

D3.2 PRELIMINARY AI-ENHANCED EXTREME EVENTS DETECTION

November 2023



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 #:	D3.2	
Deliverable Type:	Document	
Contractual Date of Delivery:	30 December 2023	
Actual Date of Delivery:	30 December 2023	
Title of Document:	Preliminary AI-enhanced Extreme Events detection	
Responsible partner:	СМСС	
Author(s):	Enrico Scoccimarro, Leone Cavicchia, Ronan McAdam, Paolo Lanteri, Antonello Squintu, Michael Maier-Gerber, Linus Magnusson, Felicitas Hansen, Jorge Pérez Aracil, César Peláez Rodríguez, Matteo Giuliani, Martina Merlo, Yiheng Du, Niklas Luther, Elena Xoplaki, Odysseas Vlachopoulos	
Content of this report:	Description of the first 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.	



Γ

Document revisions			
Author	Revision content	Date	
Antonello Squintu and Enrico Scoccimarro	D3.2_v01 – First draft of chapter 2, structure of the document	07/09/2023	
Martina Merlo, Matteo Giuliani, Andrea Castelletti	First draft of chapter 5	23/10/2023	
All authors	First drafts of all chapters	24/10/2023	
Martina Merlo, Matteo Giuliani, Andrea Castelletti	Complete draft of chapter 5	20/11/2023	
All authors	D3.2_v02 - Contributions to all chapters completed	28/11/2023	
Antonello Squintu	ntu D3.2_v03 - Finalised first draft of the deliverable		
Harilaos Loukos, Schalk Jan van Andel, Antonello Squintu, all authors	Iarilaos Loukos, chalk Jan van Andel, intonello Squintu, allD3.2_v04 - Assimilation of automated language review and internal review		
Guido Ascenso, Andrea Castelletti	D3.2_vF – Final review	22/12/2023	



TABLE OF CONTENT

TABLE OF CONTENT	4
LIST OF FIGURES	7
LIST OF ACRONYMS	13
EXECUTIVE SUMMARY	16
1 INTRODUCTION	19
2 TROPICAL CYCLONES	20
2.1 Overview	20
2.2 Datasets and candidate drivers	20
2.2.1 Datasets	20
2.2.2 Candidate drivers	20
2.3 Long term horizon	21
2.3.1 Indices	21
2.3.2 Skills and performance of existing indices	21
2.3.3 Algorithm	23
2.3.4 Results: improved indices and relevant drivers	24
2.4 Short term horizon	26
2.4.1 Target variable	26
2.4.2 Skills and performance of existing forecasts	26
2.4.3 Algorithm	28
2.4.3.1 Baselines	28
2.4.3.2 Convolutional-based approach	28
2.4.4 Results: ML-based forecasts and relevant drivers	29
2.5 Summary and outlook	31
3 EXTRATROPICAL TRANSITION OF TROPICAL CYCLONES	33
3.1 Overview	33
3.2 Datasets, candidate drivers and target variable	34
3.3 Skills and performance of existing forecasts	36
3.4 Algorithm	36
3.4.1 Decision trees and random forests	36
3.4.2 Logistic regression	36
3.5 Results: ML-based forecasts and relevant drivers	37



3.6 Summary and outlook	38
4 HEATWAVES AND WARM NIGHTS	39
4.1 Overview	39
4.2 Datasets, candidate drivers and indices	40
4.2.1 Datasets	40
4.2.2 Candidate Drivers	40
4.3 Skills and performance of existing indices and forecasts	41
4.3.1 Heatwaves	41
4.3.2 Warm Nights	42
4.4 Algorithm: feature selection framework	44
4.5 Results: relevant drivers	49
4.5.1 Feature Selection Framework: Results for Lake Como	49
4.5.2 Spatial Clustering of Heat Extremes: common drivers on regional-scales	51
4.6 Summary and outlook	54
5 EXTREME DROUGHTS	57
5.1 Overview	57
5.2 Datasets and indices	57
5.3 Skills and performance of existing indices	58
5.4 Algorithms	61
5.5 Results	64
5.5.1 Improved indices and relevant drivers	64
5.5.2 Results: AI-enhanced drought indices	67
5.6 Summary and outlook	67
6 COMPOUND EVENTS AND CONCURRENT EXTREMES	69
6.1 Overview	69
6.2 Datasets and candidate drivers	69
6.2.1 Datasets	69
6.2.2 Candidate drivers	70
6.3 Existing indices	71
6.3.1 Climate indices	71
6.3.2 Statistical measures for dependence	71
6.4 Algorithms	72
6.4.1 Time series clustering algorithms	72



