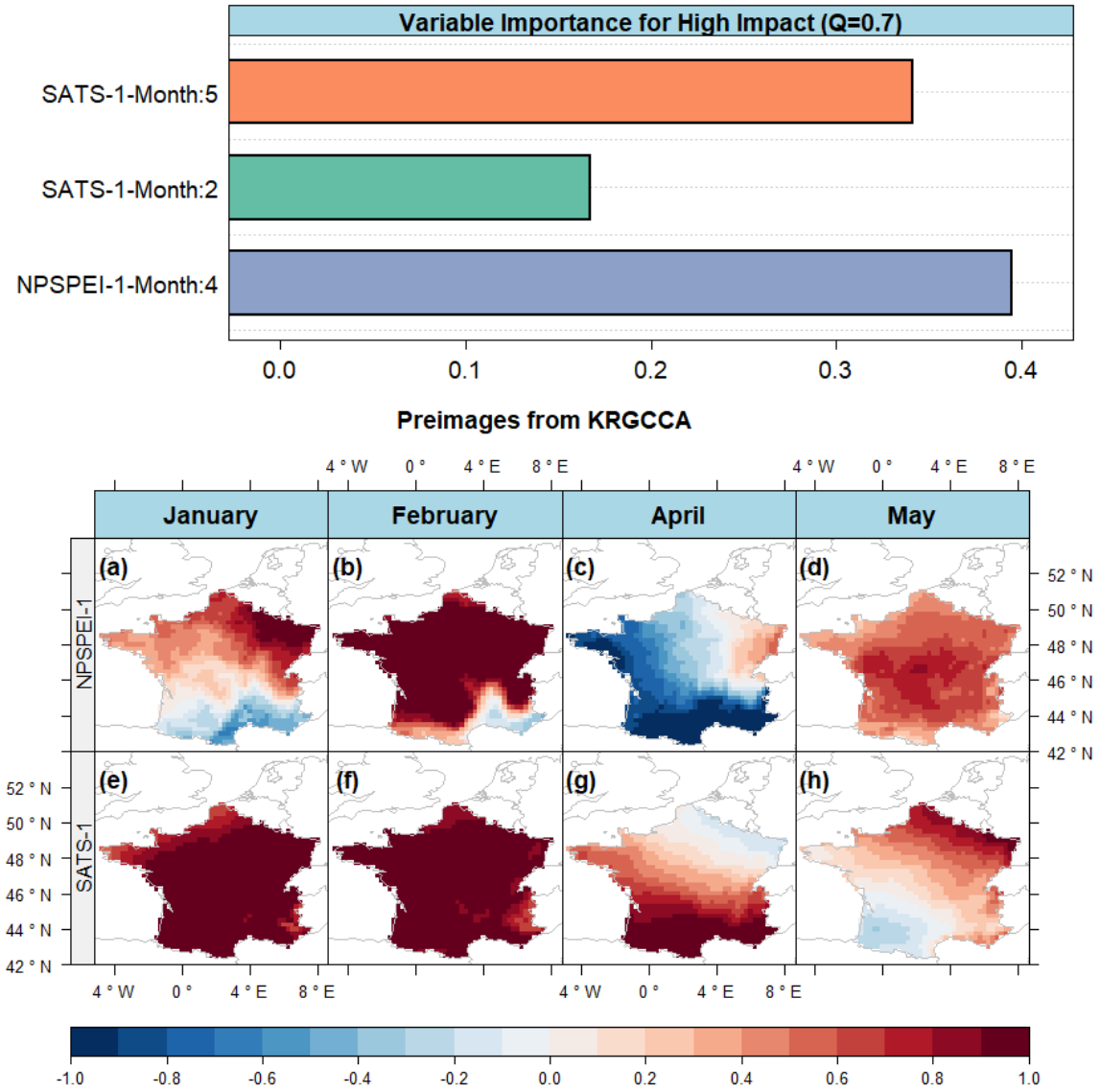


***Figure 6.2**: Top panel: variable importance obtained for the 70th conditional quantile of the crop yield anomalies using the vine copula-based quantile regression model. Bottom panel: preimages spatial patterns for the NPSPEI -1 and SATS-1 corresponding to the first component of the crop yield anomalies displayed in Figure 6.1.*

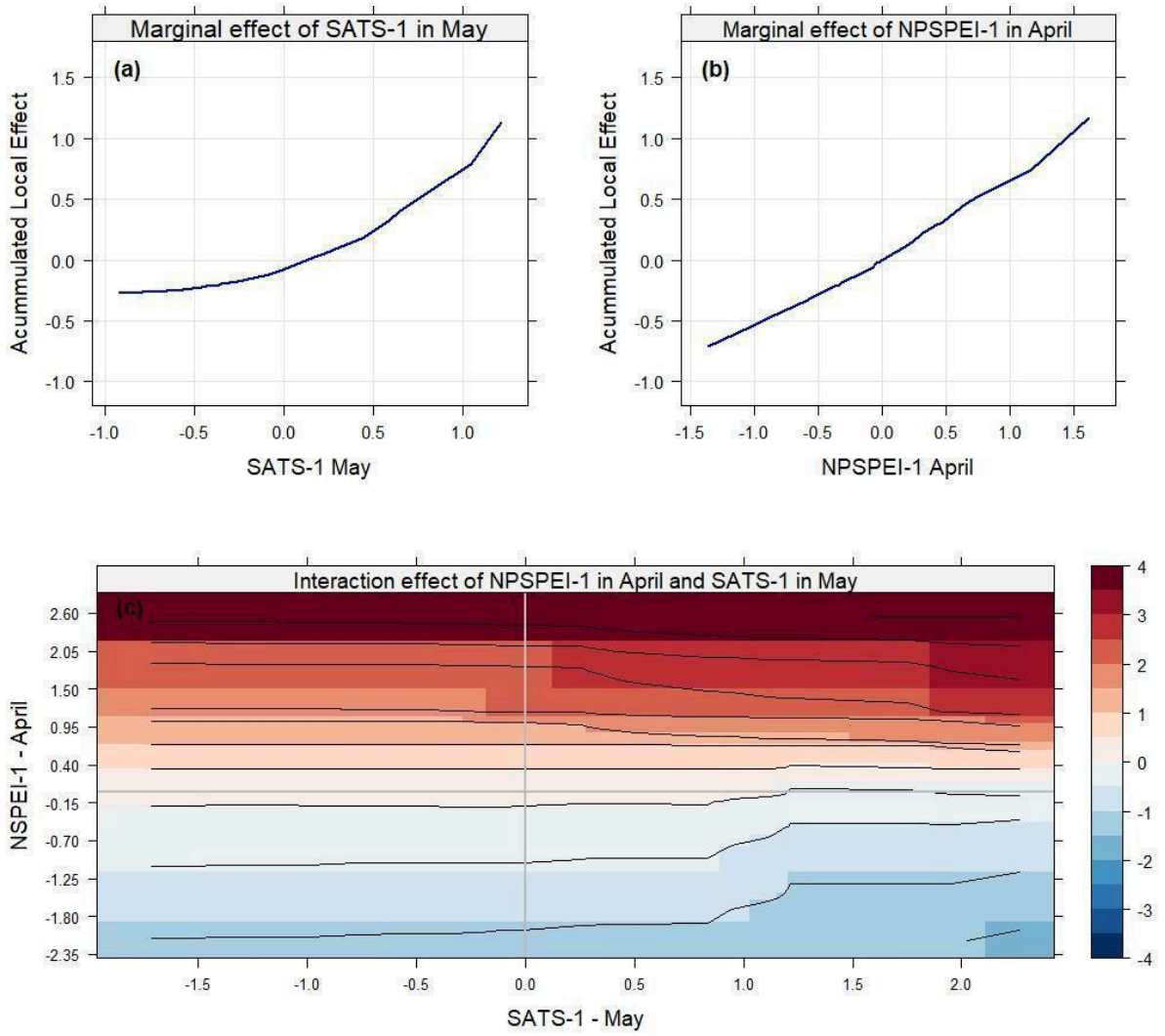


***Figure 6.3**: Marginal effects of warm May (a) and spring drought (b) on yield anomalies. Interaction effect of NPSPEI-1 and the SATS-1 in May is shown in (c).*

Focusing on the positive phase of the first KRGCCA component, which is linked to negative NPSPEI-1 values and dry conditions (Figure 6.2c), we observe a near-linear increase in ALE, indicating a corresponding rise in crop losses. Positive NPSPEI-1 values suggest that the spatial pattern in Figure 6.2, reflecting mainly dry conditions, is "active." The higher the NPSPEI-1 values, the higher the ALEs, meaning larger crop losses are expected due to the strongly pronounced droughts. For the negative values or phases of NPSPEI-1 in April, the spatial pattern has to be multiplied by -1, indicating wet conditions. In these cases, the negative ALEs suggest that losses are less likely, due to the more favourable wetter conditions. For SATS-1 in May, the ALEs show a non-linear response, particularly as they shift into the positive phase (i.e., values above zero), corresponding to warmer conditions across most of France, except for the southwestern region (Figure 6.2h). This implies that yield

losses mainly occur when May warmth is strongly pronounced, while cooler May conditions (negative SATS-1 values) have little influence, as the ALE values remain close to zero.

The interaction plot (Figure 6.3c) further reveals that the highest ALE values are typically found when the April drought is particularly severe, with the first component related to NPSPEI-1 above 2.0. These high ALE values occur almost independently of SATS-1 in May, suggesting that the April drought is the dominant hazard. However, when high SATS-1 values in May coincide with a moderately positive NPSPEI-1 phase, significant ALE values can still occur, indicating that warm May conditions can amplify the impact of the April drought, when it is less severe.

6.3.1.2 Construction of objective thresholds

The dimension reduction experiments explained in 6.3.1.1 have provided valuable insights into the most important months with the highest predictive power for crop yield losses. However, the specific intensity required for these components to trigger significant crop yields impacts remains unclear. To explore this, we focus on the local scale, and we model the impact on winter soft wheat using the climate variables for each NUTS3 region for France. Figure 6.1 suggests the existence of three clusters related to impacts. However, since KRGCCA assumes continuity across the field, precisely defining homogeneous agro-climatic regions is difficult.

To address this challenge more objectively, we apply the multivariate Soft Dynamic Time Warping (SDTW; Cuturi and Blondel, 2017) algorithm, which defines clusters based on winter soft wheat anomalies, SATS-1 and NPSPEI-1 for the most important months. The selection of these climate variables is guided by the results shown in Figure 6.2: first, with the aim of capturing warm and wet late winter conditions, we identify respectively SATS-1 and NPSPEI-1 in February as key variables. Although we have primarily used an aggregation period of one month so far, the spatial patterns in Figure 6.2 indicate that January and February share similar characteristics. This suggests that using an aggregation period of two months (i.e., NPSPEI-2 and SATS-2 in February) could be beneficial for summarizing wet and warm conditions of late winter. Statistically, higher aggregation periods often lead to more robust indices (Vicente-Serrano and Beguería, 2016). Secondly, in order to capture hazards associated with dry and warm spring conditions, we use NPSPEI-1 in April (multiplied by -1) and SATS-1 in May. All these climate variables, along with yield anomalies, are used in multivariate clustering through the SDTW approach. Hyperparameter tuning is performed using the Silhouette coefficient, which has shown strong performance in various clustering scenarios (Arbelaitz *et al.*, 2013). The resulting clusters are presented in Figure 6.4.

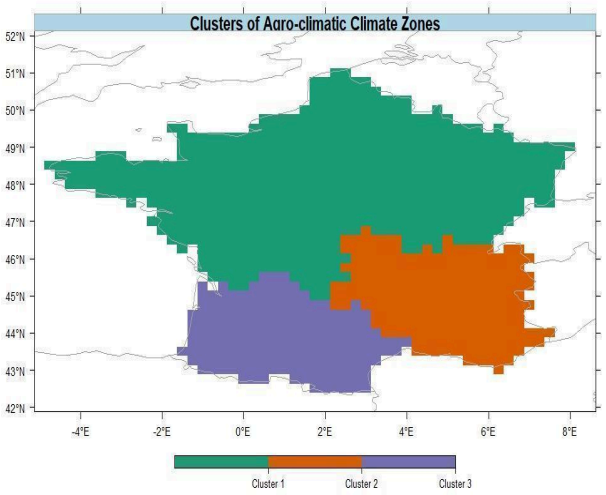


***Figure 6.4**: Obtained agroclimatic regions used for grouping winter wheat crop yield anomalies and the local climate variables from February to May described in the text.*

We observe that the determined clusters closely resemble the three regions identified in Figure 6.1, indicating the robustness of their definition.

Based on the variables we have chosen above, we define a compound event as follows:

1. SATS-2>0 in February, to define warm winter conditions.
2. NPSPEI-2>0 in February, to define wet winter conditions.
3. −NPSPEI-1>0 in April, to define the spring drought.
4. SATS-1>0 in May, to define warm conditions in May.

The choice of these thresholds comes from the definition of the indices: positive (negative) values of NPSPEI-1/NSPEI-2 coincide with wet (dry) conditions, while positive (negative) values of SATS-1/SATS-2 reflect warm (cool) conditions.

The four conditions listed above constitute altogether a meteorological event. Since in compound event research it is essential to include a high socio-economic impact feature (Zscheischler *et al.*, 2018), negative yield anomalies above the 70th percentile (chosen for the reasons explained in 6.3.1.1) are added to the requirements listed above. If all of these conditions are met, we label the event as a compound event. Next, we use decision trees to learn the definition outlined above solely from climate variables. The rationale behind this is that the definition imposes minimal conditions on wet/dry and warm/cold states, along with the impact. The decision tree should thus find thresholds for these climate conditions to reconstruct the impact, thus giving our desired objective thresholds.

As a consequence of the stringent conditions required for this definition, the classification problem is imbalanced, and using a single decision tree leads to poor performance. To

address this, we utilize an imbalanced Random Forest (RF) model based on the q*-classifier (O'Brien and Ishwaran, 2019). However, training such a model for each NUTS3 region is challenging due to the limited number of observations. For this reason, it is advantageous to group the variables according to the agro-climatic zones identified through the clustering process (Figure 6.4). In this study we focus on cluster one, since it is similar to the first component of KRGCCA (Figures 6.1.1 and 6.1.2). The KRGCCA analysis performed in section 6.3.1.1 has indicated that the patterns of SATS, NPSPEI and yield anomaly are quite similar over cluster one, so it is reasonable to assume that the link function is approximately similar throughout the corresponding NUTS3 regions. This allows us to stack the data within these regions in a panel regression setup, thus increasing the available data for model training. Next, we train the imbalanced RF model based on the q*-classifier and perform hyperparameter tuning using 10-fold stratified cross-validation, optimizing for the G-mean, as recommended by O'Brien and Ishwaran, 2019. The obtained accuracies for the test set, defined as the period 2010-2020, are shown in Table 1.

***Table 6.1**: Scores obtained from the test set (2010-2020) using the imbalanced RF approach.*

Random Forest Model		Surrogate model – Decision Tree
G-Mean	0.941	0.940
Accuracy	0.935	0.931

We observe that the G-Mean and accuracy are satisfactory, indicating strong model performance and suggesting that the setup successfully captured the definition of a compound event. Nevertheless, examining the decision thresholds of such a high-dimensional model is challenging. To address this, we employ a global surrogate model: specifically, we use the output of the complex RF model described above, and we apply a decision tree to predict its output using the same input features. The “new” decision tree is trained using 10-fold stratified cross-validation, and we tuned the model again with G-Mean based on the q*-classifier. The performance metrics are displayed in Table 6.1 and are found to be satisfactory.

The obtained decision boundaries are presented in Figure 6.5. Before proceeding to the analysis of the results, it is important to recall that all input variables are constructed as (approximately) standard Gaussian. The extracted thresholds are 1.5 for the negative NPSPEI-1 and 1.6 for SATS-1 in April and May. This indicates that relatively strong drought or warm conditions are required to observe significant impacts. Conversely, the thresholds for the winter variables are only slightly above the mean, suggesting that these events, when considered individually, might be relatively harmless. However, when they coincide with drought or heat conditions in spring, substantial impacts can be observed. This demonstrates that combinations of non-extreme events can lead to significant agricultural consequences.

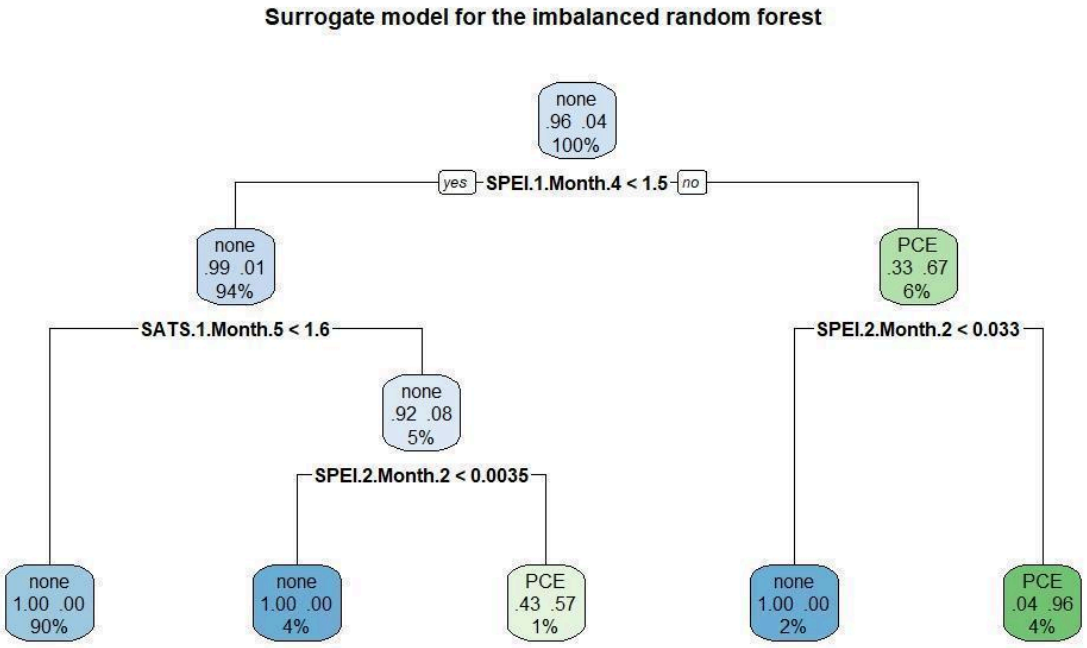


***Figure 6.5**: Global surrogate model for the imbalanced random forest.*

6.3.2 Dry winters followed by hot summers

The next compound event focuses on the superimposition of hot and dry conditions during the summer. Here, we examine how a preceding dry winter may exacerbate these impacts. Since heat in summer is an important factor, we discuss an approach for a heatwave index taking into account maximum and minimum temperature, which might be beneficial for a more holistic impact description.

6.3.2.1 The Bivariate Heat Magnitude Day

A wide range of heatwave indices has been proposed in the literature (see, e.g., Barriopedro *et al.*, 2023), each aiming to capture the complex meteorological characteristics of these events. However, many analyses tend to define heatwaves focusing on daily peak temperatures such as the maximum temperature recorded. A typical definition, for instance, is when the daily maximum temperature exceeds the 90th percentile for at least three consecutive days (e.g., Perkins and Alexander, 2013). While this approach is appealing in its simplicity and has proven effective in various studies, it raises the question of whether focusing solely on peak temperatures is sufficient for a comprehensive understanding of heatwaves. Human stress, for example, is often exacerbated by high nocturnal temperatures or unusually warm night-time conditions, emphasizing the need to consider minimum temperatures as well.

To address this, we propose a multivariate heatwave index that integrates both the maximum and minimum temperatures of the day, offering a more comprehensive definition of heatwaves. Our framework draws from multivariate risk assessment methods, as outlined by Salvadori *et al.* (2016) and considers three scenarios illustrated in Figure 6.6. The two axes in the Figure represent hypothetical thresholds between 0 and 1, conceptualized as quantiles for both maximum and minimum temperature.