6.4.2 Regularised generalised canonical correlation analysis (RGCCA)	
6.4.3 Imbalanced random forests	73
6.5 Results	75
6.5.1 Compound events	75
6.5.1.1 Agro-climatic sub-regions	75
6.5.1.2 Relevant drivers	76
6.5.1.3 Learning compound event definition	83
6.5.2 Concurrent extremes	87
6.5.2.1 Non-parametric SPEI	87
6.5.2.2 Clustering of droughts and heatwave	90
6.5.2.3 Deep Learning based interpreter of J-functions	91
6.5.2.4 Example application	92
6.6 Summary and outlook	93
7 GENERAL SUMMARY AND OUTLOOK	94
REFERENCES	95
APPENDIX A4	111
APPENDIX A5	113
A5.1	113
A5.2	118
APPENDIX A6	121
A6.1 Nonparametric SPEI for the wichita time series	121
A6.2 Number of non-extrapolatable points of the nonparametric SPEI	122
A6.3 Comparison of nonparametric SPEI and SPEI on the reference period	123
A6.4 Deep Learning based training.	123





LIST OF FIGURES

Figure 2.1: Left: Boxplots showing ensemble spread of correlation between yearly time series of number of detected TCs and number of TC detected with the three GPIs, calculated on historical simulations. Each box plot is related to a different ocean basin. The star indicates the same correlation, but calculated on ERA5 data. Right: Same as left column but for future projections. The star here represents the ensemble mean of correlations calculated on historical simulations.

Figure 2.2: Future trends of directly detected TCs and number of TC predicted with GPIs.

Figure 2.3: Pareto fronts obtained optimising the algorithm on ERA5 (left) and MERRA (right).

Figure 2.4: Top-left: comparison between spatial distribution on IBTrACS and solution (a). Topright: Interannual variability curves for solution (a). Bottom-left: Same as Top-Right but for solution (b) on ERA5. Bottom-right: Same as Bottom-left but on MERRA2.

Figure 2.5: (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 2.6: (a) Brier skill score (in %) of tropical storm strike 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 2.7: 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 2.8: Same as in Figure 2.6a, but including (a) the baseline and (b) the CNN-based ML models, respectively.

Figure 2.9: Qualitative summary of lessons learned: (a) the best-performing forecasting approaches for the lead times considered, and (b) evaluation of the impact of predictors when included in the predictor set.

Figure 3.1: IBTrACS positions cyclones in the extratropical stage for April 2016-December 2022.

Figure 3.2: 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 3.3: (a) ROC curves for all models with AUC scores in the legend. (b) As in Figure 3.2, but including the results for the ML models sorted by BS.



Figure 3.4: 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: Seasonal forecasts of the number of days of summer HWs in the Lake Como region over the 1993-2016 period. Values shown here are the annual sums of HW occurrence shown in Figure 1. The CMCC-35 ensemble spread for each year is represented by the box plots, and the median denoted (orange line).

Figure 4.2: Seasonal forecasts of daily HW occurrence in the Lake Como region over the 1993-2016 period. Forecasts are taken from the CMCC-35 dynamical system for the summer period initialised on May 1st CMCC-35. The forecast shown (red) is the ensemble member with the highest correlation score of number of summer HW days (see Figure 4.1). ERA5 (black) data is used as the benchmark for validation. HWs are defined relative to the daily 90th-percentiles of the 1993-2016 period.

Figure 4.3: Seasonal Forecast skill of warm nights over Europe. a) Domains: Northern Europe (NE), Western Europe (WE), Central Europe (CE), Eastern Europe (EE), Mediterranean (MED), Northern Africa (NAF), Middle East (ME), full European domain (ALL) b) Ensemble mean correlation values for the seasonal predictions of the b) HWMI and c) NDQ90 based on the ATn. Results for the seasonal predictions from individual C3S seasonal prediction systems (CMCC-35, DWD-21, ECMWF-5, MF-7) and their multimodel combination (MM) are shown for each region. The seasonal forecasts were issued on the 1st of May and the observational reference is ERA5. This assessment corresponds to the 15MJJA season in the 1993-2016 period. Significant correlations at the 95% confidence level are marked with an asterisk. From Torralba et al. (in review).

Figure 4.4: Schematic of Feature Selection Framework. First, the dimensionality of the problem is reduced by clustering of candidate driver variables (e.g. MSL; not all are shown here). The area-average of each cluster is used as input to the CRO optimisation algorithm, which works to identify the best combinations of predictors to optimise F1-skill score of a logistic regression model (or other, ML models). The optimal solutions can then be used as training for other ML models to recreate the test data.

Figure 4.5: Gradient Booster classifier forecast of daily HW occurrence in the Lake Como region over the 2011-2022 test period. The forecast shown (red) corresponds to the solution with the highest validation score from the evolutionary algorithm. ERA5 (black) data is used as the benchmark for validation. HWs are defined relative to the daily 90th-percentiles of the 1981-2010 period.

Figure 4.6: ML forecasts of the number of HW days in the Lake Como region over the 2011-2022 test period. The solution with the highest validation score from the evolutionary algorithm is used as input for various ML models. HWs are defined relative to the daily 90th-percentiles of the 1981-2010 period.



Figure 4.7: Solutions to optimisation of Lake Come HW occurrence predictors. The data shown corresponds to the top 10% of solutions by validation period skill (corresponding to an average F1-score of 0.49; Table 4.2). The colorbar represents the number of solutions in which the cluster and lead time is used. Maps of the clusters are shown in Appendix 4.

Figure 4.8: ML forecast scores using the optimal solution from two independent runs of the optimisation algorithm: using all clusters (ALL - grey) and excluding the most-frequently picked variables from the ALL run (NOVIP - green).

Figure 4.9: Clusters of daytime and nighttime HWs over Scandinavia and Europe. Using ERA5 over 1950-2022. Rows 1 and 3: Daytime HWs (Tmax). Rows 2 and 4: Nighttime HWs (Tmin).

Figure 4.10: Composites of variables prior to HWs. First row: GPH500 one week before and NAO index around northern Scandinavian heatwaves. Second row: PCs of EOF4 and EOF5 of GPH500 anomalies over the North Atlantic around western European (left) and Baltic (middle) HWs; frequency of southern Scandinavian HWs dependent on ENSO state in March.

Figure 4.11: NAO phase prior to summer HW/WNs. Left: Frequency of nighttime HWs during the summer based on all years (green), years in which NAO in March (mon3) was positive (red) or negative (blue). Nighttime (middle) and daytime (right) HW frequency anomalies after positive (respective lower left triangles in each square; red-blue colour scale) and negative (respective upper right triangles in each square; brown-green colour scale) NAO phase in spring and winter months before HW onset.

Figure 5.1: Occurrence of droughts (left) and mean duration of droughts measured in months (right) in 1993-2018, according to the different statistical drought indices.

Figure 5.2: Spatial distribution of Clusters. The 11 clusters are identified by ascending numbers from 1 to 11.

Figure 5.3: Heatmap: Cluster 10 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).

Figure 5.4: Scatterplot matrix: Cluster 10.

Figure 5.5: FRamework for Index-based Drought Analysis (FRIDA).

Figure 5.6: Selection matrix: Cluster 10.

Figure 5.7: Qualitative improvement obtained via FRIDA: on the left, the initial correlation values between traditional drought indices and FAPAN; on the right, the final correlation values obtained via FRIDA.

Figure 6.1: Obtained agro-climate regions from multivariate clustering approach.

Figure 6.2: Centroids corresponding to the Clusters displayed in Figure 6.1.



Figure 6.3: Obtained projected components from the non-linear RGCCA for the winter wheat series of individual clusters. Orange lines indicate time points at which the associated SATS and NP-SPEI time series were in a positive phase.

Figure 6.4: Spearman correlation coefficients for obtained projected time series of non-linear RGCCA. Only statistically significant values at the 90% confidence level are displayed.

Figure 6.5: Same as figure 6.4, but for SATS.