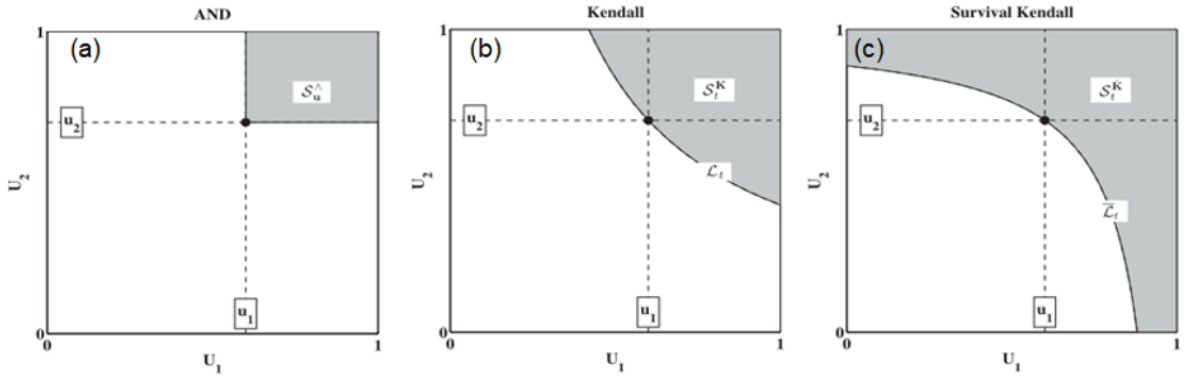


***Figure 6.6**: Schematic Visualisation of multivariate thresholds taken from Salvadori et al. 2016.*

When aiming to identify multivariate thresholds for maximum temperature and minimum temperature, the simplest method is to apply thresholds to both variables as displayed in panel (a) and check for joint exceedances (e.g., both temperatures exceed their 90th percentile, Lavaysse *et al.*, 2018). However, this imposes a stringent constraint on the detection method, requiring both variables to be in an extreme state to be recognized as a multivariate extreme event, which can be potentially problematic. For instance, if maximum temperature exceeds the 98th percentile while minimum temperature is below the 90th percentile, the event is not considered extreme, despite maximum temperature being extremely high and potentially hazardous to certain socio-economic sectors. Therefore, it may be beneficial to allow some tolerance if one variable is in an extreme state while the other remains high, but not extreme. Statistically, this issue can be addressed by examining the dependence structure of the two variables. Utilizing copulas, Salvadori et al. (2016) summarizes methods for identifying multivariate thresholds. The so-called Kendall function and the survival Kendal (SK) function are two possible approaches and are depicted in Figures 6.6(b) and 6.6(c). These scenarios classify an event as multivariate extreme allowing for the possibility that one variable may be extremely high while the other is below the conventional threshold. We have explored the detection of bivariate heatwave indices using three different methods, defining a heatwave event as occurring when both maximum and minimum temperatures exceed the bivariate 90th percentile for at least three consecutive days. Among the approaches considered, the SK scenario consistently produced the best results for our purposes.

While the detection of heatwaves is crucial, it is also necessary to classify the magnitude of these events, as this is most important for assessing impacts. For instance, sectors like agriculture can be significantly affected by large deviations from the mean (e.g., Porter and Semenov, 2005). We adopt the Heat Magnitude Day (HMD) introduced by Zampieri *et al.,* (2017) and construct a Bivariate Heat Magnitude Day (BVHMD, equation 6.1): we first detect the maximum and minimum temperature-based events using the bivariate thresholds defined in the scenarios shown in Figure 6.5 and check if it is exceeded for at least three days. Then, for both variables, we calculate the HMD, resulting in two time series called HMD_{TMAX} and HMD_{TMIN}. To combine these and to find a link function f:R->R such that

$$BVHMD = f(HMD_{Tmax}, HMD_{TMin}). \quad (6.2)$$

We choose f:R->R through a supervised learning model using HMD_{Tmax} and HMD_{Tmin} as predictors and indicators for reflecting socio-economic impact (e.g., crop yield anomalies for agriculture) as response. The function f:R->R is then directly estimated by the model.

6.3.2.2 Nonlinear compound stress indices

After introducing the novel heatwave indices, we shift focus to the impacts of dry winters followed by hot summers on both the agricultural and energy sectors. While the combined effect of heat and drought on agriculture has been extensively documented (e.g., Hao et al., 2022), the study of the energy sector is primarily motivated by drought-related factors. For instance, a reduction in water availability can hinder the cooling systems vital for power generation. Moreover, the interplay between agriculture and energy sectors can result in cascading impacts, such as water shortages limiting irrigation, further exacerbating agricultural losses.

In this chapter, we specifically examine the effects of these climate events on grain maize, the most important summer crop in Europe. The yield data is obtained from AGRI4CAST at the NUTS3 subnational level. To isolate climate-related impacts, yield anomalies are calculated by detrending the yield data, as in the previous chapter. Here, we adopt a parametric detrending approach based on the model:

$$log(yield) = \beta_0 + \beta_1 year + \beta_2 year^2. \quad (6.3)$$

In this chapter, we adopt this approach as we also analyze regions with relatively small sample sizes (e.g., Germany often only having 15 observations per NUTS3 region), unlike the data from France in the previous chapter, which benefits from longer records (N=32 for all NUTS3 region). This detrending method has been shown to perform well compared to nonparametric approaches (Ceglar *et al.*, 2017).

Instead of applying this detrending approach separately to each region, we estimate the model jointly for all regions using the Seemingly Unrelated Regression (SUR; e.g., Fiebig,

2007) approach. SUR offers significant efficiency gains compared to single regression models when their error terms—here, the yield anomalies—are correlated, which is likely the case given the spatial dependency of crop yield failures in response to climate conditions. As displayed in Figure A.6.2 high correlations among the yield anomalies can be observed. Thus, we can expect important efficiency gains of SUR in comparison to single regressions by making use of those spatial correlations, and we utilize SUR for the following two subchapters focussing on agricultural impact.

6.3.2.2.1 Local indices

To analyse the impact of hot and dry summers with preceding dry winters on the agricultural impacts in summer, we examine the effects of these events on grain maize, the predominant summer crop in Europe. Our benchmark index for assessing agricultural impact is the Compound Stress Index (CSI; Zampieri *et al.*, 2017), which employs a linear superimposition of drought and heatwave indicators. To incorporate the stress induced by dry winters, we consider soil moisture levels across four different strata (Layer 1: 0-7cm; Layer 2: 7-28 cm; Layer 3 28 -100 cm; Layer 4: 100 - 289 cm) available in the ERA5 dataset. Furthermore, we utilize our nonparametric NPSPEI, which will be discussed in 6.4.1, and the BVHMD based on the SK scenario to measure climate-related stress during summer.

Following Zampieri *et al.* (2017) we construct the compound stress index as a linear superimposition of the BVHMD, NPSPEI as well as the use the four soil moisture variables, which we standardize to standard Gaussian variables using the same methods as for NPSPEI (see D2.3 and Chapter 6.4). To find the linear model, we adopt the adaptive elastic net (ADNET), which not only addresses correlations between predictors but also includes a shrinking mechanism to prevent overfitting. The ADNET possesses the oracle property (Zou and Zhang, 2009), meaning it can (asymptotically) identify the active features (i.e. those features whose coefficients are not null) with a probability approaching one. We perform hyperparameter tuning by calculating a sequence of models through a regularization path (Tay *et al.*, 2023), and we choose the “best” model among those using the BIC as information criterion, ensuring the oracle property is maintained (Fan *et al.*, 2020). The final retained model provides our benchmark index for forecasting. Furthermore, we employ the same pre-processing steps as in Zampieri *et al.*, 2017.

While the above approach has compelling properties—such as the oracle feature, data-driven variable selection, and interpretability—it also presents certain drawbacks. The linearity assumption implies that the effects of variables remain constant across their range, which is less realistic, as crops may exhibit heightened sensitivity to threshold exceedances or abrupt spikes in climate variables (Lavaysse *et al.*, 2018). Additionally, the standard regression setting provides only a single forecast. To address these limitations, we apply a vine-copula-based regression, which models the CDF of crop yields given the input variables by decomposing them into vine copulas, offering several advantages. By modelling the full

distribution, we can directly quantify uncertainty, assess the effects of input variables across different parts of the distribution, and account for non-linear relationships via the copulas.

Model selection for the vine-copula based regression is performed using an information criterion. We utilize the Akaike Information Criterion (AIC; Akaike, 1974), as recommended by Kraus and Czado, 2017. The analysis is conducted across NUTS3 regions in Europe, with grain maize data and climate data aggregated from ERA5. To evaluate the performance of both models, we compute the R^2 values for the models. Due to the highly uneven number of observations per NUTS3 region, we limited our analysis to regions with at least 30 observations, encompassing regions in Austria, Portugal, and France utilising the overlapping period of 1989-2020 for all regions. The results are shown in Figure 6.7.

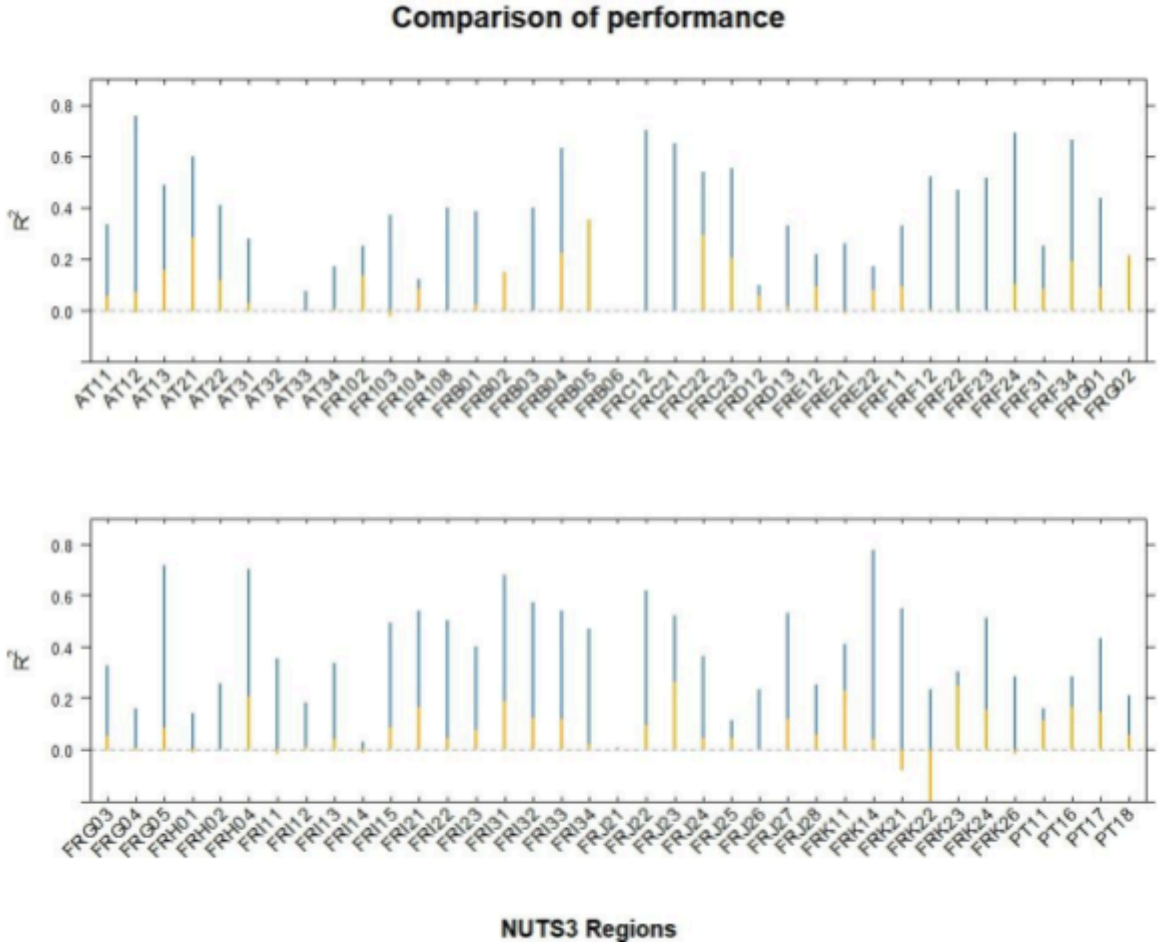


***Figure 6.7**: Explained Variance by regressing the nonparametric indices on grain maize anomalies using d-vine copula-based quantile regression. Blue lines correspond to the explained variance (R^2) of this model, while the superimposed orange lines describe the increase of the latter in comparison to the CSI. The x-axis displays the evaluated NUTS3 regions.*

We observe that most of the R^2 values tend to increase, with only a few exceptions, demonstrating the added value of accounting for non-linearities in the proposed index. To verify whether the model accurately captures the CDF of the grain maize anomalies, we use the PIT, where the predicted conditional probabilities should follow a standard uniform distribution. We applied the Anderson-Darling test to each model, and all returned p-values greater than 0.1, indicating no significant deviations from uniformity, suggesting that the CDF is also well modelled.

There are two main advantages of our index. First, by modelling the full CDF of grain maize anomalies, we can target specific aspects of the distribution, such as the upper tails, which correspond to high-impact events. This feature also enhances our ability to estimate uncertainties. Second the index captures non-linear relationships between inputs and output and those are nonparametrically modelled thus making minimal assumptions about the functional form.

6.3.2.2.2 A large-scale model

While the models introduced in the previous chapter show promising attributes, their applicability is limited to local scales, where a sufficient number of observations exists. Unfortunately throughout Europe the amount of observations is quite inhomogeneous as shown in Figure A6.3 and the experiments performed in the previous chapter cover only roughly 58 % of observations. To address this, we aim to develop a large-scale model capable of capturing the dynamics across most of the NUTS3 regions in Europe. For this purpose, we organize the agricultural data into a vector format: data from all NUTS3 regions are combined into a single large vector, forming a 1-D time series within total N=3415 observations. The same is done for the climate variables, which are the same as in the previous chapter, such that each column of the resulting predictor matrix corresponds to climate features for each NUTS3 region. The goal is to use a neural network to predict yields from these given covariates and we utilize the Quantile Regression using I-Spline Neural Network (QUINN; Xu and Reich, 2023).

However, to accurately reproduce these patterns, we should incorporate additional characteristics to the predictors that describe both spatial and temporal information. To include temporal information, we add the year of the yield for each NUTS3 region. Spatial information is incorporated by adding soil data of LDAS to describe agricultural conditions across Europe. For further spatial insights, particularly with respect to climate, we construct agricultural climate zones using the SDTW approach (Cuturi and Blondel, 2017). This method is well-suited for our case, as it accommodates different time series lengths and multiple variables, allowing us to cluster yields and climate variables jointly across all NUTS3 regions as we did it in chapter 6.3.2.1. The number of clusters is again determined using the silhouette coefficient, resulting in ten clusters (not shown). To characterize the compound nature of the events, we include dummy variables that indicate when associated heatwave and drought conditions are present. Since the soil moisture variables and NPSPEI are

normalized to standard gaussian variables, we apply a threshold of -1, commonly used to define extreme conditions, and characterize heatwaves by checking if the HMDs for maximum and minimum temperature exceed zero. This information will later be used to calculate the likelihood of crop failures under different combinations of compound events or extremes.

We train the model, perform hyperparameter tuning with the Widely Applicable Information Criterion (WAIC; Vehtari *et al.,* 2017), and use the No-U-Turn Sampler (NUTS; Hoffman and Gelman, 2011) algorithm with 2000 warm-up steps followed by 8000 iterations, discarding every eighth sample to retain 1000 MCMC samples. Figure 6.8, panels (a) and (b), display the Q-Q plots of the ensemble mean and the full ensemble, both of which show very satisfactory results, indicating that the model adequately captures the distribution of the yields. When comparing model performance in terms of R^2 against the two indices, there is a clear advantage over the classical CSI. However, the non-linear CSI introduced in chapter 6.3.2.2.1 surpasses the QUINN in 40 % of the cases (Figure 6.8, panel (c)). Thus, for local decision-making, the non-linear CSI may offer better insights in certain cases. Nevertheless, both indices significantly outperform the classical CSI.

An advantage of the QUINN model is that it allows us to examine how different variables influence various parts of the distribution, enabling us to compute variable importance criteria through ALE. The resulting importance values are displayed in Figure 6.8 choosing a couple of representative percentiles to assess the impact of the climate variable on the distribution of yields.

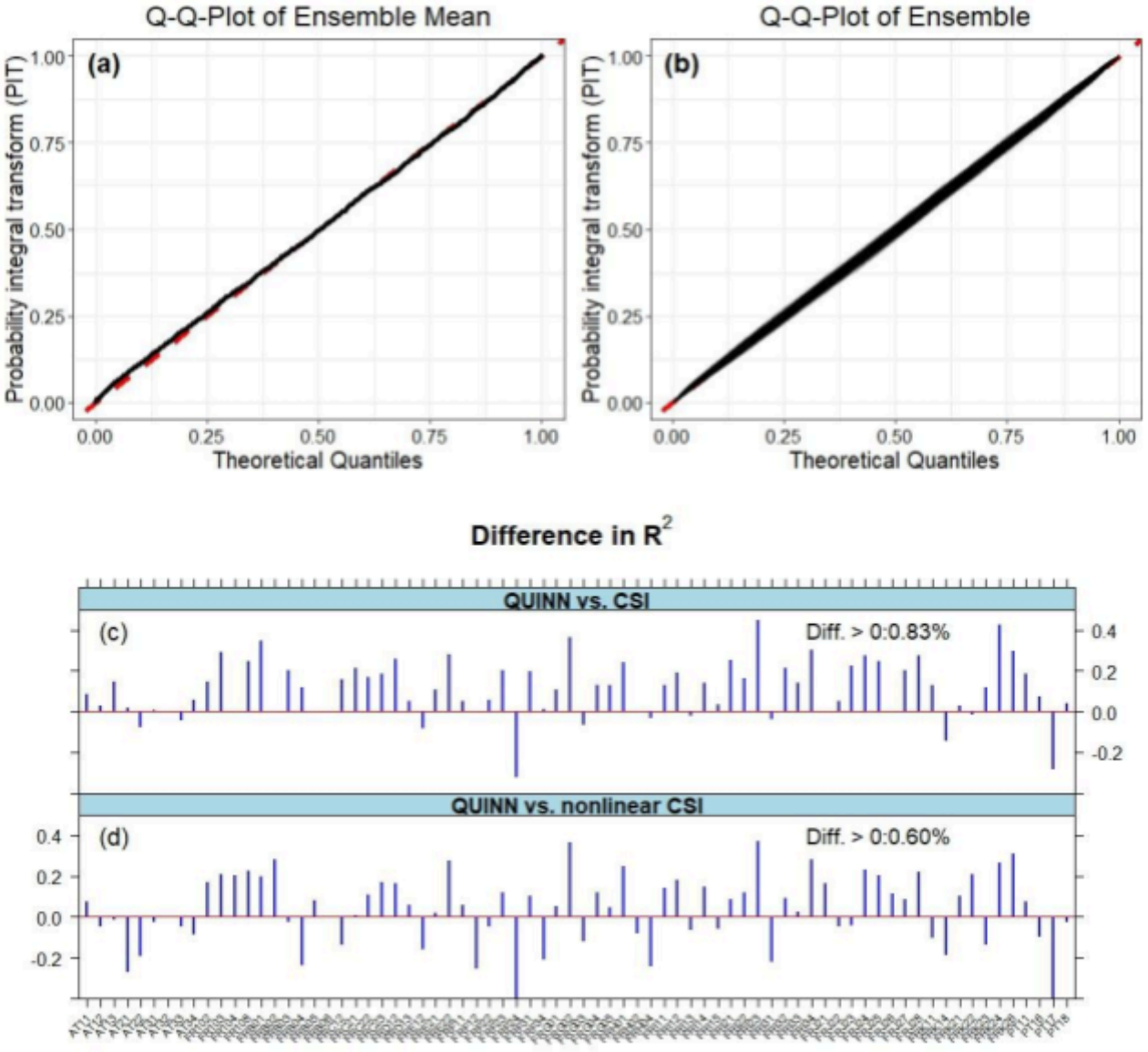


***Figure 6.8**: Model diagnostics from the QUINN model used for predicting grain maize anomalies. Panel (a) and (b) show Q-Q-plots of the model and (c) and (d) compares the QUINN prediction with the CSI and non-linear extension (Section 6.3.2.2.1).*

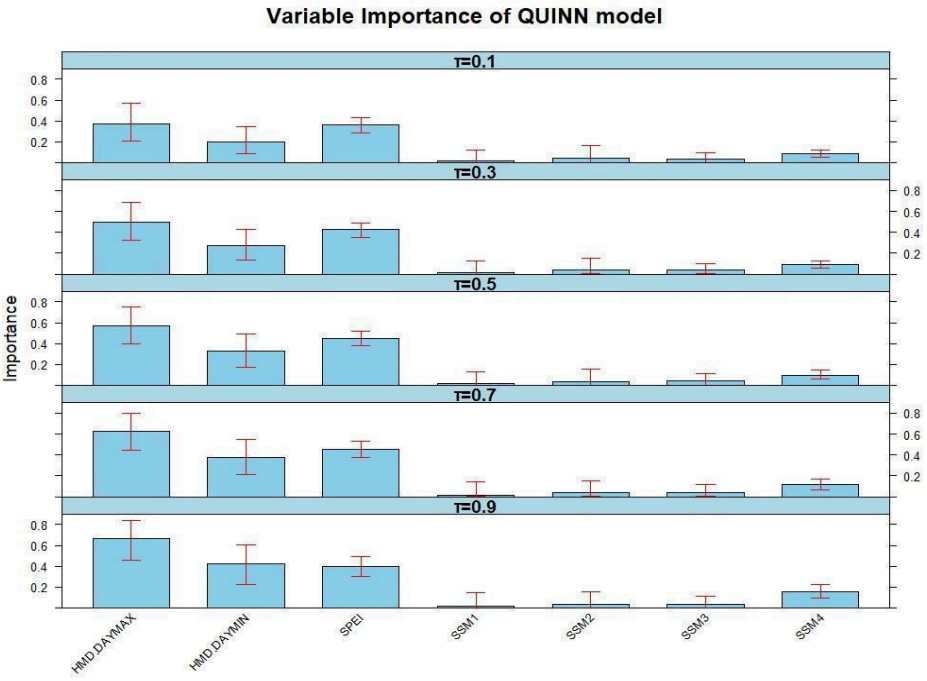


***Figure 6.9**: Variable importance for the grain maize based QUINN model utilizing the 90th Percentile. SSM denoted soil moisture in layer 1, 2, 3 and 4.*

We observe that the HMDs based on maximum and minimum temperature, and the NPSPEI are the most influential variables across all percentiles. Hence, the inclusion of minimum temperature as discussed in chapter 6.3.2.1 seemed to be beneficial for predictive purposes. However, the HMD based on maximum temperature becomes particularly important in the upper percentiles, which correspond to larger crop yield failures. A similar pattern is observed for the soil moisture in layer four, indicating that this is the most critical soil moisture layer.