Figure 6.6: Same as Figure 6.4, but for 500 hPa geopotential height. Contour lines indicate statistical significance at the 90 % confidence level.

Figure 6.7: Same as Figure 6.6, but for relative humidity on the 700 hPa level.

Figure 6.8: Same as Figure 6.6, but for SSTs.

Figure 6.9: Estimated Marginal effects of D-Vine-Copula based quantile regression model for predicting yield variability in cluster three. Values of alpha represent the conditional quantile for which the model was estimated.

Figure 6.10: Retained splitting bounds from RF for the SATS. Light blue lines indicate the PCE and orange lines the MCE. Points are displayed below together with a kernel density estimator for graphical orientation.

Figure 6.11: Same as Figure 6.10, but for NP-SPEI.

Figure 6.12: Number of non-extrapolatable points, when the log-logistic distribution is used as a mapping function for the SPEI.

Figure 6.13: Pearson Correlation coefficient, upper and lower tail dependence of SPEI and NP-SPEI, when the indices are calibrated on the full period. Pearson correlation values are unitless, while tail dependence is expressed in probability. Since both indices have results between 0 and 1, they are plotted with the same scale.

Figure 6.14: Difference of Cramer-von-Mises Statistics for the SPEI and NP-SPEI. Positive values indicate that the Cramer-von-Mises statistic is smaller for the NP-SPEI, suggesting a better mapping to the standard normal distribution.

Figure 6.15: Obtained clusters from the multivariate clustering of heatwaves and droughts based on soft-dynamic time warping for k=10 clusters.

Figure 6.16: Bivariate centroids of cluster six from the soft-dynamic time warping (figure 6.15) approach.

Figure 6.17: Results of the J-Function based interpreter applied for the three regions of interest.

Figure A4.1: Clusters of European predictor variables for the heatwave driver Feature Selection Framework (Section 4). K-means clustering is applied to ERA5 daily data over the



period 1951-2010. The domain covers [30N,70N], [-15E,46E]. Values over the ocean are removed for 2m temperature and Soil Moisture.

Figure A4.2: Clusters of North Atlantic predictor variables for the heatwave driver Feature Selection Framework (Section 4). K-means clustering is applied to ERA5 daily data over the period 1951-2010. The domain covers [0N,70N], [90W,46E].

Figure A4.3: Clusters of Arctic Sea Ice Concentration for the heatwave driver Feature Selection Framework (Section 4). K-means clustering is applied to ERA5 daily data over the period 1951-2010. The domain covers the northern polar region; parts of the domain which have never experienced sea ice are removed from the clustering.

Figure A4.24 Clusters of global predictor variables for the heatwave driver Feature Selection Framework (Section 4). K-means clustering is applied to ERA5 daily data over the period 1951-2010. The domain covers all latitudes and longitudes.

Figure A5.1: Heatmap: Cluster 1 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).

Figure A5.2: Heatmap: Cluster 2 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).

Figure A5.3: Heatmap: Cluster 3 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).

Figure A5.4: Heatmap: Cluster 4 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).

Figure A5.5: Heatmap: Cluster 5 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).

Figure A5.6: Heatmap: Cluster 6 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018)

Figure A5.7: Heatmap: Cluster 7 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).

Figure A5.8: Heatmap: Cluster 8 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).



Figure A5.9: Heatmap: Cluster 9 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).

Figure A5.10: Heatmap: Cluster 11 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).

Figure A5.2.1: Selection matrix: Cluster 1.

Figure A5.2.2: Selection matrix: Cluster 2.

Figure A5.2.3: Selection matrix: Cluster 3.

Figure A5.2.4: Selection matrix: Cluster 4.

Figure A5.2.5: Selection matrix: Cluster 5.

Figure A5.2.6: Selection matrix: Cluster 6.

Figure A5.2.7: Selection matrix: Cluster 7.

Figure A5.2.8: Selection matrix: Cluster 8.

Figure A5.2.9: Selection matrix: Cluster 9.

Figure A5.2.10: Selection matrix: Cluster 11.

Figure A.6.1: Example calibration of SPEI and NP-SPEI for the wichita of the SPEI-demo package. Upper panels show the calculated SPEI based on the log-logistic distribution, while the lower panels display the calculated SPEI based on the nonparametric approach. Red dots in the upper-right panel show values which cannot be mapped.

Figure A.6.2: Same as Figure 6.13, but for calibration on the reference period.

Figure A.6.3: Same as Figure 6.15, but for calibration on the reference period.

Figure A.6.4: Evaluation plot of the deep neural network used. The blue line represents the accuracy of the test set, while the green line represents the accuracy of the validation set.

LIST OF TABLES

Table 2.1: Performance of the selected solution. Each row reports a solution with corresponding equations, and scores results.

Table 3.1: Total and train-test-split number statistics of all TCs and TCs reaching extratropical stage.



Table 4.1: F1 scores of various ML models using the best solution from the evolutionary algorithm run.

Table 4.2: Solutions to experiments on diverse heat extremes in the Lake Como region. Lake Como HW occurrence is depicted in Figure 4.6. NDQ90 refers to the number of days above the 90th percentile (1981-2010). WNs are the equivalent to HWs but with the average apparent temperature at night. S2S removes the first 20 days lag time as possible solutions to the evolutionary algorithm.

Table 5.1: Numerical improvement obtained via FRIDA. 2nd column: best model structure selected; 3rd column: selected predictors for each cluster, 4th column: best initial correlation value between traditional drought indices and FAPAN; 5th column: final correlation value between FRIDA index and FAPAN; 6th column: improvement obtained for each cluster via FRIDA.

Table 6.1: Total and relative contribution of each cluster to the winter wheat yield in France.

Table 6.2: Performances of RF for each cluster. Cluster one to five correspond to the clusters in Figure 6.1 and cluster zero means that the training is done for the full region.

Table 6.3: Test performance of deep neural network.

LIST OF ACRONYMS

WP: Work Package **AI: Artificial Intelligence** AIC: Akaike Information Criterion AMO: Atlantic Multidecadal Oscillation ANN: Artificial Neural Network **ATS: Active Temperature Sum** AUC: Area Under the Curve **BIC: Bayesian Information Criterion** BS: Brier score BSS: Brier skill score CCA: Canonical Correlation Analysis CCEW: Convectively Coupled Equatorial Wave CI: Coupling Index CMCC: Centro Euro-Mediterraneo sui Cambiamenti Climatici CMIP6: Coupled Model Intercomparison Project Phase 6 **CNN: Convolutional Neural Network CSI: Compound Stress Index** CVM: Cramer von Mises D: Deliverable DTW: Dynamic Time Warping ECMWF: European Centre for Medium-Range Weather Forecasts



EN-GPI: Emanuel-Nolan Genesis Potential Index **EE: Extreme Events ELM: Extreme Learning Machine ENSO: El Nino Southern Oscillation EOF: Empirical Orthogonal Function** ERA5: ECMWF Reanalysis v5 **ET: Extratropical Transition** FFNN: Feed-forward Neural Network HYPE: HYdrological Predictions for the Environment HydroGFD: Hydrological Global Forcing Data FAPAN: Fraction of Absorbed Photosynthetically Active radiation aNomaly FAPAR: Fraction of Absorbed Photosynthetically Active Radiation FRIDA: FRamework for Index-based Drought Analysis GA: Grant Agreement GHG: Greenhouse Gas **GPI:** Genesis Potential Index HMD: Heat Magnitude Day HWMI: Heatwave Magnitude Index HW: Heatwaves hPa: Hectopascal IBTrACS: International Best Track Archive for Climate Stewardship IFS: Integrated Forecasting System **IVS: Input Variable Selection** LSTM: Long Short-Term Memory MJO: Madden-Julian Oscillation **ML: Machine Learning** MOEA: Multi-Objective Evolutionary Algorithm MPI: Maximum Potential Intensity MS: Milestone **MSLP: Mean Sea Level Pressure** MCE: Meteorological Compound Event MWMOTE: Majority Weighted Minority Oversampling Technique NAO: North Atlantic Oscillation NOAA: National Oceanic and Atmospheric Administration NP-SPEI: Nonparametric Standardised Evapotranspiration and Precipitation Index NSGA-II: Non-dominated Sorting Genetic Algorithm **OLR: Outgoing Longwave Radiation** PCE: Pure Compound Event PDO: Pacific Decadal Oscillation PMIP4: Paleoclimate Modelling Intercomparison Project phase 4 **PV: Potential Vorticity QBO:** Quasi Biennial Oscillation **ROC:** Receiver operating characteristic **RGCCA: Regularized Generalised Canonical Correlation Analysis**