To assess the impacts of the combination of these variables, we compute the likelihoods of crop failures under various configurations of the input variables, with a particular emphasis on the increased risk posed by compound events. Specifically, we calculate the probabilities of observing the following scenarios:

1. A crop failure when none of the climate variables indicate an extreme event (REF).
2. A crop failure during a heatwave, based on maximum temperatures (HW).
3. A crop failure during a heatwave combined with a summer drought (HW-DR).
4. A crop failure during a heatwave, where the fourth layer of soil moisture experiences anomalous dry conditions, but we do not condition on a summer drought (HW-WIN).
5. A crop failure during a heatwave and a summer drought, with the soil moisture layer four experiencing anomalous dry conditions (HW-DR-WIN).

A crop failure is defined as occurring when the yield anomaly exceeds the 70th percentile, consistent with the definition used in the previous chapter, and the results are shown in Table 6.2.

***Table 6.2**: Likelihood of grain maize crop yield failures in each event. The Reference ratio is defined as the median likelihood (column 1) of observing the desired event, divided by the likelihood of observing a failure in the REF scenario.*

Event	Likelihood Crop Fail.	Credible Interval	Reference Ratio
REF	20.38	[14.42, 27.10]	1.00
HW	31.19	[21.70, 41.50]	1.53
HW-DR	62.84	[49.70, 74.66]	3.08
HW-WIN	46.25	[34.52, 58.02]	2.26
HW-DR-WIN	69.42	[56.57, 80.30]	3.41

We observe that the likelihood of crop failure during a heatwave is approximately 1.5 times higher than under normal conditions. When preceded by a dry winter, this likelihood increases to more than double (2.26). The worst-case scenario, however, arises when a summer drought and heatwave coincide, particularly if preceded by a dry winter, resulting in a three- to three-and-a-half-fold higher chance of crop failure—double the risk observed during heatwaves alone. Hence, compounding conditions are typically met with the highest risks, although the summer variables seem to play the most important role for agriculture.

6.3.2.3 Impacts on the energy sector

In this chapter, we focus on the impacts of hot and dry summers on the energy sector, utilizing data on power outages reported by operators to the European Network of ENTSO-E. The dataset includes unplanned outages that occurred within the European bidding zone between 2015-12-31 and 2022-03-31, excluding scheduled maintenance. To characterize the climate conditions, we use winter soil moisture anomalies to represent dry conditions, and total precipitation, maximum temperature, and minimum temperature to capture hot and dry conditions during the summer. We convert all climate variables to anomalies using the

1991-2020 reference period, the only common reference period overlapping with the outages observations.

To analyze how climate affects these outages, we first prepare the outages for impact estimation by labeling each day as either 0 or 1, with 1 indicating that an outage is observed during that day. Hence, we can try to identify the climate conditions corresponding outages (or labels of one) indirectly assuming, however, that the relevant climate conditions persisted for the entire day. This assumption is reasonable, given that heatwaves and droughts typically last for several days (in the case of heatwaves) or even weeks to months (for droughts), and both phenomena are generally characterized by persistent conditions (e.g., Zargat *et al.,* 2011, Perkins and Alexander 2013). Our focus is on the summer months (June, July, August).

However, after encoding the data in this way, we found that the number of outages is relatively low, with imbalance ratios often below 1%. This makes it challenging to estimate a model with such low occurrence rates. To address this issue, we aggregate the data on a weekly basis by summing the number of outages, enabling us to model the data using Poisson or negative binomial distributions. We opt for the negative binomial distribution, as it is not sensitive to overdispersion compared to the Poisson model (e.g., Warton *et. al,* 2012).

In the next step, we aim to obtain structural insights into the spatial and temporal patterns of the outages by employing dimension reduction techniques. Since we are working with count data, we apply the generalized linear model PCA (glmPCA; Townes, 2019; Townes *et al.*, 2019), which extends the classical PCA approach to generalized linear models (GLMs). The model creates real-valued principal components and eigenvectors, which can be interpreted like in the classical PCA. This method also allows the inclusion of covariates, and we use for the week of the year account for seasonal effects, enabling the extraction of anomalies related to outages. The glmPCA shows good convergence and through graphical evaluation, we retain the first three components, as higher components showed minimal variability (not shown).

After identifying the dominant large-scale patterns for energy outages, we seek to determine the corresponding meteorological patterns. To achieve this, we aggregate the climate variables on a weekly scale using the European domain (Figure 6.10) by building weekly sums for precipitation and means for the other variables. Since, the number of power plants is limited over Europe, we likely have a lot of inactive features for the climate covariates if we take the full European domain. To account for the latter, we perform feature screening through Sure Independence Screening (SIS), based on the projection correlation (Zhu *et al.*, 2017; Liu *et al.*, 2022), which is suitable for GLMs, for each climate variable utilizing the three principal components from glmPCA as output. Having obtained those active features, we employ KRGCCA with the same setup as in Chapter 6.3.2.1. Considering

the used kernels, we employ the Gaussian kernel for the climate variables, with bandwidths derived from the method in Chaudhuri et al. (2021) and we apply a linear kernel for the outages principal components.

Figure 6.10 (a) displays the spatial pattern of the first outage component, explaining 89% of the variability of the latent features extracted by glmPCA and thus seems sufficient to characterise the outage variability. Examining the distribution of points, we identify two distinct zones: one in the eastern to south-eastern part of Europe and the other in the western region, potentially extending towards Scandinavia.

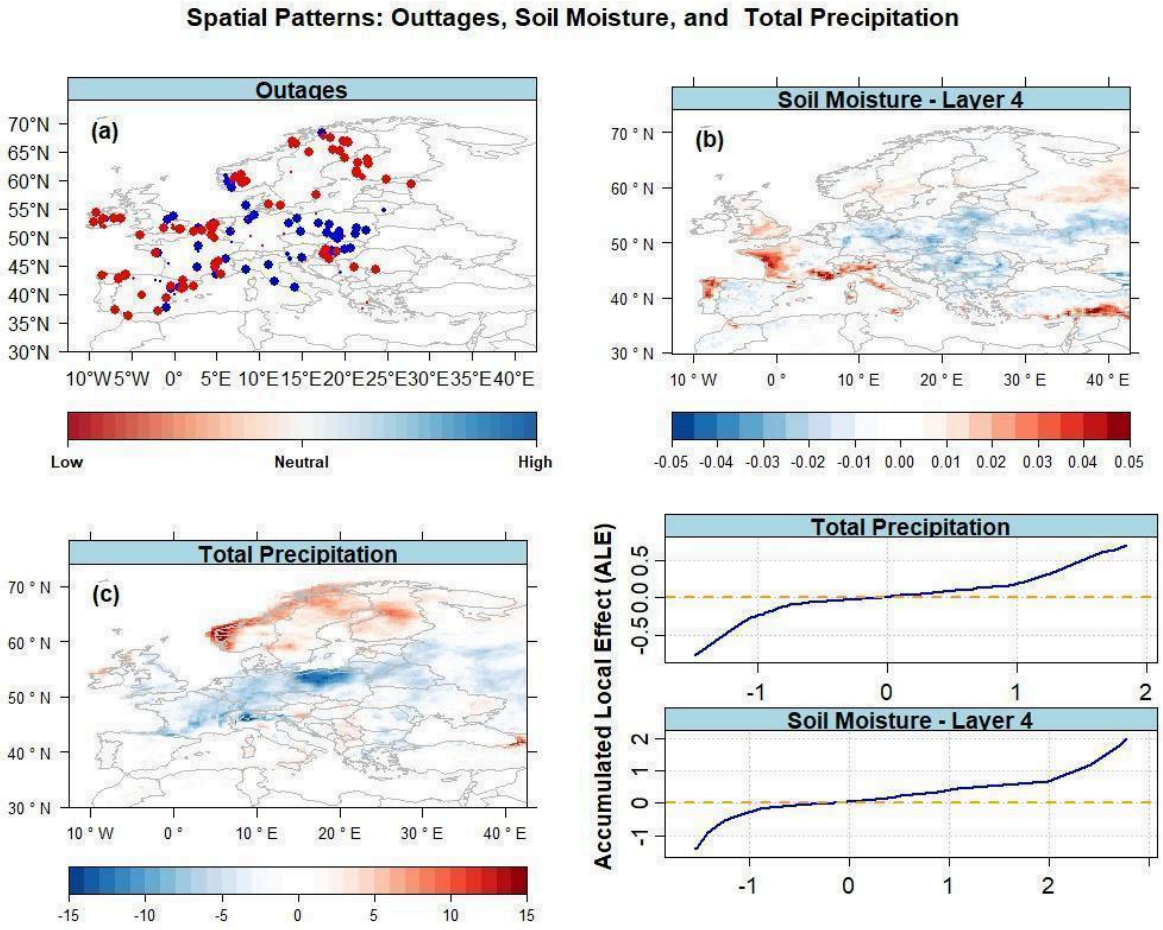


Figure 6.10: *Large-scale component extracted for the KRGCCA analysis of outages. Panel (a) shows the component of the outages obtained from glmPCA and KRGCCA. Panel (b) and (c) correspond to preimages of the first component of the KRGCCA for the (b) soil moisture layer 4 and (c) total precipitation. The panel on the bottom right shows ALEplots obtained for total precipitation and soil moisture in the fourth layer.*

Next, we identify the most relevant climate variables for the outage components using vine-copula-based quantile regression as in chapter 6.3.1. Model selection is done with the BIC and we predict the first outage component from the constructed first components of the

climate variables. The model selects soil moisture layer #4, total precipitation, and minimum temperature as the most influential variables. We begin by focusing on total precipitation and soil moisture, with the corresponding preimages or spatial patterns and ALE plots shown in Figure 6.10.

We observe that both climate variables primarily reflect dry conditions across Europe, which largely coincide with the blue points, representing areas with a higher likelihood of outages. The ALE plots indicate that an increase in the magnitude of both variables raises the likelihood of outages. However, soil moisture exhibits significantly higher ALE values, suggesting a more pronounced impact compared to summer precipitation. This highlights the compounded effect of these variables in contributing to outages. Interestingly, as observed in previous agricultural studies, the fourth layer of soil moisture appears to be the most influential, reinforcing the notion that dry winter conditions play a crucial role in driving socio-economic impacts. Here, however, it seems to be much more important than for agriculture (see previous chapter) highlighting that preceding dry winters are very relevant for accurate risk assessment.

Finally, we note that we have also examined the impacts of minimum temperature. Focusing on its interaction effects with total precipitation and soil moisture (Figure A6.4), we find that its influence is clearly overshadowed by these two variables. As a result, we do not discuss the role of minimum temperature further, as it appears to be less significant compared to the other variables.

6.3.3 Wet warm springs

The analysis of wet and warm springs focuses on the Alpine region and examines how excessive snowfall, subsequent melting, and concurrent rainfall contribute to severe flooding events. These flooding events primarily arise from the combined effects of melting snow in higher altitudes and precipitation in lower-elevation areas. To explore this phenomenon, we utilize daily data provided by the HERA dataset, focusing on the Alpine region (see Figure 6.11). River discharge patterns are analyzed alongside temperature and total precipitation data, both sourced from the EMO-5 dataset (Thiemig et al., 2022), covering the period from 1990 to 2020. These datasets, available at a 1x1 km resolution, present a high-dimensional challenge, so we apply Empirical Orthogonal Function (EOF) analysis to reduce dimensionality. The number of selected components is determined using the truncation criterion of Wilks (2016), with the explained variance displayed in Appendix, Table A6.1.

After applying the EOF analysis, we proceed with KRGCCA to extract the key patterns that describe the essential dynamics of the system. Large-scale variables from Chapter 6.3.1.1 based on the ERA5 data set are included as well. Given the relatively small spatial extent of the Alpine region, we begin by using the SIS procedure to identify the primary drivers influencing the Alpine region from the broader Euro-Atlantic domain. For this, we implement the Conditional SIS using Reflection via data splitting (CIS-ReDs; Tong *et al.*, 2023) for the

spring months (March, April, May - MAM), which allows for conditioning on relevant variables or known relationships. Specifically, we use the temperature and total precipitation-related EOFs to condition on local variables responsible for discharges, aiming to identify the corresponding large-scale patterns while taking the regional variables into account. After completing the CSIS-ReDs procedure, we apply KRGCCA. In this step, the discharges from the HERA data set is used as the response variable, while the reduced set of input features from the CSIS-ReDs process is used as predictors as well as the maximum, minimum temperature and total precipitation EOFs. This enables us to model the relationship between local discharges and large-scale climatic drivers, providing a more nuanced understanding of the dynamics governing flooding events in the Alpine region. Since there can also be lagged relationships we shift in time the climate variables using lags of $l=0,1,2,3$ where l indicates units of 30 days. For instance, a lag of two temporal units corresponds to a shift of 60 days (approximately two months) with respect to MAM, thus mostly describing JFM (January, February, March). Otherwise, we use the same setup as of chapter 6.3.1.

We employ the non-stationary Multi-Layer Perceptron kernel for all models when employing KRGCCA again with the connection matrix shown in Figure A6.1. We retain the first two components, which together account for approximately 88.60% of the variance in the reduced system of the discharges. In total, these components explain 79.20% of the variance across the system (as detailed in Table A6.1). To capture the spatial information, we generate preimages, which are displayed in Figure 6.11.

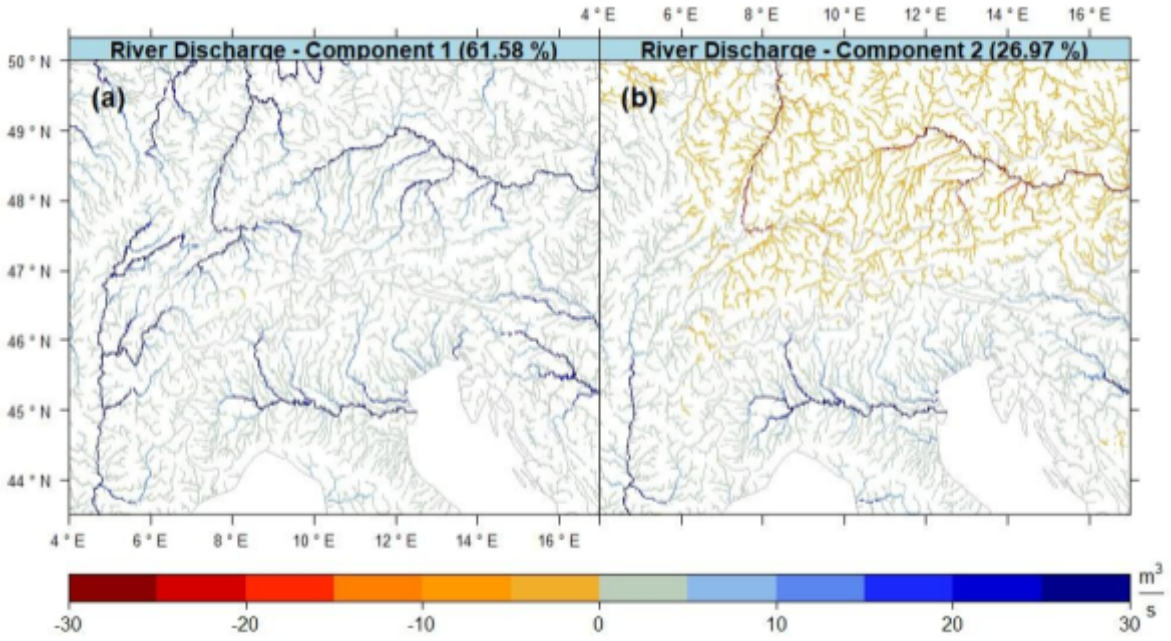
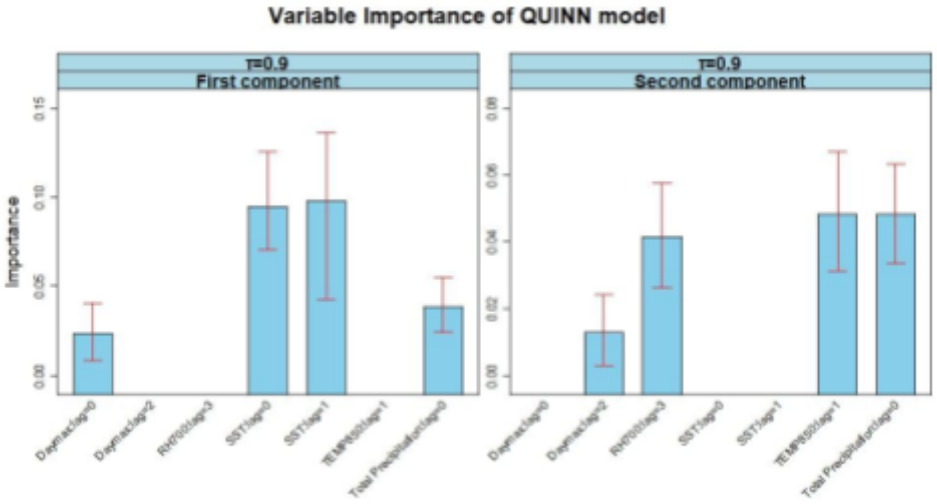


Figure 6.11: *The first two extracted components of river run-off from the KRGCCA approach taking the E-Hype model as output.*

We observe that the first component (when in a positive mode) primarily reflects large-scale discharges in the entire region. The second mode, on the other hand, regionalizes the patterns, capturing southern to southwestern discharges in a positive mode and northeastern discharges in the Alps when negative mode occurs. This regional differentiation makes it potentially relevant for the Lake Como case studies in WP7. We omit to study the third component, since it explains only 8.90% of the variance.

We apply the QUINN approach to model the two discharges components, using all the meteorological variables with which we have fed the KRGCCA. A dummy variable is included as well to indicate whether the predictors and output correspond to the first or second component. For the estimation, we use Monte-Carlo Markov Chain (MCMC) sampling with the NUTS algorithm, conducting 2000 warm-up iterations followed by 8000 iterations, discarding every eighth iteration to yield an ensemble of 1000 members. Hyperparameter tuning is performed using the WAIC. Figure A6.5 presents the Q-Q plot of the model, indicating satisfactory performance for both components. We then turn our attention to identifying the most important features for high discharges, and we compute variable importances for their conditional 10th and 90th percentile primarily because of the second component (Figure 6.11), which captures the extreme states of the positive phase (discharges south of the Alps) and the negative phase (discharges north of the Alps). We focus on the two most important large-scale and regional variables for these percentiles. Since we observed that the variable importances are similar across both percentiles, we are

only presenting the 90th percentile (corresponding to the phases depicted in Figure 6.11).



The results are displayed in Figure 6.12.

Figure 6.12: *Feature importance for the river discharges using the QUINN model based on the 90th percentile. Whiskers indicate 95 % credible intervals*

We see that the most important features differ for the two components: for the first, which is reflecting large-scale variability, large-scale variables are found to be more important (mainly the SSTs during MAM and FMA). Warm SSTs can transfer energy to the cooler atmosphere, enriching it with moisture. As this moisture-laden air moves towards land, particularly the Alpine region, it can result in heavy precipitation, including snowfall at higher altitudes. For this deliverable, we will, however, focus on the local variables. For the second component, we observe that total precipitation plays a significant role for both components, along with daily maximum temperature during JFM. The corresponding preimages are presented in Figure 6.13.

In JFM, the regionalization becomes quite clear: in a positive phase, the component reflects cold and wet conditions in the southern Alps (the opposite in negative phase). These conditions are conducive to snow accumulation, establishing pre-conditions for spring floods. The MAM (March, April, May) components, on the other hand, indicate warm and wet conditions, which likely contribute to melting snow and heightened flood risk. The combination of accumulated snow from late winter, coupled with melting and increased precipitation, sets the stage for compound conditions that lead to extreme flooding events.