RF: Random Forest RH: Relative Humidity RWB: Rossby Wave Breaking SATS: Standardised Active Temperature Sum SDTW: Soft-Dynamic Time Warping SIC: Sea Ice Concentration SMHI: Swedish Meteorological and Hydrological Institute SMA: Soil Moisture Anomaly SPEI: Standardised Precipitation and Evapotranspiration Index **SPI: Standardised Precipitation Index** SSI: Standardised Streamflow Index. SST: Sea Surface Temperature **TC: Tropical Cyclone** TC-GPI or TCGI: Tippett Genesis Potential Index Th: Thickness asymmetry W-QEISS: Wrapper for Quasi-Equally Informative Subset Selection WM-GPI: Wang & Murakami Genesis Potential Index



EXECUTIVE SUMMARY

Detection of extreme events is of primary importance in climate science. Identifying the features that contribute to the occurrence of these phenomena can help improve early actions and prompt communication to institutions and stakeholders. This deliverable aims at applying and customising ML algorithms from WP2 that can enhance existing methods, mostly based on observations or dynamical models, or that can work alongside them. On the one hand, these studies corroborate expected relations. On the other hand, these methods 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.

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 footprint of models. The resulting models may improve or compete with existing models, which are often based on dynamical systems.

This deliverable describes the first steps performed in the selection of drivers and machine learning algorithms to detect and predict the four types of climate Extreme Events (EE) considered in CLINT. The extreme events are:

- Tropical cyclones: 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).

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 at their best. Finally, it was possible to highlight the implications of these findings on the physical understanding of the phenomena.

In the case of the detection of tropical cyclones at long time scales, the focus was on the analysis of the capability of the current Genesis Potential Indices (GPI) to indicate the correct number of TC generated in the corresponding model grid cell. This revealed different skills according to the type of GPI and the basin considered. Furthermore, it was found that GPIs had low skill in reproducing interannual variability and trends in future cyclone activity. The



Emanuel-Nolan GPI formula was optimised to perform better in terms of spatial and temporal correlation, obtaining relevant information about the weight of the drivers included in the expression.

In the medium range, operational forecasts of tropical cyclone activity are largely based on dynamical models. The ECMWF ensemble was used as a skilful representative to evaluate the predictive performance of such models in comparison with the prediction obtained from climatological probability. Starting from these two benchmarks, various types of ML models were trained to improve the skill at various lead-times. More complex architectures performed best at the shortest lead times, followed by plain neural networks, and eventually dropped in skill below the climatological prediction. Even though a variety of predictors were tested, considerable improvements were found when including previous predictions or near real-time observations.

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

The analysis of drivers and predictability of heatwaves was extended to other relevant indices of heat extremes, such as Warm Nights. The skill of European warm night forecasts in operational dynamical seasonal forecast systems was assessed and complements the existing knowledge of the dynamical skill of (daytime) heatwaves. Meanwhile, a two-step feature selection and optimization framework to identify drivers and produce forecasts of HW occurrence using a data-driven approach was developed. Despite the initial tests reducing the dimensionality considerably (i.e., using regional-scale area-averages as predictors), the method skilfully recreated the past decade of HW events in Lake Como. This framework identified drivers from local atmospheric conditions (e.g., SST and Z500) to global teleconnections. The results include quantification of the role of S2S drivers, which provides the foundation for building seasonal forecasts from this framework.

The detection of Extreme Droughts has been the subject of extensive research in the past decades, with the development of a wide set of indices (e.g., SPI, SPEI, and SMA). However, these methods failed to 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) was used to construct new composite drought indices. In particular, it helped identify the basin characteristics and extract the most relevant features. This process helped detect the relevant drivers for each cluster and enhanced the impact-based drought indices.