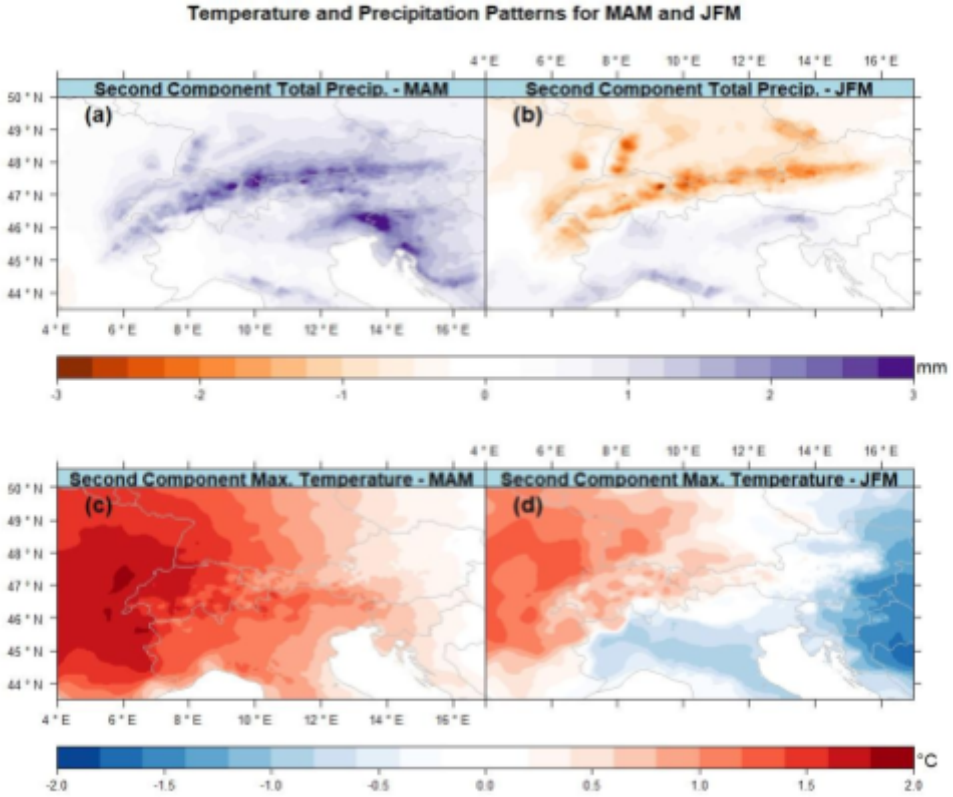


Figure 6.13: *Constructed preimages for the second component of the KRGCCA analysis reflecting Total Precipitation for (a) MAM and (b) JFM. (c) and (d) correspond to preimages of maximum temperature in MAM and JFM.*

To verify this hypothesis, we use ALE plots to examine the interaction effects. Since the variable importance analysis in Figure 6.12 indicates that maximum temperature in JFM and total precipitation in MAM are significant, we focus on these two variables. To describe this in a holistic manner, we compute the interaction for a sequence of percentiles and the ALE plot is shown in Figure 6.14. Note that high percentiles (in tendency) correspond to an extreme state of Figure 6.13(b), namely strong discharges South of the Alps. On the other hand, low percentiles are related to a pronounced negative (i.e. multiplied by -1) pattern of Figure 6.13(b), representing strong discharges North of the Alps. Hence, the obtained values of ALE should be read keeping in mind which percentile they have been calculated for. In Figure 6.14 positive values of an ALE are observed for large percentiles, implying that the large South discharges values become in tendency even higher. Conversely, negative ALEs for low percentiles indicate that the 10th percentile is expected to become even lower, representing stronger discharges North of the Alps.

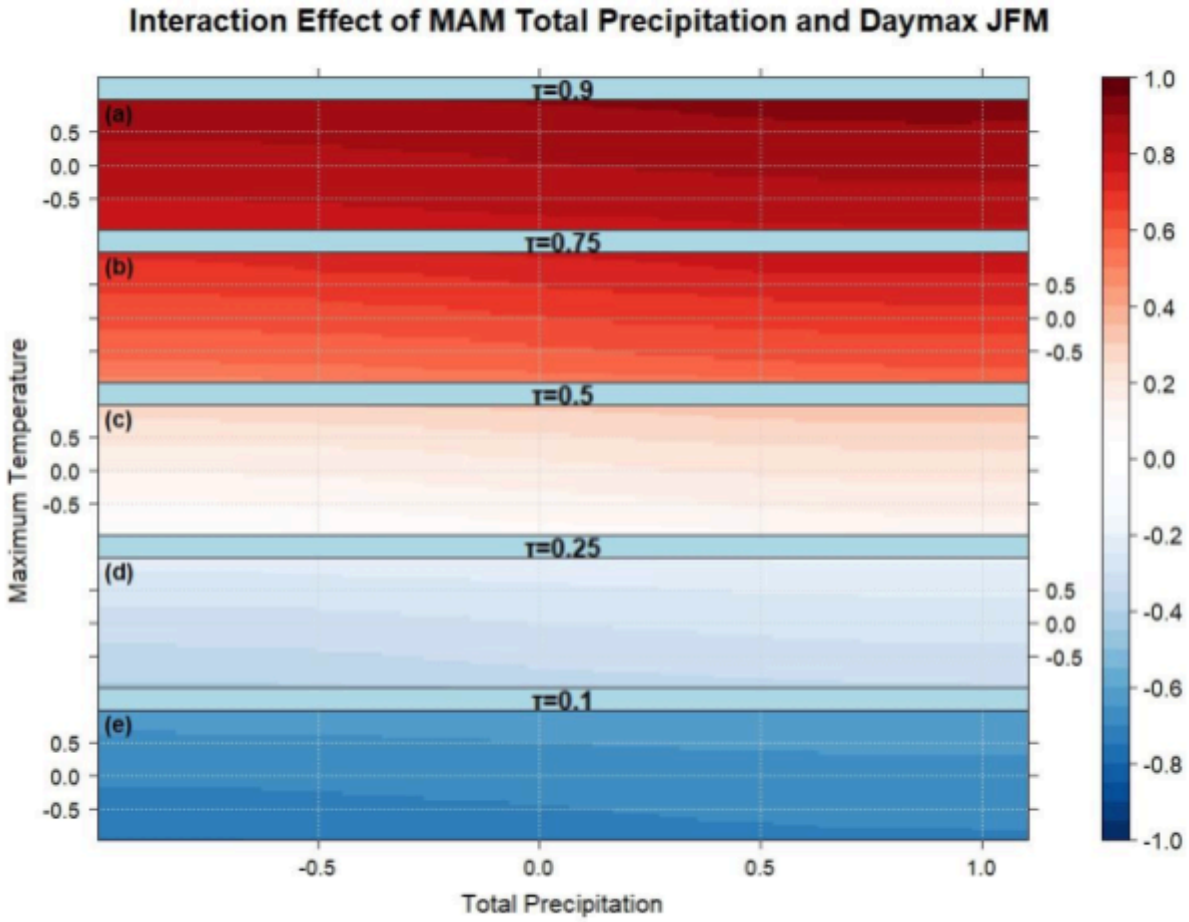


Figure 6.14: *Second order interaction of MAM Total Precipitation and JFM maximum temperature.*

Considering the 90th percentile: the highest ALEs values appear in the upper right panel of the plot indicating that the strongest effects on floods South of Alps occur when these regions experience a wet spring with preceding cold late winter. When focusing on the lower tail, specifically the 10th percentile, the effect is reversed (lowest ALEs in the lower left panel), which reflects that all spatial patterns (Figures 6.11 and 6.13) are in a negative phase. This indicates that the same phenomenon described above occurs north of the Alps. Interestingly, for the intermediate percentiles (0.25, 0.50, and 0.75), the effect is dampened, suggesting that the most extreme discharges events are primarily influenced by the compounding effects. These significantly exacerbate impacts at the most extreme states, underscoring the importance of identifying precursors for effective risk assessment.

Finally, we inspect whether increased discharges affect the hydropower sector, using the ENTSO-E dataset spanning the period from 2015 to 2021 (Chapter 6.2). The observational data in the study region is quite limited, and machine learning as well as statistical methods (e.g., imbalanced random forests, support vector machines, over/under sampling techniques, glmPCA) result in poor predictions of outages (accuracy ≤ 50%). Figure A6.6

shows the second discharge component superimposed with the outages in 2016 and 2020, which are selected because these years record a relatively high number of outages (n=24) compared to other years (n=15 for 2019, n≤9 for the others). For these two years, we compute lagged biserial correlations between the discharges and the outages, yielding quite different values (Figure A6.6), with only the one for 2020 being significant. This suggests that there is likely not a stationary link function between the two quantities, potentially explaining why the models fail in prediction. Given that there are only two years with a relatively high number of outages due to the limited observational record, and a causal link appears doubtful, we do not pursue a further in-depth analysis.

6.4 Concurrent Extremes

6.4.1 Nonparametric climate indices with an application to the SPEI

One of the most widely used indices for drought assessment is the SPEI introduced by Vicente-Serrano *et al.,* in 2010. SPEI relies on the PIT and often utilizes a log-logistic distribution due to its proven effectiveness. However, this distribution has bounded support, limiting the extrapolation characteristics hereby depending on the distribution parameters. We have discussed this issue in D3.2 using ERA5 data. Other benchmark studies (e.g., Vicente-Serrano *et al.,* in 2010; Begueria *et al.,* 2014) use CRU TS4.07 data set and we have reperformed the analysis on the latter, finding a similar phenomenon, even though to a lesser extent (see Figure A6.7 and A6.8 in the Annex). However, this shows that the parametric approaches used in those studies might be less effective on datasets on which they still have to be tested. Hence, it is worth investing into nonparametric approaches.

To address these limitations, we propose a nonparametric local likelihood-based approach (Loader, 1999), which accommodates higher moments and thereby outperforms classical kernel density methods. Using the Gaussian kernel, this method effectively represents a superimposition of Gaussian functions, ensuring mass distribution across the entire real line and enabling extrapolation. D3.2 demonstrated that this issue is fully resolved with our NPSPEI. The NPSPEI also shows more accurate PIT results, as indicated by improved Anderson-Darling statistics (Figure A6.9). Since these findings are already discussed in D3.2, we will not elaborate further.

To evaluate whether NPSPEI is better at capturing extreme events, we follow the approach outlined in Vicente-Serrano *et al.* (2010), which selected eleven representative stations worldwide as proxies for global wet and dry conditions. We compute both NPSPEI and SPEI using data from the stations available at:
https://github.com/sbegueria/SPEI/tree/master/data. Given that local likelihood estimators are known to perform better in the tails of the distribution (e.g., Geenens and Wang, 2018; Geenens, 2014; Nagler, 2018b; Nagler, 2018a), we expect our method to outperform other nonparametric approaches, such as kernel density estimators (KDE). To verify this, we also computed a KDE-based SPEI (KDESPEI) and compared it with SPEI and NPSPEI using Q-Q

plots for the eleven stations in a way with which we also produce a benchmark for nonparametric indices. These experiments are conducted for aggregation periods of 1, 3, 6, 12, and 48 months, with results only shown for six-month aggregation in Figure 6.15 as the results for the other months are quite similar.

We observe that the NPSPEI remains closest to the identity line and the tails (i.e., absolute values greater than one) in most cases, indicating that it adapts best to the data among all versions considered. This observation is consistent for most of the aggregation intervals, except for an aggregation scheme of one, where the performance between the indices is relatively balanced. This highlights three key improvements of the NPSPEI:

1. It offers superior extrapolation, making it more suitable for calibration on reference periods and avoiding the limitations imposed by the log-logistic distribution used by SPEI.
2. It produces smaller Anderson-Darling distances almost globally, suggesting a better overall fit to the normal distribution.
3. It performs better in the tails of the distribution, even when compared to classical kernel density estimators.

Finally, it is worth noting that these methods can be easily extended to other climate variables, including discrete variables as well as mixtures of continuous and discrete ones, making the NPSPEI concept broadly applicable to a wide range of climate-related data.

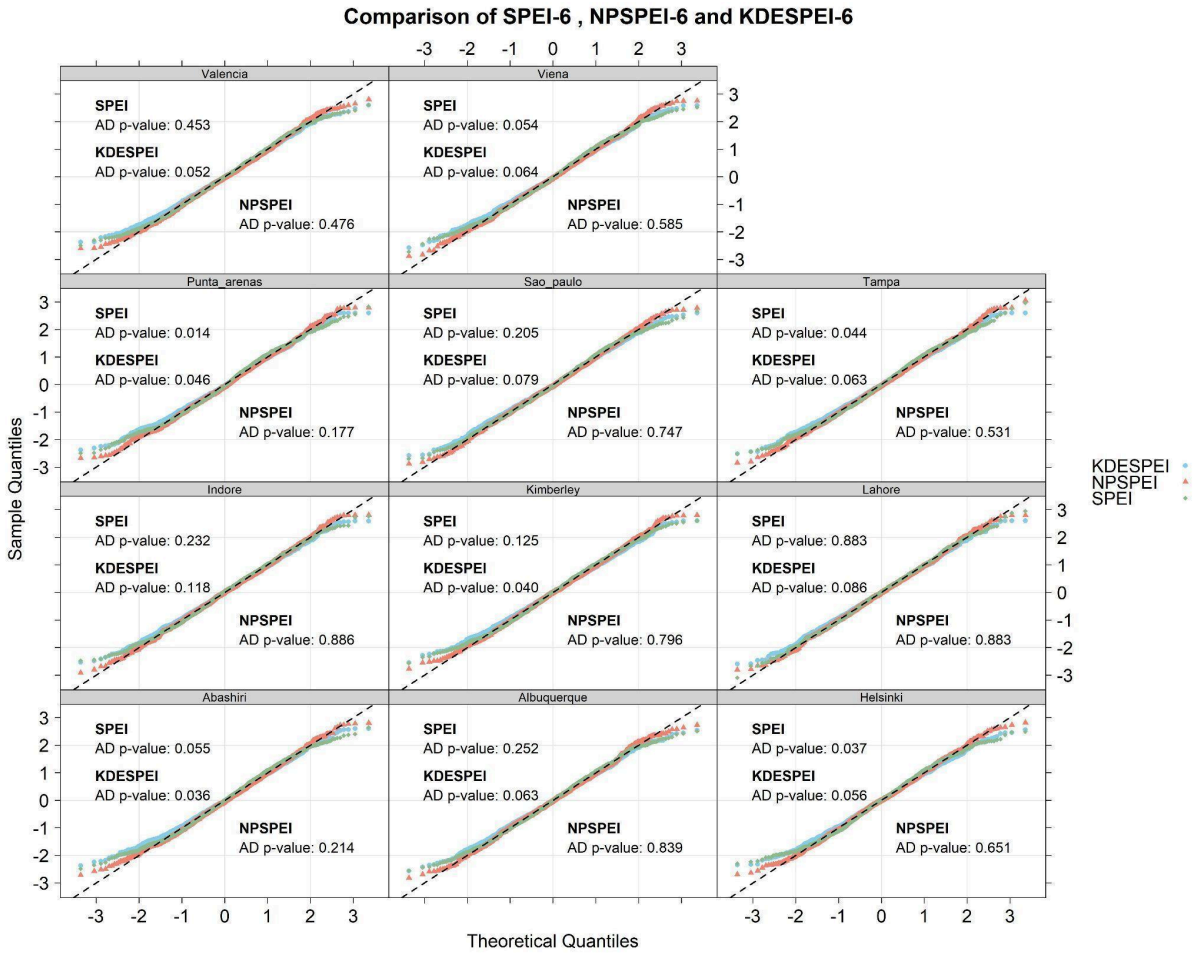


Figure 6.15: *Q-Q-plots of the considered SPEI version for eleven representative regions in the world.*

6.4.2 Detection of dependencies using AI-enhanced point process approaches

Heatwaves and droughts are often closely connected due to their inherent physical relationships (see D.4.1 for a comprehensive review), and they frequently have significant socio-economic impacts, particularly in agriculture. For instance, studies indicated that these events can reduce cereal yields by 9-10% at the national level and explain up to 40% of the interannual variability in crop yields (Lesk *et al.*, 2016; Zampieri *et al.,* 2017). This is especially concerning as the frequency and intensity of heatwaves and droughts are expected to increase under future climate change scenarios (Zscheischler and Seneviratne, 2017; Toreti *et al.,* 2019a; Alizadeh *et al.*, 2020; Vogel *et al.*, 2020; Meng *et al.*, 2022).

While many studies examine the co-variability of heatwaves and droughts at regional or national scales, they often overlook how large-scale teleconnections might propagate these impacts to other regions. For instance, research has explored the links between the North Pacific Oscillation, the El Niño Southern Oscillation, and the Arctic Oscillation, which may

connect heatwaves and droughts across continents, such as in Europe and Australia (Chen *et al.*, 2013). Toreti *et al.,* (2019b) showed how these teleconnections can be examined using the inhomogeneous J-function (van Lieshout, 2011; Cronie and van Lieshout, 2015, 2016) which identifies whether extreme events are independent, clustered, or inhibited. The J-function also accounts for temporal variability, an important aspect given the influence of climate change. However, the method's reliance on graphical interpretation makes it prone to subjectivity, a limitation discussed in D2.3. To address this, in this chapter, we will apply an AI-based approach discussed in D2.3 to detect these dependencies more objectively.

To conduct this analysis, we first calculate the NPSPEI and HMD indices globally using the ERA5 dataset, followed by grid-point level detection of droughts and heatwaves. Droughts are identified when the NPSPEI drops below -1, while heatwaves are defined as periods when the HMD exceeds the 90th percentile. A large-scale event is classified when at least 20% of the grid points in a region of interest meet these conditions. For these large-scale events, J-functions are computed and subsequently classified using the neural network model, as previously described.

The analysis is carried out based on IPCC regions. Given the asymmetric nature of the J-function, we treated droughts and heatwaves in each region as distinct drivers. We calculated the J-function for all variable combinations, with the regions in Europe (Northern Europe (NEU), Southern Europe/Mediterranean (MED), Central Europe (CEU)) being employed as response variables. The analysis is performed on a seasonal scale, and the results are presented in Figure 6.16.

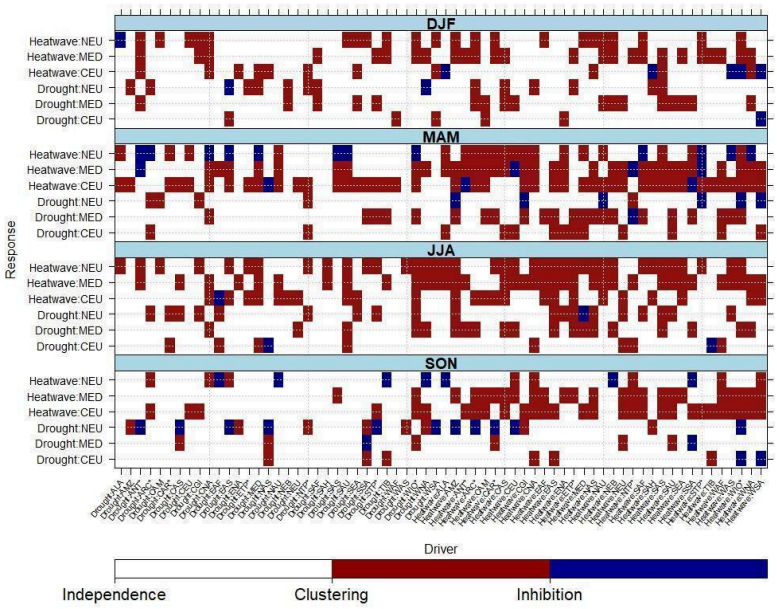


***Figure 6.16**: Identification of dependencies between heat waves and droughts in the IPCC regions through the AI-based J-function interpreter.*

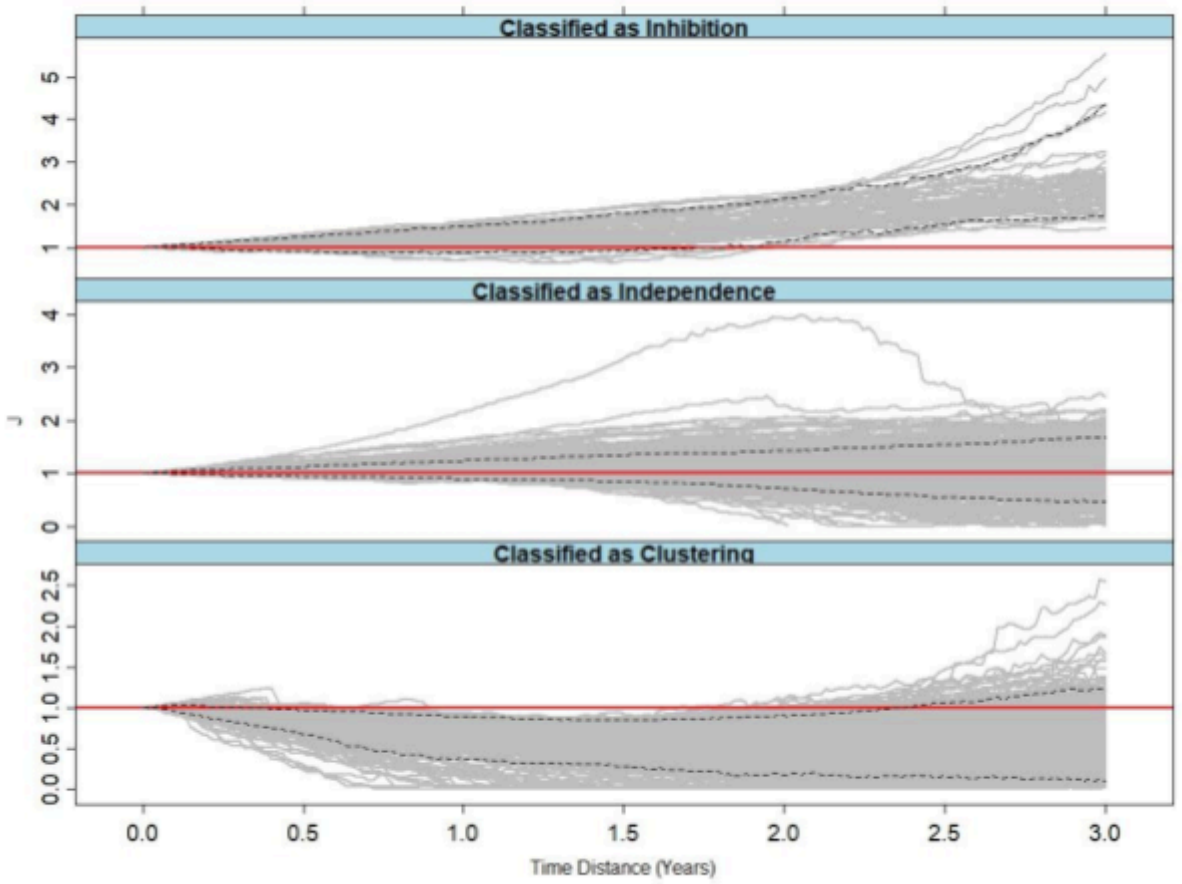


Figure 6.17: *All classified J-functions of figure 15 plotted with respect to their identified class. The dotted lines correspond to pointwise 10th and 90th percentiles, indicating that most of the identified functions follow their theoretical trajectories.*

We observe that many interdependencies or clustering phenomena are present. Specifically, independence is observed the most frequently (N=1042), followed by clustering (N=454), with inhibition being quite rare (N=61). These findings clearly underscore the heightened risk associated with compound events, as their occurrence is likely to impact multiple vulnerable sectors in different spatial regions of the world.

To ensure the validity of these results, we plot the J-functions for each classified group, and the results are shown in Figure 6.17. As described in section 6.2, J-functions that have been identified as clustering should be smaller than one, the ones identified as independent approximately equal to one and those for inhibition larger than one. We observe that for clustering most of the J-function remain below the one line. We observe that the functions largely follow these trajectories and closely resemble their theoretical counterparts (see D2.3), indicating that the developed algorithm works well. Remarkably, the evaluation of all functions is completed in just 8.2 seconds, demonstrating that global interdependencies can now be assessed in a matter of seconds, while accounting for the non-stationarity of climate change signals.

6.5 Conclusion

Our studies demonstrate that focusing on the combination of climate events can be beneficial for risk assessment. The analysis of wet and warm late winters followed by dry springs reveals that high impacts can arise from combinations of events that are not individually extreme, underscoring the importance of such combinations in holistic risk assessment. In our second case study, a more flexible approach to defining multivariate thresholds, which better accounts for the dependence structure between maximum and minimum temperatures, is proved to be beneficial for identifying heatwaves. By using these detection methods and integrating drought indicators for both winter and summer, we significantly improve benchmark indices for modelling the climate impact on agriculture through the incorporation of non-linear models. With the aid of AI, we develop a large-scale model for European crops, demonstrating that the likelihood of crop failures during compound events are significantly higher compared to single extremes. Also, we find significant linkages between the energy sector and both dry winters and dry summers. Additionally, results from our third case study show that strong river runoffs are strongly favoured by a compounding effect of cold late winters followed by wet springs.

In our analysis of concurrent extremes, we introduce new nonparametric methods for creating climate indices. As an example, we benchmark these methods against the SPEI, which has been shown to struggle with extrapolating values beyond chosen reference periods. Our index is able to overcome this burden, while also improving the statistical properties (i.e., behaviour in tails) of the indices. Moreover, we develop a tool for analysing the interconnections between extreme events while accounting for the non-stationarity of the climate, revealing that heatwaves and droughts in many parts of the world are interconnected with those occurring in Europe.

The methods developed in our research can significantly improve event management, forecasting, and climate adaptation to reduce future disaster risk. Our newly developed indices (NPSPEI, BVHMD, non-linear CSI) offer enhanced predictability for droughts, heatwaves, and yield anomalies. By integrating these indices into WP6, we aim to better identify “areas of concern” for the food sector and other critical areas. These indices can be applied to various forecasting time horizons, improving the accuracy of extreme event predictions and bolstering resilience in vulnerable regions. In addition, our studies on compound events have revealed preconditions that can lead to severe socio-economic impacts across sectors like food, energy, and water. Recognizing these preconditions early is essential for the development of effective early warning systems, as they often highlight risks that might otherwise be underestimated. Some of these preconditions also offer seasonal predictability, enabling earlier preparedness and more effective risk mitigation. Finally, our research into the interdependencies of large-scale extreme events, such as droughts and heatwaves, demonstrates how these connections can lead to cascading effects across regions. For instance, simultaneous crop failures in key agricultural areas can trigger global spikes in food prices, leading to longer-term socio-economic consequences.

Understanding these global linkages allows for improved forecasting and better preparation, helping to mitigate the broader impacts of interconnected disasters.

7 CONCLUSIONS

This deliverable reports the final results regarding the improvement of extreme events detection with ML algorithms and in some cases its implication for predictions. The study has focused on different categories of extreme events:

- Tropical cyclones: genesis indices (Chapter 2), TC activity and extratropical transitions (Chapter 3)
- Heatwaves and warm nights (Chapter 4)
- Extreme droughts (Chapter 5)
- Compound events and concurrent extremes (Chapter 6)

Many aspects of detection have been tackled with existing ML-based methods and those developed in WP2. In some cases, they are used to refine existing indices, (e.g. the selection of parameter values for GPI of Tropical Cyclones), to identify thresholds which are used to detect the EE themselves (e.g. RF for compound events), or to identify which indicators explain variability in impacts of EE (e.g. HW indicators and crop yield). ML approaches have also been used for dimensionality reduction, a key step in ML-based detection, such as for clustering of drought occurrences over Europe.

In all cases, the use of ML has allowed identification of key drivers to target EEs, and in particular has focused on optimising detection skill. Besides various sensitivity analysis performed in WP2, here feature engineering is used to determine key set-up parameters based on physical understanding of the EEs. Novel evolutionary/genetic algorithms are used as feature selection tools to highlight key variables from a pool of potential predictors (e.g. for TCs, HWs, and droughts). Modifications to AI-enhanced short-term forecasts of TCs have also identified the optimal set-up parameters, such as the importance of near-term or global-scale input and various climate indices. In all cases, discussions are opened on the physical meaning of the selected features, which can form the basis of future work. Despite the highly imbalanced nature of EE datasets, event recreation or forecasting using dimensionality reduction of predictors was proven possible for the whole range of EE studied here.

The diverse range of ML approaches has employed an equally diverse range of climate datasets for training. When possible, training with ERA5 has ensured training on real-world data. Meanwhile, certain problems are tested using model world training datasets, such as CMIP6 or paleo simulations; it is demonstrated that learning in the model world is transferable to real world detection problems. In the case of droughts and compound events, impact-based variables (e.g. crop yield or energy usage) are incorporated into detection algorithms to provide more effective and relevant indices.

Regarding prediction of EE, a key activity is the creation of new data-driven forecasting systems or the development of AI-enhancements to existing systems. Forecasting horizons covered short-term and S2S tropical cyclones forecasts to seasonal forecasts of HWs. In this activity, a key concept is “added value” (explored in greater detail in Deliverables on case studies in WP6 and WP7). Here, novel forecasting approaches display diverse types of added value. Crucially, some systems outperform existing dynamical forecasts (for example, TCs in ECMWF short-term) in terms of prediction skill. In the case of seasonal forecasts of HWs, while there are regional-scale patterns of significant correlation in the data-driven approach, existing operational systems still outperform on a European scale. In such cases, the DD approaches still add value by reducing computational expense and moreover provide complementary information (of scientific and practical relevance) on the drivers used to make predictions.

Finally, many of the detection methods presented here are being applied across the project in either pan-European (e.g. compound events in WP6) or local-scale (e.g. droughts and heatwaves in WP7) case studies. Others are being prepared for deployment in the Climate Services Information Systems (WP8), or in the demonstration of prototype operational prediction services (WP9). The application and operationalisation of these systems will ensure their continued development within and beyond the project.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.

Alizadeh, M. R., Adamowski, J., Nikoo, M. R., AghaKouchak, A., Dennison, P., & Sadegh, M. (2020). A century of observations reveals increasing likelihood of continental-scale compound dry-hot extremes. *Science Advances*, 6(39). https://doi.org/10.1126/sciadv.aba0787

Allstadt, A. J., Vavrus, S. J., Heglund, P. J., Pidgeon, A. M., Thogmartin, W. E., & Radeloff, V. C. (2015). Spring plant phenology and false springs in the conterminous US during the 21st century. *Environmental Research Letters*, 10(10), 104008. https://doi.org/10.1088/1748-9326/10/10/104008

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4), 1059–1086. https://doi.org/10.1111/rssb.12374

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1), 243–256. https://doi.org/10.1016/j.patcog.2012.04.027

Ascenso, G., Cavicchia, L., Scoccimarro, E., & Castelletti, A. (2023). Optimisation-based refinement of genesis indices for tropical cyclones. *Environmental Research Communications*, 5(2), 021001. https://doi.org/10.1088/2515-7620/acb52a

Ault, T. R., Henebry, G. M., Beurs, K. M. de, Schwartz, M. D., Betancourt, J. L., & Moore, D. (2013). The false spring of 2012, earliest in North American record. *Eos, Transactions American Geophysical Union*, 94(20), 181–182. https://doi.org/10.1002/2013EO200002

Bárdossy, A., Stehlík, J., & Caspary, H., 2002. Automated objective classification of daily circulation patterns for precipitation and temperature downscaling based on optimized fuzzy rules. Climate Research, 23: 11-22.

Baker, A. J., Hodges, K. I., Schiemann, R. K. H., & Vidale, P. L. (2021). Historical variability and lifecycles of North Atlantic midlatitude cyclones originating in the tropics. *Journal of Geophysical Research: Atmospheres*, 126, e2020JD033924. https://doi.org/10.1029/2020JD033924

Baldi, Pierre. "Autoencoders, unsupervised learning, and deep architectures." Proceedings of ICML workshop on unsupervised and transfer learning. JMLR Workshop and Conference Proceedings, 2012.

Barriopedro, D., García–Herrera, R., Ordóñez, C., Miralles, D. G., & Salcedo–Sanz, S. (2023). Heat waves: Physical understanding and scientific challenges. *Reviews of Geophysics*, 61(2). https://doi.org/10.1029/2022RG000780

Beguería, S., Vicente-Serrano, S. M., Reig, F., & Latorre, B. (2014). Standardized precipitation evapotranspiration index (SPEI) revisited: Parameter fitting, evapotranspiration models, tools, datasets and drought monitoring. *International Journal of Climatology*, 34(10), 3001–3023. https://doi.org/10.1002/joc.3887

Ben-Ari, T., Boé, J., Ciais, P., Lecerf, R., van der Velde, M., & Makowski, D. (2018). Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France. *Nature Communications*, 9(1), 1627. https://doi.org/10.1038/s41467-018-04062-0

Beran, J., Feng, Y., Ghosh, S., & Kulik, R. (2013). *Long-memory processes: Probabilistic properties and statistical methods*. Softcover reprint of the original 1st ed. 2013. Springer Berlin; Springer, Berlin.

Bosart, L. F., & Lackmann, G. M. (1995). Postlandfall tropical cyclone reintensification in a weakly baroclinic environment: A case study of Hurricane David (September 1979). *Monthly Weather Review*, 123(11), 3268–3291. https://doi.org/10.1175/1520-0493(1995)123<3268>2.0.CO;2

Cavicchia, L., Scoccimarro, E., Ascenso, G., Castelletti, A., Giuliani, M., & Gualdi, S. Tropical cyclone genesis potential indices in a new high-resolution climate models ensemble (2023). Limitations and way forward. Geophysical Research Letters, 50, https://doi.org/10.1029/2023GL103001

Ceglar, A., Toreti, A., Lecerf, R., van der Velde, M., & Dentener, F. (2016). Impact of meteorological drivers on regional inter-annual crop yield variability in France. *Agricultural and Forest Meteorology*, 216, 58–67. https://doi.org/10.1016/j.agrformet.2015.10.006

Ceglar, A., Turco, M., Toreti, A., & Doblas-Reyes, F. J. (2017). Linking crop yield anomalies to large-scale atmospheric circulation in Europe. *Agricultural and Forest Meteorology*, 240-241, 35–45. https://doi.org/10.1016/j.agrformet.2017.03.002

Chamberlain, C. J., Cook, B. I., García de Cortázar-Atauri, I., & Wolkovich, E. M. (2019). Rethinking false spring risk. *Global Change Biology*, 25(7), 2209–2220. https://doi.org/10.1111/gcb.14665

Chaudhuri, A., Sadek, C., Kakde, D., Wang, H., Hu, W., Jiang, H., Kong, S., Liao, Y., & Peredriy, S. (2021). The trace kernel bandwidth criterion for support vector data description. *Pattern Recognition*, 111, 107662. https://doi.org/10.1016/j.patcog.2021.107662

Chen, S., Chen, W., Yu, B., & Graf, H.-F. (2013). Modulation of the seasonal footprinting mechanism by the boreal spring Arctic Oscillation. *Geophysical Research Letters*, 40(24), 6384–6389. https://doi.org/10.1002/2013GL058207

Coumou, D., Di Capua, G., Vavrus, S., Wang, L., & Wang, S. (2018). The influence of Arctic amplification on mid-latitude summer circulation. *Nature Communications*, 9(1), 1–12. doi:10.1038/s41467-018-05256-8.

Cronie, O., & van Lieshout, M. N. M. (2015). A J-function for inhomogeneous spatio-temporal point processes. *Scandinavian Journal of Statistics*, 42(2), 562–579. https://doi.org/10.1111/sjos.12129

Cronie, O., & van Lieshout, M. N. M. (2016). Summary statistics for inhomogeneous marked point processes. *Annals of the Institute of Statistical Mathematics*, 68(4), 905–928. https://doi.org/10.1007/s10237-015-0670-3

Cuturi, M., & Blondel, M. (2017). Soft-DTW: A differentiable loss function for time-series. *Proceedings of the 34th International Conference on Machine Learning*.

Dainelli F., Ascenso G., Cavicchia L. & Scoccimarro E. ML-Improved indices for Tropical Cyclone Genesis. In preparation.

Davis, R.E., McGregor, G.R. and Enfield, K.B. (2016), "Humidity: A review and primer on atmospheric moisture and human health", *Environmental research*, Vol. 144 Pt A, pp. 106–116

Della-Marta, P. M., Luterbacher, J., von Weissenfluh, H., Xoplaki, E., Brunet, M., & Wanner, H. (2007). Summer heat waves over western Europe 1880–2003, their relationship to large-scale forcings and predictability. *Climate Dynamics*, *29*(2), 251-275. doi:10.1007/s00382-007-0233-1

Denzil, G. Fiebig (2007). Seemingly unrelated regression. In *A Companion to Theoretical Econometrics* (pp. 101–121). John Wiley & Sons, Ltd.

Dimitriadis, T., Gneiting, T., & Jordan, A. I. (2021). Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, *118*(8), doi: 10.1073/pnas.2016191118.

Domeisen, D. I., Eltahir, E. A., Fischer, E. M., Knutti, R., Perkins-Kirkpatrick, S. E., Schär, C., & Wernli, H. (2023). Prediction and projection of heatwaves. *Nature Reviews Earth & Environment*, 4(1), 36–50. https://doi.org/10.1038/s43017-022-00350-1

Dong, L., Mitra, C., Greer, S., & Burt, E. (2018). The dynamical linkage of atmospheric blocking to drought, heatwave, and urban heat island in southeastern US: A multi-scale case study. *Atmosphere*, 9(1), 33. https://doi.org/10.3390/atmos9010033

Duchez, A., Frajka-Williams, E., Josey, S. A., Evans, D. G., Grist, J. P., Marsh, R., & Hirschi, J. J. (2016). Drivers of exceptionally cold North Atlantic Ocean temperatures and their link to the 2015 European heat wave. *Environmental Research Letters*, 11(7), 074004. https://doi.org/10.1088/1748-9326/11/7/074004

Emanuel, K., & Nolan, D. S. (2004). Tropical cyclone activity and the global climate system. *Nature*, 426(6964), 134–135. https://doi.org/10.1038/426134a

Evans, C., et al. (2017). The Extratropical Transition of Tropical Cyclones. Part I: Cyclone Evolution and Direct Impacts, *Monthly Weather Review*, 145(11), 4317-4344. doi: 10.1175/MWR-D-17-0027.1

Fan, J., Li, R., Zhang, C.-H., & Zou, H. (2020). Statistical foundations of data science. *Chapman & Hall/CRC Data Science Series*. CRC Press Taylor & Francis Group, Boca Raton, London, New York.

Feng, Y., Gries, T., & Fritz, M. (2020). Data-driven local polynomial for the trend and its derivatives in economic time series. *Journal of Nonparametric Statistics*, 32(2), 510–533. https://doi.org/10.1080/10485252.2019.1631703

Fischer, E. M., & Schär, C. (2010). Consistent geographical patterns of changes in high-impact weather. *Nature Geoscience*, 3(3), 206–210. https://doi.org/10.1038/ngeo763

Fischer, E. M., & Schär, C. (2016). Increased intensity of precipitation extremes in a warmer climate. *Nature Climate Change*, 6(6), 579–586. https://doi.org/10.1038/nclimate2960

Frank, W. M., & Roundy, P. E. (2006). The Role of Tropical Waves in Tropical Cyclogenesis, Monthly Weather Review, 134(9), 2397-2417, doi: 10.1175/MWR3204.1.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences, 55*, 119–139.

Freychet, N., Tett, S., Wang, J., & Hegerl, G. (2017). Summer heat waves over eastern China: Dynamical processes and trend attribution. *Environmental Research Letters*, *12*(2), 024015, doi: 0.1088/1748-9326/aa5ba3

Fudeyasu, H. (2014). *A Global View of the Landfall Characteristics of Tropical Cyclones*. *3*(3).

García-Martínez, I. M., & Bollasina, M. A. (2021). Identifying the evolving human imprint on heat wave trends over the United States and Mexico. *Environmental Research Letters*, *16*(9), 094039. doi:10.1088/1748-9326/ac1edb

Geenens, G. (2014). Probit transformation for kernel density estimation on the unit interval. *Journal of the American Statistical Association*, 109(505), 346–358. https://doi.org/10.1080/01621459.2013.860677

Geenens, G., & Wang, C. (2018). Local-likelihood transformation kernel density estimation for positive random variables. *Journal of Computational and Graphical Statistics*, 27(4), 822–835. https://doi.org/10.1080/10618600.2017.1366899

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning, 63*, 3–42.

Ghosh, S., & Beran, J. (2009). Estimation of long memory and its application to modeling financial time series. *Statistical Modelling*, 9(3), 207–222. https://doi.org/10.1177/1471082X0900900301

Giuliani, M., Castelletti, A., Pianosi, F., Mason, E., & Reed, P. M. (2016). Curses, tradeoffs, and scalable management: Advancing evolutionary multiobjective direct policy search to improve water reservoir operations. Journal of Water Resources Planning and Management, 142(2), 04015050.

Giuliani, M., Carola, C., McAdam, R., Squintu, A., Scoccimarro, E, and Casteletti, A. Discovering the impacts of temperature extremes on agricultural production to support farmers' adaptation to climate change. In preparation.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org

Gray, W. M. (1979). Hurricanes: Their formation, structure and likely role in the tropical circulation. Meteorology over the tropical oceans, 155, 218

Gray, W. M. (1984). Atlantic Seasonal Hurricane Frequency. Part I: El Niño and 30 mb Quasi-Biennial Oscillation Influences. Monthly Weather Review, 112(9), 1649-1668, doi:10.1175/1520-0493(1984)112<1649:ASHFPI>2.0.CO;2.

Guo, X., Ren, H., Zou, C., & Li, R. (2023). Threshold selection in feature screening for error rate control. Journal of the American Statistical Association, 118(543), 1773–1785. https://doi.org/10.1080/01621459.2022.2120913

Hansen, F., Feser, F., Zorita, E., McAdam, R. Daytime and nighttime heatwave clusters over Europe and their physical drivers. In preparation.

Hao, Z., Hao, F., Xia, Y., Feng, S., Sun, C., Zhang, X., Fu, Y., Hao, Y., Zhang, Y., & Meng, Y. (2022). Compound droughts and hot extremes: Characteristics, drivers, changes, and impacts. Earth-Science Reviews, 235, 104241. https://doi.org/10.1016/j.earscirev.2022.104241

Hao, Z., Singh, V. P., & Hao, F. (2018). Compound extremes in hydroclimatology: A review. *Water*, 10(6), 718. https://doi.org/10.3390/w10060718

Harris, I., Osborn, T. J., Jones, P., & Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data*, 7(1), 109. https://doi.org/10.1038/s41597-020-0460-0

Hastie, T., Tibshirani, R., & Friedman, J.H. (2009). *The elements of statistical learning: Data mining, inference, and prediction, Springer Series in Statistics.* Second edition, Springer Science & Business Media, 745 pp.

Henderson, S. A., & Maloney, E. D. (2013). An intraseasonal prediction model of Atlantic and east Pacific tropical cyclone genesis. Monthly Weather Review, 141(6), 1925-1942, doi: 10.1175/MWR-D-12-00268.1.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., & Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hoffman, M., & Gelman, A. (2011). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15, 1593–1623.

Honeine, P., & Richard, C. (2011). Preimage problem in kernel-based machine learning. *IEEE Signal Processing Magazine*, 28(3), 77–88. https://doi.org/10.1109/MSP.2011.941027

Huang, B., Liu, C., Banzon, V., Freeman, E., Graham, G., Hankins, B., Smith, T., & Zhang, H.-M. (2021). Improvements of the Daily Optimum Interpolation Sea Surface Temperature (DOISST) Version 2.1. *Journal of Climate*, 34(8), 2923–2939. https://doi.org/10.1175/JCLI-D-20-0455.1

Jungclaus, J. H., K. Lohmann, and D. Zanchettin, 2014: Enhanced 20th-century heat transfer to the Arctic simulated in the context of climate variations over the last millennium. Clim. Past, 10, 2201–2213, https://doi.org/10.5194/cp-10-2201-2014.

Jungclaus, J.H., Bard, E., Baroni, M., Braconnot, P., Cao, J., Chini, L.P., Egorova, T., Evans, M., González-Rouco, J.F., Goosse, H., Hurtt, G.C., Joos, F., Kaplan, J.O., Khodri, M., Klein Goldewijk, K., Krivova, N., LeGrande, A.N., Lorenz, S.J., Luterbacher, J., Man, W., Maycock, A.C., Meinshausen, M., Moberg, A., Muscheler, R., Nehrbass-Ahles, C., Otto-Bliesner, B.I., Phipps, S.J., Pongratz, J., Rozanov, E., Schmidt, G.A., Schmidt, H., Schmutz, W., Schurer, A., Shapiro, A.I., Sigl, M., Smerdon, J.E., Solanki, S.K., Timmreck, C., Toohey, M., Usoskin, I.G., Wagner, S., Wu, C.-J., Yeo, K.L., Zanchettin, D., Zhang, Q., & Zorita, E. (2017). The PMIP4 contribution to CMIP6–Part 3: The Last Millennium, Scientific Objective, and Experimental Design for the PMIP4 Past1000 Simulations. *Geoscientific Model Development*, 10(11), 4005-4033. https://doi.org/10.5194/gmd-10-4005-2017.

Kämäräinen, M., Uotila, P., Karpechko, A. Y., Hyvärinen, O., Lehtonen, I., & Räisänen, J. (2019). Statistical Learning Methods as a Basis for Skillful Seasonal Temperature Forecasts in Europe. *Journal of Climate*, 32(17), 5363-5379.

Katsafados, P., Papadopoulos, A., Varlas, G., Papadopoulou, E., & Mavromatidis, E. (2014). Seasonal predictability of the 2010 Russian heat wave. *Natural Hazards and Earth System Sciences*, *14*(6), 1531-1542., doi: 10.5194/nhess-14-1531-2014

Kautz, L.-A., Martius, O., Pfahl, S., Pinto, J.G., Ramos, A.M., Sousa, P.M. & Woollings, T. (2022). Atmospheric Blocking and Weather Extremes over the Euro-Atlantic Sector – A Review. *Weather and Climate Dynamics*, 3(1), 305–336.

Keller, J. H., et al. (2019). The Extratropical Transition of Tropical Cyclones. Part II: Interaction with the Midlatitude Flow, Downstream Impacts, and Implications for Predictability, *Monthly Weather Review*, 147(4), 1077-1106, doi:10.1175/MWR-D-17-0329.1

Kendrovski, V., Baccini, M., Martinez, G. S., Wolf, T., Paunovic, E., & Menne, B. (2017). Quantifying Projected Heat Mortality Impacts under 21st-Century Warming Conditions for Selected European Countries. *International Journal of Environmental Research and Public Health*, 14(7), 729. https://doi.org/10.3390/ijerph14070729.

Kenyon, J., & Hegerl, G. C. (2008). Influence of Modes of Climate Variability on Global Temperature Extremes. *Journal of Climate*, 21(15), 3872-3889. https://doi.org/10.1175/2008JCLI2125.1.

Kiladis, G. N., Wheeler, M. C., Haertel, P. T., Straub, K. H., and Roundy, P. E. (2009), Convectively coupled equatorial waves, *Rev. Geophys.*, 47, RG2003, doi: 10.1029/2008RG000266.

Kipf, T. N., & Welling, M. (2017). *Semi-Supervised Classification with Graph Convolutional Networks* (arXiv:1609.02907). arXiv. http://arxiv.org/abs/1609.02907

Kirkpatrick, S., Gelatt Jr, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598), 671-680.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). Logistic regression. Springer.

Klotzbach, P. J. (2014). The Madden–Julian oscillation's impacts on worldwide tropical cyclone activity. Journal of Climate, 27(6), 2317–2330, doi:/10.1175/JCLI-D-13-00483.1

Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J., & Neumann, C. J. (2010). The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying Tropical Cyclone Data. *Bulletin of the American Meteorological Society*, *91*(3), 363–376. https://doi.org/10.1175/2009BAMS2755.1

Kornhuber, K., Coumou, D., Vogel, E., Lesk, C., Donges, J. F., Lehmann, J., & Horton, R. M. (2020). Amplified Rossby waves enhance risk of concurrent heatwaves in major breadbasket regions. *Nature Climate Change*, *10*(1), 48-53, doi:10.1038/s41558-019-0637-z

Kraus, D. and Czado, C. (2017). D-vine Copula Based Quantile Regression. *Computational Statistics & Data Analysis*, 110, 1–18.

Lavaysse, C., Cammalleri, C., Dosio, A., van der Schrier, G., Toreti, A., & Vogt, J. (2018). Towards a Monitoring System of Temperature Extremes in Europe. *Natural Hazards and Earth System Sciences*, 18(1), 91–104.

Lawton, Q. A., Majumdar, S. J., Dotterer, K., Thorncroft, C., & Schreck III, C. J. (2022). The influence of convectively coupled kelvin waves on african easterly waves in a wave-following framework. *Monthly Weather Review*. doi:10.1175/MWR-D-21-0321.1

LeCun, Y., Jackel, L. D., Boser, B., Denker, J. S., Graf, H. P., Guyon, I., Henderson, D., Howard, R. E., & Hubbard, W. (1989). *Handwritten Digit Recognition: Applications of Neural Net Chips and Automatic Learning*. *27*(11), 41–46.

Leroy, A., & Wheeler, M. C. (2008). Statistical Prediction of Weekly Tropical Cyclone Activity in the Southern Hemisphere, *Monthly Weather Review*, 136(10), 3637-3654, doi:10.1175/2008MWR2426.1.

Leonard, M., Westra, S., Phatak, A., Lambert, M., van den Hurk, B., McInnes, K., Risbey, J., Schuster, S., Jakob, D., & Stafford-Smith, M. (2014). A Compound Event Framework for Understanding Extreme Impacts. *WIREs Climate Change*, 5(1), 113–128.

Lesk, C., Rowhani, P., & Ramankutty, N. (2016). Influence of Extreme Weather Disasters on Global Crop Production. *Nature*, 529(7584), 84–87.

Lindström, G., Pers, C., Rosberg, J., Strömqvist, J., and Arheimer, B. (2010). Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales. *Hydrology Research*, 41(3-4):295–319.

Liu, W., Ke, Y., Liu, J., & Li, R. (2022). Model-Free Feature Screening and FDR Control With Knockoff Features. *Journal of the American Statistical Association*, 117(537), 428–443.

Loader, C. (1999). Local Regression and Likelihood. *Statistics and Computing Series*, Springer New York, New York, NY.

Lobell, D.B. (2013). Errors in Climate Datasets and Their Effects on Statistical Crop Models. *Agricultural and Forest Meteorology*, 170, 58–66.

Lowe, R., García-Díez, M., Ballester, J., Creswick, J., Robine, J. M., Herrmann, F. R., & Rodó, X. (2016). Evaluation of an early-warning system for heat wave-related mortality in Europe: Implications for sub-seasonal to seasonal forecasting and climate services. *International journal of environmental research and public health*, *13*(2), 206.

Luo, M., Lau, N. C., & Liu, Z. (2022). Different mechanisms for daytime, nighttime, and compound heatwaves in Southern China. *Weather and Climate Extremes*, 100449. doi:10.1016/j.wace.2022.100449

Magnusson, L., Bidlot, J., Lang, S. T. K., Thorpe, A., Wedi, N., & Yamaguchi, M. (2014). Evaluation of medium-range forecasts for hurricane Sandy. Monthly Weather Review, 142(5), 1962-1981. https://doi.org/10.1175/MWR-D-13-00228.1

Magnusson, L., Doyle, J. D., Komaromi, W. A., Torn, R. D., Tang, C. K., Chan, J. C., Yamaguchi, M., & Zhang, F. (2019). Advances in understanding difficult cases of tropical cyclone track forecasts. Tropical Cyclone Research and Review, 8(3), 109-122, doi: 10.1016/j.tcrr.2019.10.001

Magnusson, L., Majumdar, S., Emerton, R., Richardson, D., Alonso-Balmaseda, M., Baugh, C., Bechtold, P., Bidlot, J., Bonanni, A., Bonavita, M., Bormann, N., Brown, A., Browne, P., Carr, H., Dahoui, M., De Chiara, G., Diamantakis, M., Duncan, D., English, S., … Zsoter, E. (2021). Tropical cyclone activities at ECMWF. ECMWF Technical Memorandum 888.

Mahony, C.R., & Cannon, A.J. (2018). Wetter Summers Can Intensify Departures from Natural Variability in a Warming Climate. *Nature Communications*, 9(1), 783.

Maier-Gerber, M., Fink, A. H., Riemer, M., Schoemer, E., Fischer, C., & Schulz, B. (2021). Statistical-Dynamical Forecasting of Subseasonal North Atlantic Tropical Cyclone Occurrence. *Weather and Forecasting*, *36*(6), 2127-2142, doi:10.1175/WAF-D-21-0020.1.

Materia, S., Ardilouze, C., Prodhomme, C., Donat, M. G., Benassi, M., Doblas-Reyes, F. J., ... & Gualdi, S. (2021). Summer temperature response to extreme soil water conditions in the

Mediterranean transitional climate regime. *Climate Dynamics*, 1-21. doi:10.1007/s00382-021-05815-8

Mendelsohn, R., Emanuel, K., Chonabayashi, S., & Bakkensen, L. (2012). The impact of climate change on global tropical cyclone damage. *Nature Climate Change*, 2(3), 205–209. https://doi.org/10.1038/nclimate1357

Matsuno, T. (1966), Quasi-geostrophic motions in the equatorial area, Journal of the Meteorological Society of Japan., 44, 25–43, doi:10.2151/jmsj1965.44.1_25.

McAdam, R., Pérez-Aracil, J., Pelaez-Rodriguez, C., Squintu, A., Hansen, F., Torralba, V., Cavicchia, L., Zorita, E., Salcedo-Sanz, S., Scoccimarro, E. Optimisation-based driver detection and seasonal forecasting of European heatwaves. In preparation.

McNally, T., Bonavita, M., & Thépaut, J. (2014). The Role of Satellite Data in the Forecasting of Hurricane Sandy. *Monthly Weather Review*, *142*(2), 634-646. doi:10.1175/MWR-D-13-00170.1

McTaggart-Cowan, R., Deane, G. D., Bosart, L. F., Davis, C. A., & Galarneau, T. J., Jr. (2008). Climatology of Tropical Cyclogenesis in the North Atlantic (1948–2004), *Monthly Weather Review*, 136(4), 1284-1304, doi: 10.1175/2007MWR2245.1.

McTaggart-Cowan, R., Galarneau, T. J., Jr., Bosart, L. F., Moore, R. W., & Martius, O. (2013). A Global Climatology of Baroclinically Influenced Tropical Cyclogenesis, Monthly Weather Review, 141(6), 1963-1989, doi:10.1175/MWR-D-12-00186.1.

Meng, Y., Hao, Z., Feng, S., Zhang, X., & Hao, F. (2022). Increase in Compound Dry-Warm and Wet-Warm Events under Global Warming in CMIP6 Models. *Global and Planetary Change*, 210, 103773.

Mishra, A.K., & Singh, V.P. (2010). A Review of Drought Concepts. *Journal of Hydrology*, 391, 202-216. https://doi.org/10.1016/j.jhydrol.2010.07.012.

Nagler, T. (2018a). A Generic Approach to Nonparametric Function Estimation with Mixed Data. *Statistics & Probability Letters*, 137, 326–330.

Nagler, T. (2018b). Asymptotic Analysis of the Jittering Kernel Density Estimator. *Mathematical Methods of Statistics*, 27(1), 32–46.

Nairn, J.R., Fawcett, R.J.B., 2014. The Excess Heat Factor: A Metric for Heatwave Intensity and its Use in Classifying Heatwave Severity. International Journal of Environmental Research and Public Health, doi:10.3390/ijerph120100227.

O'Brien, R. and Ishwaran, H. (2019). A Random Forests Quantile Classifier for Class Imbalanced Data. *Pattern Recognition*, 90, 232–249.

Pal, J.S. and Eltahir, E.A.B. (2016), "Future temperature in southwest Asia projected to exceed a threshold for human adaptability", *Nature Climate Change*, Vol. 6 No. 2, pp. 197–200.

Pedro-Monzonís, M., Solera, A., Ferrer, J., Estrela, T., & Paredes-Arquiola, J. (2015). A Review of Water Scarcity and Drought Indexes in Water Resources Planning and Management. *Journal of Hydrology*, 527, 482-493. https://doi.org/10.1016/j.jhydrol.2015.05.003.

Peng, J., Muller, J.P., Blessing, S., Giering, R., Danne, O., Gobron, N., Kharbouche, S., Ludwig, L., Müller, B., Leng, G., You, Q., Duan, Z., and Dadson, S. (2019). Can We Use Satellite-Based FAPAR to Detect Drought? *Sensors (Basel, Switzerland)*, 19(17):3662.

Pérez-Aracil, J., Camacho-Gómez, C., Lorente-Ramos, E., Marina, C. M., Cornejo-Bueno, L. M., & Salcedo-Sanz, S. (2023). New Probabilistic, Dynamic Multi-Method Ensembles for Optimization Based on the CRO-SL. *Mathematics*, *11*(7), 1666.

Pérez-Aracil, J., Pelaez-Rodriguez, C., McAdam, R., Squintu, A., Torralba, V., Marina, C. M., Lorente-Ramos, E., Cavicchia, L., Giuliani, M., Zorita, E., Hansen, F., Barriopedro, D., Garcia-Herrera, R., Gutiérrez, P. A., Xoplaki, E., Castelletti, A., Salcedo-Sanz, S. Unveiling Large-Scale Heatwave Drivers with Spatio-Temporal Cluster-Optimization-Based Feature Selection. In preparation.

Perkins, S.E., & Alexander, L.V. (2013). On the Measurement of Heat Waves. *Journal of Climate*, 26(13), 4500–4517.

Porter, J.R., & Semenov, M.A. (2005). Crop Responses to Climatic Variation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 360(1463), 2021–2035.

Prodhomme, C., Doblas-Reyes, F., Bellprat, O., & Dutra, E. (2016). Impact of land-surface initialization on sub-seasonal to seasonal forecasts over Europe. *Climate dynamics*, *47*, 919-935.

Prodhomme, C., Materia, S., Ardilouze, C., White, R. H., Batté, L., Guemas, V., Fragkoulidis, G., & García-Serrano, J. (2022). Seasonal prediction of European summer heatwaves. *Climate Dynamics*, *58*(7), 2149-2166. doi:10.1007/s00382-021-05828-3

Qin, H., Wang, C., Zhao, K., & Xi, X. (2018). Estimation of the Fraction of Absorbed Photosynthetically Active Radiation (fPAR) in Maize Canopies Using LiDAR Data and Hyperspectral Imagery. *PLoS ONE*, 13(5): e0197510.

Raymond, C., Horton, R. M., Zscheischler, J., Martius, O., AghaKouchak, A., Balch, J., Bowen, S. G., Camargo, S. J., Hess, J., Kornhuber, K., Oppenheimer, M., Ruane, A. C., Wahl, T., &

White, K. (2020). Understanding and managing connected extreme events. *Nature Climate Change*, 10(7), 611–621.

Riemer, M., & Jones, S. C. (2014). Interaction of a tropical cyclone with a high-amplitude, midlatitude wave pattern: Waviness analysis, trough deformation and track bifurcation. *Quarterly Journal of the Royal Meteorological Society*, 140, 1362–1376.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference*, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, 234–241.

Russo, S., Dosio, A., Graversen, R. G., Sillmann, J., Carrao, H., Dunbar, M. B., ... & Vogt, J. V. (2014). Magnitude of extreme heat waves in present climate and their projection in a warming world. *Journal of Geophysical Research: Atmospheres*, 119(22), 12-500. doi:10.1002/2014JD022098.

Russo, S., Sillmann, J., & Fischer, E. M. (2015). Top ten European heatwaves since 1950 and their occurrence in the coming decades. *Environmental Research Letters*, 10(12), 124003. doi:10.1088/1748-9326/10/12/124003.

Russo, S., Sillmann, J., & Sterl, A. (2017). Humid heat waves at different warming levels. *Scientific Reports*, 7(1), 1–7. doi:10.1038/s41598-017-07536-7.

Salcedo-Sanz, S., Del Ser, J., Landa-Torres, I., Gil-López, S., & Portilla-Figueras, J. A. (2014). The coral reefs optimization algorithm: a novel metaheuristic for efficiently solving optimization problems. *The Scientific World Journal*, 2014.

Salcedo-Sanz, S. (2017). A review on the coral reefs optimization algorithm: new development lines and current applications. *Progress in Artificial Intelligence*, 6, 1–15.

Schäfler, A., Craig, G., Wernli, H., Arbogast, P., Doyle, J. D., McTaggart-Cowan, R., Methven, J., Rivière, G., Ament, F., Boettcher, M., & Bramberger, M. (2018). The North Atlantic waveguide and downstream impact experiment. *Bulletin of the American Meteorological Society*, 99(8), 1607–1637. doi:10.1175/BAMS-D-17-0003.1.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks, 61*, 85–117.

Schreck, C. J., III, Molinari, J., & Mohr, K. I. (2011). Attributing Tropical Cyclogenesis to Equatorial Waves in the Western North Pacific. *Journal of the Atmospheric Sciences*, 68(2), 195–209. doi:10.1175/2010JAS3396.1.

Schreck, C. J., III, Molinari, J., & Aiyyer, A. (2012). A Global View of Equatorial Waves and Tropical Cyclogenesis. *Monthly Weather Review*, 140(3), 774–788. doi:10.1175/MWR-D-11-00110.1.

Schumacher, D. L., Keune, J., Van Heerwaarden, C. C., Vilà-Guerau de Arellano, J., Teuling, A. J., & Miralles, D. G. (2019). Amplification of mega-heatwaves through heat torrents fuelled by upwind drought. *Nature Geoscience*, 12(9), 712–717. doi:10.1038/s41561-019-0431-6.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.

Scoccimarro, E., Fogli, P. G., & Gualdi, S. (2017). The role of humidity in determining scenarios of perceived temperature extremes in Europe. *Environmental Research Letters*, *12*(11), 114029. doi:10.1088/1748-9326/aa8cdd

Scoccimarro, E., Cattaneo, O., Gualdi, S. et al. (2023) Country-level energy demand for cooling has increased over the past two decades. Commun Earth Environ 4, 208 https://doi.org/10.1038/s43247-023-00878-3

Scoccimarro E., Lanteri P., Cavicchia L. (2024). Freddy: breaking record for tropical cyclone precipitation? Environ. Res. Lett. 19 064013 DOI 10.1088/1748-9326/ad44b5

Serifi, A., Günther, T., & Ban, N. (2021). Spatio-Temporal Downscaling of Climate Data Using Convolutional and Error-Predicting Neural Networks. *Frontiers in Climate*, 3.

Song, J., Klotzbach, P. J., & Duan, Y. (2022). Statistical linkage between coastal El Niño–Southern Oscillation and tropical cyclone formation over the western North Pacific. *Atmospheric Science Letters*, 23(2). doi:10.1002/asl.1071

Sousa P.M., Trigo R. M., Barriopedro D., Soares P. M. M., Santos J. A. (2018). European temperature responses to blocking and ridge regional patterns. *Climate Dynamics*, 50, 1-2, 457-477, doi: 10.1007/s00382-017-3620-2

Spinoni, J., Naumann, G., Vogt, J. V., & Barbosa, P. (2016). Meteorological droughts in Europe: Events and impacts—past trends and future projections. Publications Office of the European Union. ISBN-13: 978-92-79-55097-3.

Stefanon, M., D'Andrea, F., & Drobinski, P. (2012). Heatwave classification over Europe and the Mediterranean region. *Environmental Research Letters*, *7*(1), 014023.

Tay, J. K., Narasimhan, B., & Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, 106.

Tenenhaus, A., & Tenenhaus, M. (2011). Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2), 257–284.

Thiemig, V., Gomes, G. N., Skøien, J. O., Ziese, M., Rauthe-Schöch, A., Rustemeier, E., Rehfeldt, K., Walawender, J. P., Kolbe, C., Pichon, D., Schweim, C., & Salamon, P. (2022). EMO-5: A high-resolution multi-variable gridded meteorological dataset for Europe. *Earth System Science Data*, 14(7), 3249–3272.

Thomas, N. P., Bosilovich, M. G., Marquardt Collow, A. B., Koster, R. D., Schubert, S. D., Dezfuli, A., & Mahanama, S. P. (2020). Mechanisms associated with daytime and nighttime heat waves over the contiguous United States. Journal of Applied Meteorology and Climatology, 59(11), 1865-1882. doi:10.1175/JAMC-D-20-0053.1

Tilloy, A., Paprotny, D., Grimaldi, S., Gomes, G., Bianchi, A., Lange, S., Beck, H., & Feyen, L. (2024). HERA: A high-resolution pan-European hydrological reanalysis (1950–2020). *Earth System Science Data Discussions*, pp. 1–38.

Tong, Z., Cai, Z., Yang, S., & Li, R. (2023). Model-free conditional feature screening with FDR control. *Journal of the American Statistical Association*, 118(544), 2575–2587.

Toreti, A., Belward, A., Perez-Dominguez, I., Naumann, G., Luterbacher, J., Cronie, O., Seguini, L., Manfron, G., Lopez-Lozano, R., Baruth, B., van den Berg, M., Dentener, F., Ceglar, A., Chatzopoulos, T., & Zampieri, M. (2019a). The exceptional 2018 European water seesaw calls for action on adaptation. *Earth's Future*, 7(6), 652–663.

Toreti, A., Cronie, O., & Zampieri, M. (2019b). Concurrent climate extremes in the key wheat producing regions of the world. *Scientific Reports*, 9(1), 5493.

Torralba, V., Materia, S., Cavicchia, L., Álvarez-Castro, M.C., Prodhomme, C., McAdam, R., Scoccimarro, E. and Gualdi, S., 2024. Nighttime heat waves in the Euro-Mediterranean region: definition, characterisation, and seasonal prediction. Environmental Research Letters, 19(3), p.034001.

Townes, F. W. (2019). Generalized principal component analysis.

Townes, F. W., Hicks, S. C., Aryee, M. J., & Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, 20(1), 295.

van Lieshout, M. N. M. (2011). A J-function for inhomogeneous point processes. *Statistica Neerlandica*, 65(2), 183–201.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432.

Vicente-Serrano, S., Beguería, S., & López-Moreno, J. I. (2010). A multiscalar drought index sensitive to global warming: The standardized precipitation evapotranspiration index. *Journal of Climate*, 23, 1696–1718.

Vicente-Serrano, S. M., & Beguería, S. (2016). Comment on "Candidate distributions for climatological drought indices (SPI and SPEI)" by James H. Stagge et al. *International Journal of Climatology*, 36(4), 2120–2131.

Vogel, M. M., Hauser, M., & Seneviratne, S. I. (2020). Projected changes in hot, dry, and wet extreme events' clusters in CMIP6 multi-model ensemble. *Environmental Research Letters*, 15(9), 094021.

Wahl, T., Jain, S., Bender, J., Meyers, S. D., & Luther, M. E. (2015). Increasing risk of compound flooding from storm surge and rainfall for major US cities. *Nature Climate Change*, 5(12), 1093–1097.

Walsh, K. J., McBride, J. L., Klotzbach, P. J., Balachandran, S., Camargo, S. J., Holland, G., et al. (2016). Tropical cyclones and climate change. *Wiley Interdisciplinary Reviews: Climate Change*, 7(1), 65–89. https://doi.org/10.1002/wcc.371

Wang, Z., Zhang, G., Dunkerton, T. J., & Jin, F. F. (2020). Summertime stationary waves integrate tropical and extratropical impacts on tropical cyclone activity. *Proceedings of the National Academy of Sciences*, *117*(37), 22720-22726, doi:10.1073/pnas.2010547117.

Warton, D. I., Wright, S. T., & Wang, Y. (2012). Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, 3, 89–101.

Wilks, D. S. (2016). Modified "Rule N" procedure for principal component (EOF) truncation. *Journal of Climate*, 29(8), 3049–3056.

Xu, S. G., & Reich, B. J. (2023). Bayesian nonparametric quantile process regression and estimation of marginal quantile effects. *Biometrics*, 79(1), 151–164.

Zadeh, L.A., 1965. Fuzzy sets. Information and control, 8(3), pp.338-353.

Zampieri, M., Ceglar, A., Dentener, F., & Toreti, A. (2017). Wheat yield loss attributable to heat waves, drought, and water excess at the global, national, and subnational scales. *Environmental Research Letters*, 12(6), 064008.

Zhang, G., Wang, Z., Dunkerton, T. J., Peng, M. S., & Magnusdttir, G. (2016). Extratropical Impacts on Atlantic Tropical Cyclone Activity, *Journal of the Atmospheric Sciences*, 73(3), 1401-1418, doi:10.1175/JAS-D-15-0154.1.

Zhang, G., Wang, Z., Peng, M. S., & Magnusdottir, G. (2017). Characteristics and Impacts of Extratropical Rossby Wave Breaking during the Atlantic Hurricane Season, *Journal of Climate*, 30(7), 2363-2379, doi:10.1175/JCLI-D-16-0425.1.

Zhang, R., Sun, C., Zhu, J., Zhang, R., & Li, W. (2020). Increased European heat waves in recent decades in response to shrinking Arctic Sea ice and Eurasian snow cover. *NPJ Climate and Atmospheric Science*, *3*(1), 7.

Zhang, R. Z., Jia, X. J., & Qian, Q. F. (2022). Seasonal forecasts of Eurasian summer heat wave frequency. *Environmental Research Communications*, *4*(2), 025007.

Zhang, W., Luo, M., Gao, S., Chen, W., Hari, V., & Khouakhi, A. (2021). Compound hydrometeorological extremes: Drivers, mechanisms, and methods. *Frontiers in Earth Science*, 9.

Zhu, L., Xu, K., Li, R., & Zhong, W. (2017). Projection correlation between two random vectors. *Biometrika*, 104(4), 829–843.

Zou, H., & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4), 1733–1751.

Zscheischler, J., & Fischer, E. M. (2020). The record-breaking compound hot and dry 2018 growing season in Germany. *Weather and Climate Extremes*, 29, 100270.

Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R. M., van den Hurk, B., AghaKouchak, A., Jézéquel, A., Mahecha, M. D., Maraun, D., Ramos, A. M., Ridder, N. N., Thiery, W., & Vignotto, E. (2020). A typology of compound weather and climate events. *Nature Reviews Earth & Environment*, 1(7), 333–347.

Zscheischler, J., & Seneviratne, S. I. (2017). Dependence of drivers affects risks associated with compound events. *Science Advances*, 3(6), e1700263.

Zscheischler, J., Westra, S., van den Hurk, B. J. J. M., Seneviratne, S. I., Ward, P. J., Pitman, A., AghaKouchak, A., Bresch, D. N., Leonard, M., Wahl, T., & Zhang, X. (2018). Future climate risk from compound events. *Nature Climate Change*, 8(6), 469–477.

Zuo, J., Pullen, S., Palmer, J., Bennetts, H., Chileshe, N., & Ma, T. (2015). Impacts of heat waves and corresponding measures: a review. *Journal of Cleaner Production*, *92*, 1-12. doi: 10.1016/j.jclepro.2014

APPENDIX A4

A4.1 Data-Driven Forecast Skill (2004-2022)

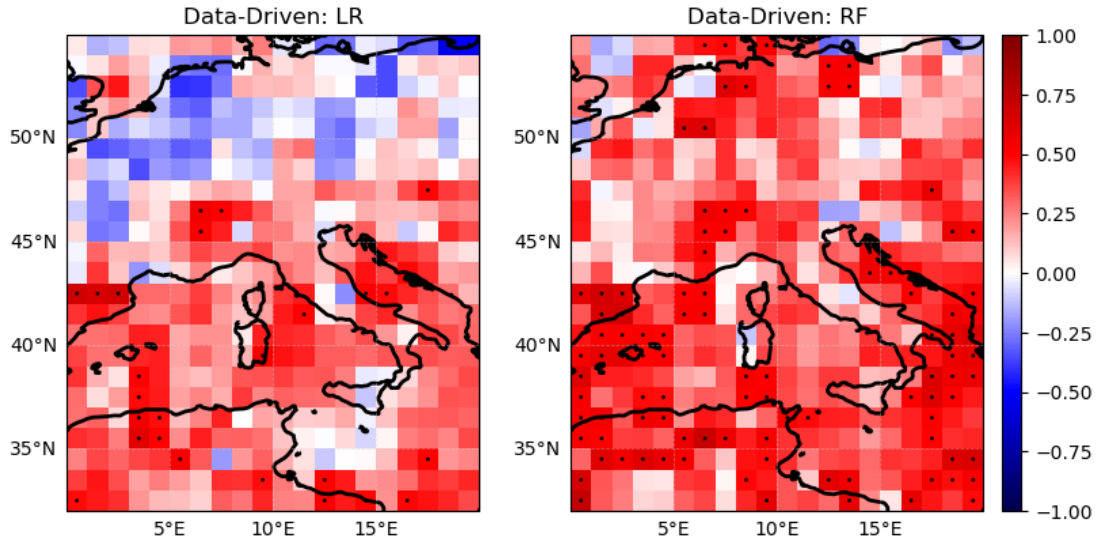


Figure A4.1: *Correlation skill of Data-Driven HW Seasonal Forecasts over 2004-2022, for comparison with equivalents from the dynamical (ECMWF-SEA5) and hybrid systems (Fig 4.11). LR - Logistic Regression; RF - Random Forest.*

A4.2 Night-time heatwave clusters

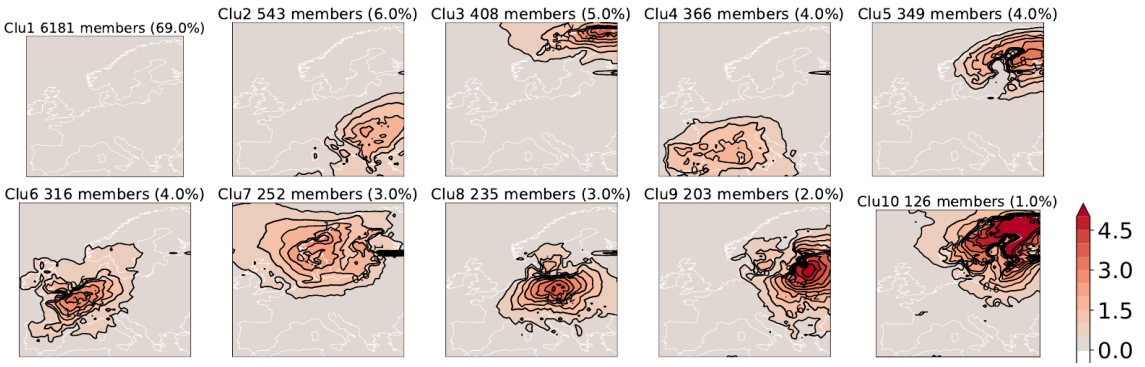


Figure A4.2*: Night-time HW clusters over the European domain, coloured by their average intensity (contours correspond to 0.3°C intervals).*

A4.3 Seasonal Forecast skill of day and night-time extremes

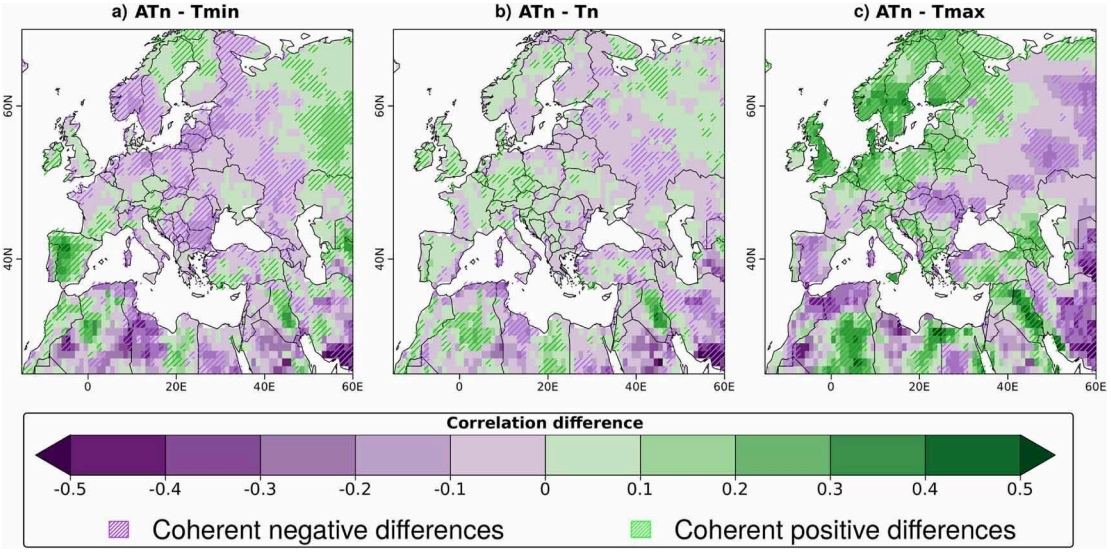


Figure A4.3: *Differences between the correlation maps of the multi-model seasonal predictions for the Apparent temperature at night (ATn) and the corresponding correlation maps but for (a) Tmin, (b) Tn, and (c) Tmax for the 1993–2016 period in the 15MJJA season. These correlation maps are computed with ERA5 as an observational reference. Hatched areas indicate where the four individual prediction systems agree in the positive (green lines) or negative (purple lines) correlation differences. The seasonal forecasts are issued on the 1st of May.*

APPENDIX A6

A6.1 Relatively wet and warm late winters followed by dry and warm springs

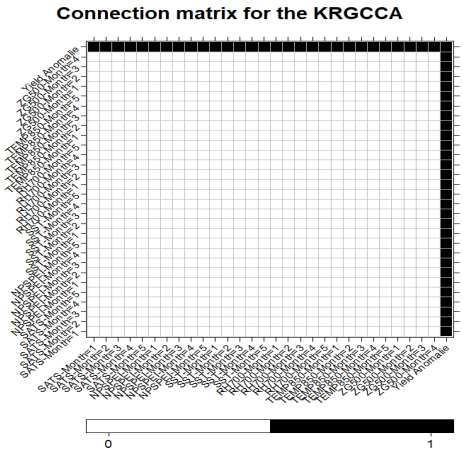


Figure A6.1: *Connection matrix used for the KRGCCA employed for the analysis of the relatively wet and warm late winters followed by dry springs.*

A6.2 Dry winters followed by hot summers

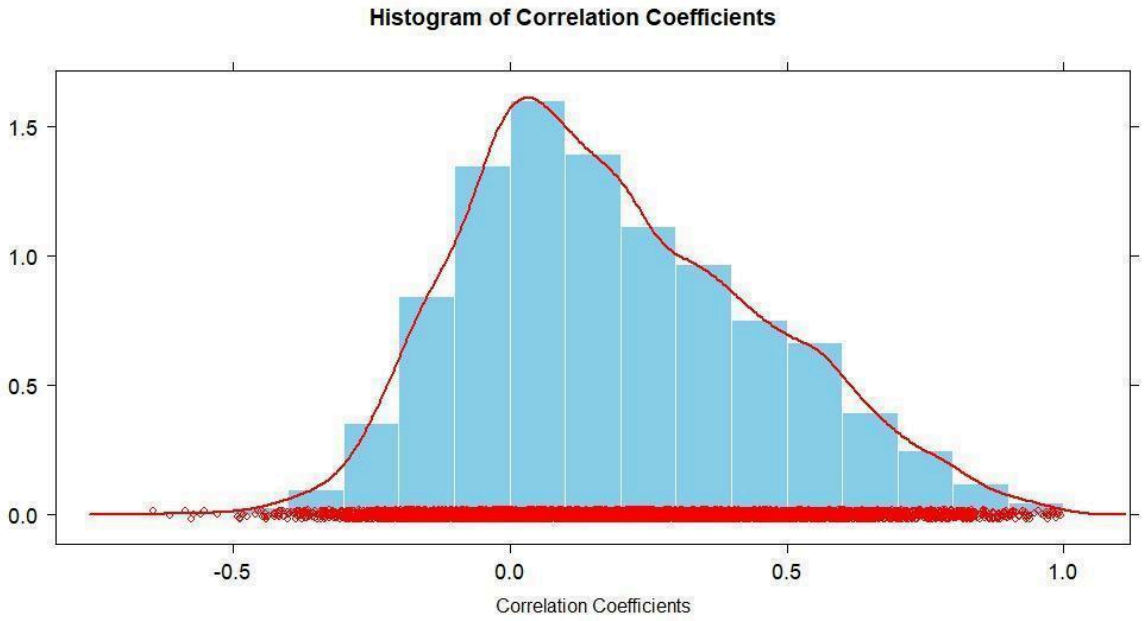


***Figure A6.2**: Density histogram of correlation of residuals from the NUTS3 regions utilized for the SUR model.*

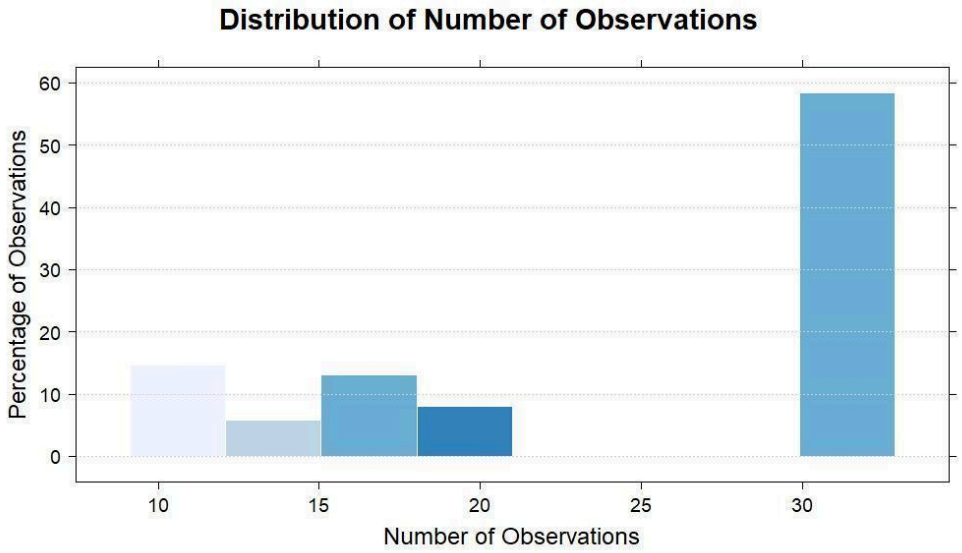


***Figure A6.3:** Histogram for the number of observations in the NUTS3 regions.*

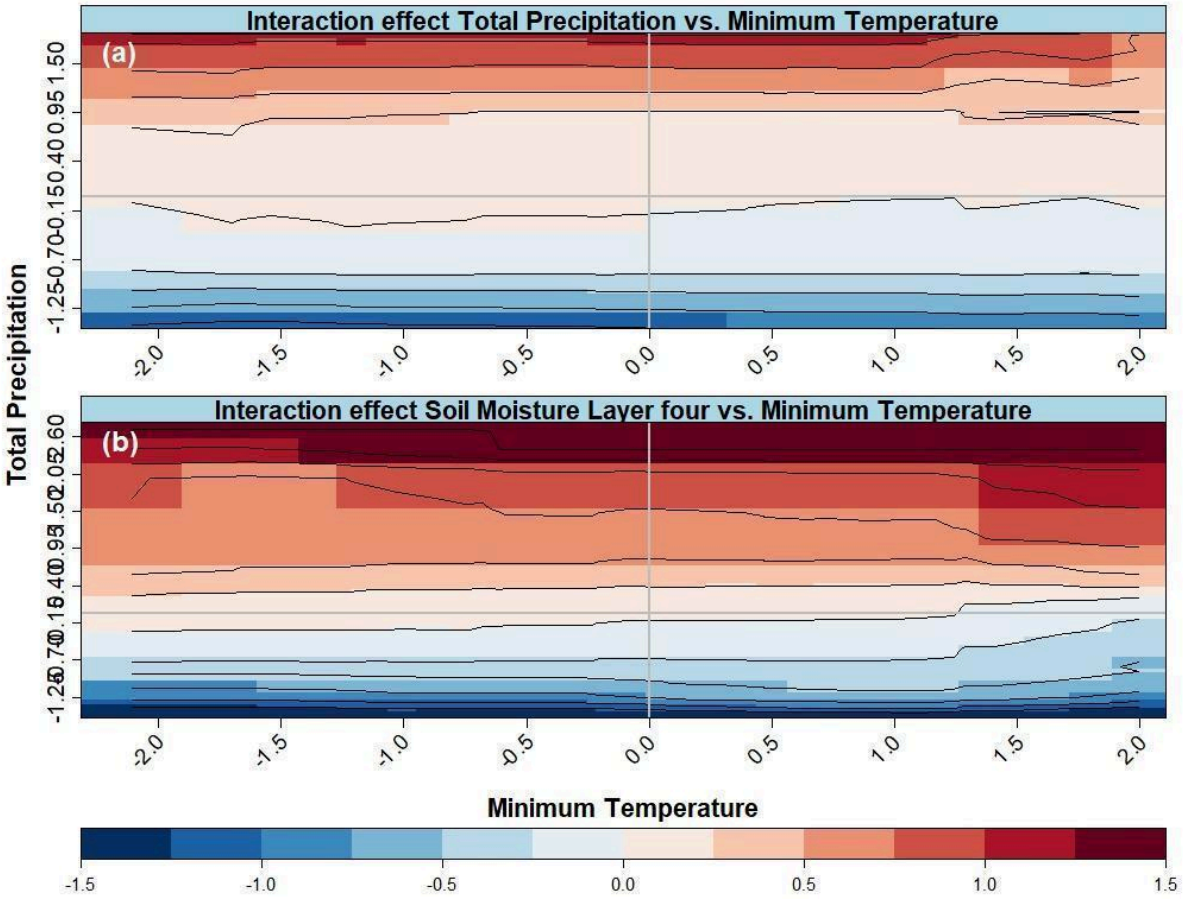


Figure A6.4: *Interaction Effect of Minimum Temperature with (a) Total Precipitation and (b) Soil Moisture Layer 4.*

A6.3 Wet and warm springs

Table A6.1: *Statistics for the EOF analysis performed for the Alpine region variables.*

Variable	Lag	Number Components	Explained Variance (%)
Runoff	0	4	89.40
Maximum Temperature	0	3	85.59
Maximum Temperature	1	3	85.02
Maximum Temperature	2	3	82.84
Maximum Temperature	3	2	75.73
Minimum Temperature	0	3	77.49
Minimum Temperature	1	3	78.96
Minimum Temperature	2	4	82.37
Minimum Temperature	3	3	78.84
Total Precipitation	0	7	61.77
Total Precipitation	1	6	63.44
Total Precipitation	2	6	66.37
Total Precipitation	3	4	61.02

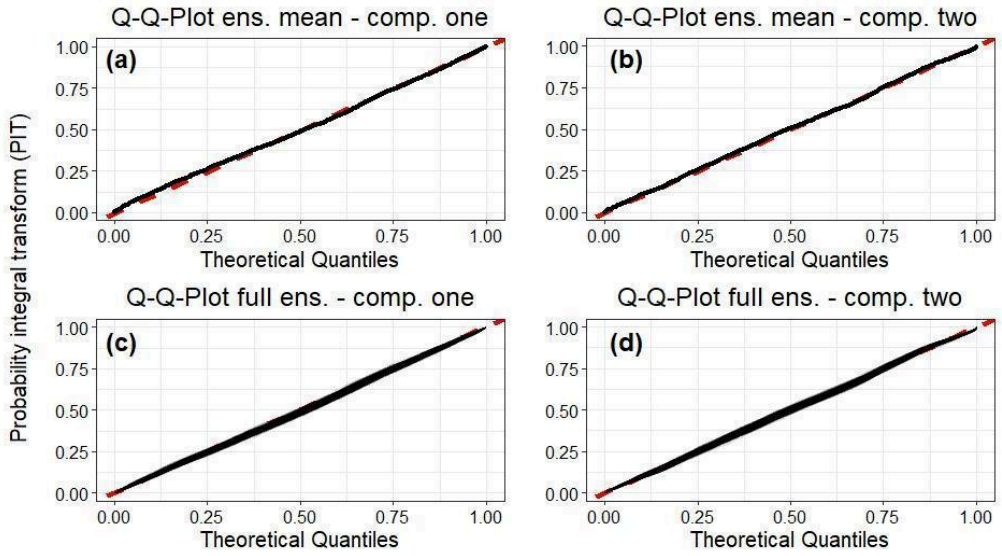


Figure A6.5: *Q-Q-plot of the estimated QUINN model for the river runoffs for the first component displayed in (a) and (b) and the second component in (c) and (d).*

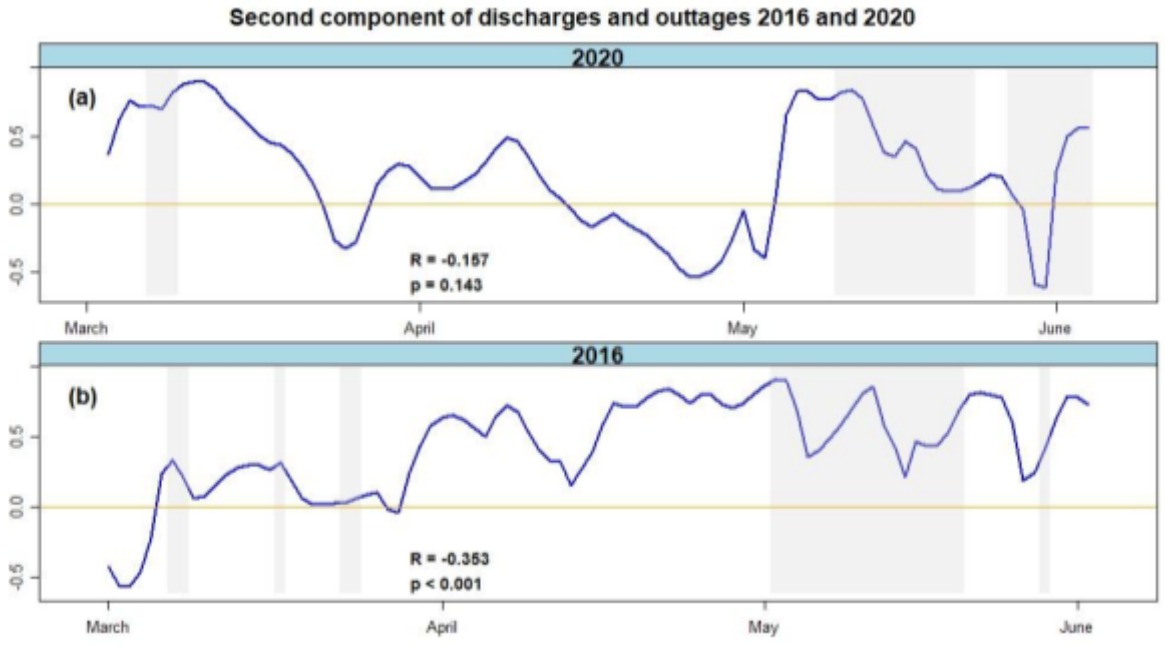


Figure A6.6: *Displayed in blue is the second component of discharges (Figure 6.13) and overlaid in gray shadows the reported outages from ENTSO-E for (a) 2020 and (b) 2016. R corresponds to the biserial correlation and p denotes the p-value. Both correlations have been calculated for a lag of three days for which the maximum lagged correlation is observed.*

A6.4 Nonparametric SPEI

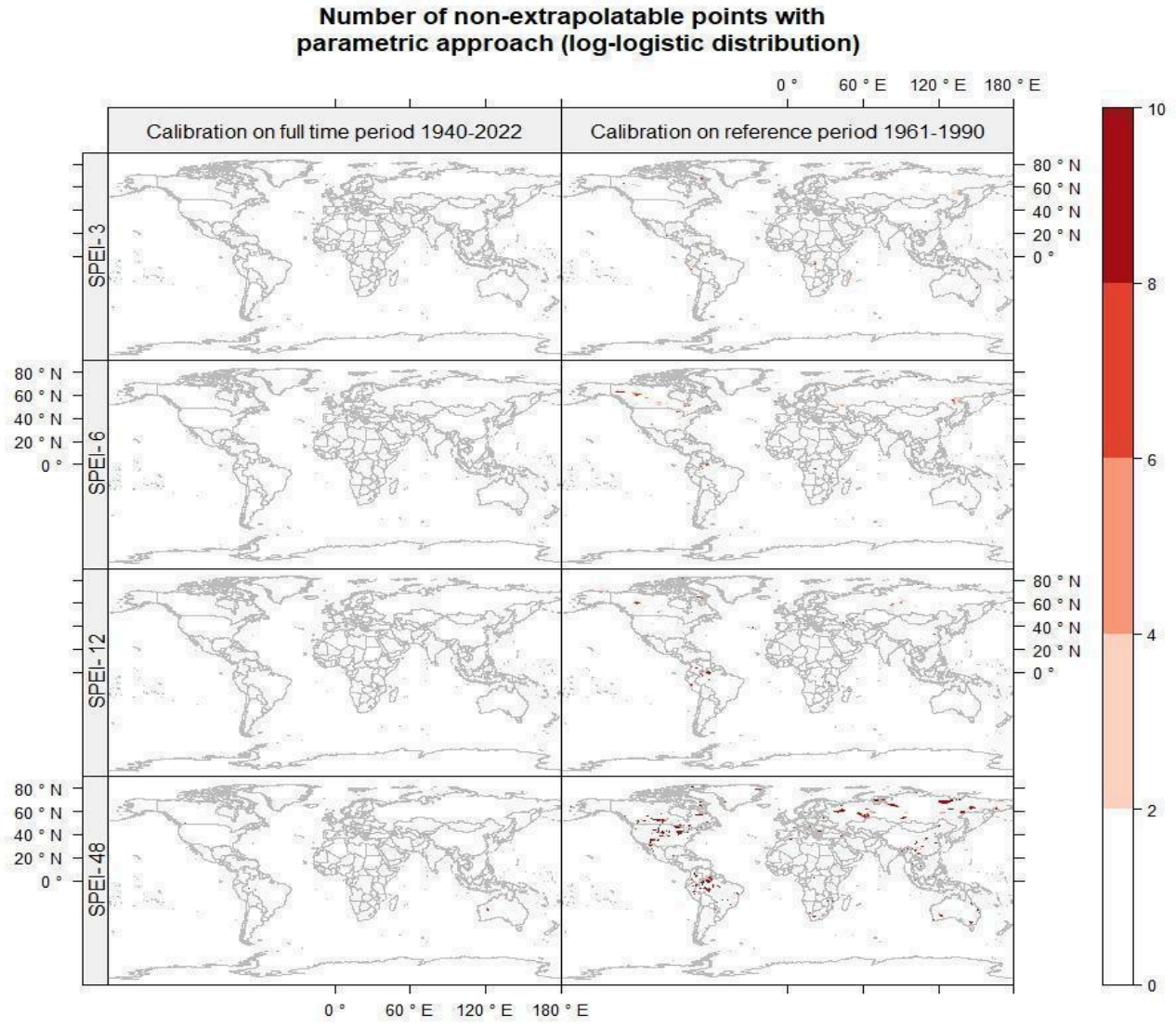


Figure A6.7: *Number of non-extrapolatable points of SPEI using the log-logistic distribution*

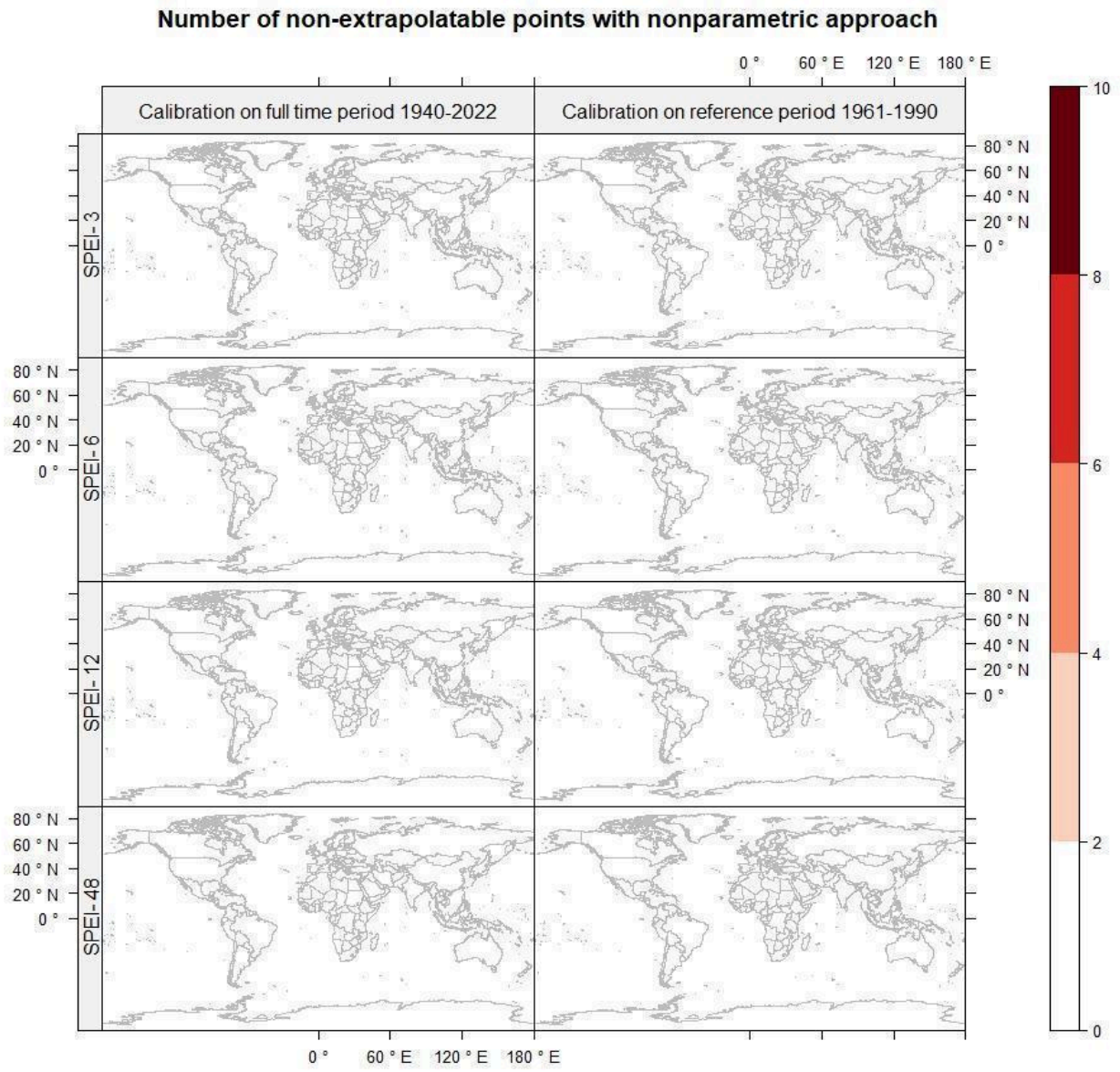


Figure A6.8: *Same as Figure A6.6, but for the NPSPEI.*

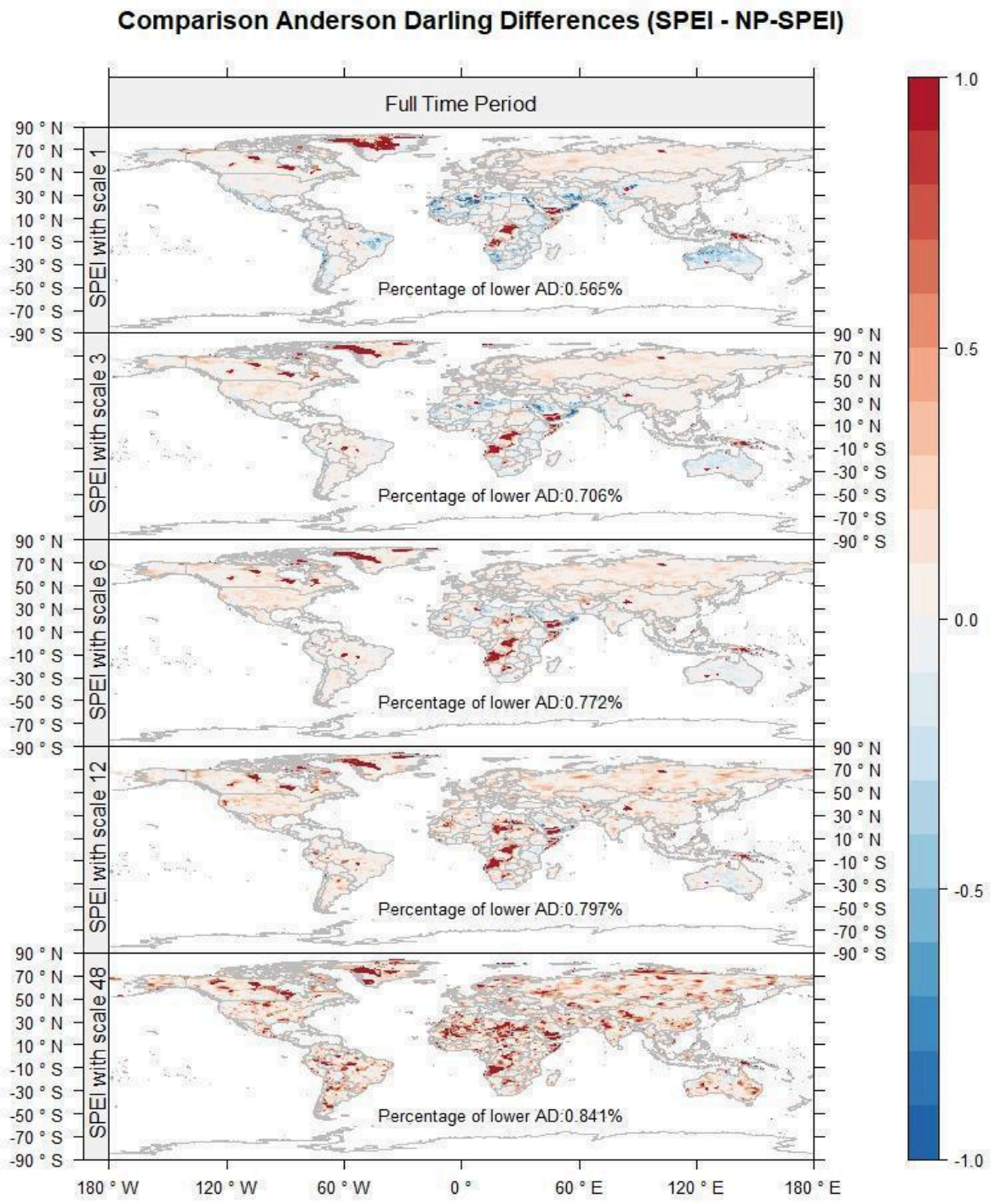
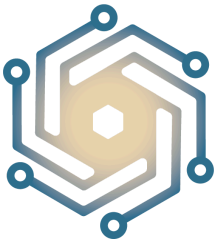


Figure A6.9: *Difference of Anderson Darling statistics for SPEI and NPSPEI. Positive value indicate that the NPSPEI produces a smaller statistics and hence a better fit with respect to the standard normal distribution.*



CLINT

CLIMATE INTELLIGENCE

This project is part of the H2020 Programme supported by the European Union, having received funding from it under Grant Agreement No 101003876