The study of compound events focused on multiple temporally and/or spatially overlapping/connected climate events and their corresponding impacts on the food, water, and energy sectors in Europe. A case study of compound event impacts on winter wheat in France revealed that large fractions of crop variability can be explained by the combined impact of wet and warm conditions in January and February, followed by warm and dry conditions in April. Furthermore, experiments with Random Forests suggested that these types of events can be accurately predicted at a local scale. These methods can be used to construct objective bounds and define thresholds for climate variables associated with substantial agricultural impacts. For the study of interconnected large-scale droughts and heatwaves on a global scale, state-of-the-art non-stationary statistical methods enhanced with Deep Learning allowed the identification of these relationships while considering the climate change signal. Furthermore, a nonparametric version of the well-known standardised precipitation and evapotranspiration index (SPEI) was proposed, with particularly enhanced performance for drought detection and better extrapolation characteristics in reference periods, thereby opening new opportunities to study interconnected heatwaves and droughts.

Further work will be performed to improve the detection systems presented herein. This will be done according to the peculiarities of each extreme event, for example considering different ML routines, testing on different areas or applying detection algorithms to "prediction" mode. These will be the objects of the upcoming Deliverable.



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 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 variables can obtain reliable information from specific datasets (e.g., IBTrACS and FAPAR), which provide more on-point information and can be used as a benchmark for the training of ML 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 considered phenomenon. In addition, the detection and the prediction of events imply the use of different drivers. While the former requires local predictors, the latter needs the inclusion of remote predictors in terms of time and space. 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.



2 TROPICAL CYCLONES

2.1 Overview

Tropical Cyclones (TC) form at a rate of approximately 80-90 times per year globally in the tropical latitude bands on both sides of the equator. TCs that make landfall are among the costliest and deadliest natural disasters due to the combination of strong winds, heavy precipitation, and eventual storm surges. Therefore, it is of paramount importance to accurately predict their activity on several timescales, ranging from a few days to seasonal and climate projections.

To date, a comprehensive theory of TC formation is lacking. Several indicators were developed that relate the spatiotemporal distribution of TC formation to large-scale atmospheric and oceanic variables, such as atmospheric humidity, vorticity, and SST. These indices generally have good skill at the climatological/global scales at which they were trained, but their performance tends to degrade at interannual scales and varies from basin to basin. Another issue with genesis potential indices is that their future trends are often inconsistent with the estimates of future TC activity.

A useful predictor for TC genesis/activity is one that has a reasonable correlation with the target and shows predictive skill. As the predictive skill is different for different predictors, the choice will differ according to the timescales of interest (i.e., a few days ahead vs. the climate change timescale). On sub-seasonal timescales, the Madden-Julian Oscillation is a powerful predictor (Klotzbach, 2014) and possesses predictive skill on a 3-4-week timescale (Vitart, 2009). The open question for ML design is whether to train on the underlying gridded data that form the indices or directly on the indices.

Within Task 3.1, Subtask 3.1.1 focuses on the long climatological scales, using ML algorithms to discover improved formulations of genesis potential indices based on large-scale climate variables. Subtask 3.1.2, on the other hand, focuses on short time scales, developing ML predictors of TC activity for weather prediction and sub-seasonal time scales.

2.2 Datasets and candidate drivers

2.2.1 Datasets

The datasets used as benchmarks included reanalyses (e.g., ERA5 and MERRA) and IBTrACS, a dataset that collects estimates of positions and intensity of TC in the considered basins. For the improvement of genesis indices, drivers were selected from the simulations included in ScenarioMIP and HighResMIP. On the other hand, the ECMWF-ENS was used to improve ML-based forecasts at sub-seasonal time scales. The datasets used are listed in Deliverable 3.1.

2.2.2 Candidate drivers

Subtask T3.1.2 aims to identify the key structures of TC genesis and activity. Such a goal required defining candidate drivers that could refer to local or remote conditions and could



be related to atmospheric or oceanic variables. In particular, local drivers can include variables such as humidity, vorticity, and SST. On the other hand, large-scale field potential drivers consider tropical and 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, Frank and Roundy 2006, Maier-Gerber et al. 2021, Lawton et al. 2022, Schreck et al. 2011), empirical wave-like phenomena such as the Madden-Julian Oscillation (MJO, Klotzbachm, 2004), the African Easterly wave and, even though often rejected (Leroy and Wheeler 2008, Henderson and Maloney 2013), and Quasi-Biennial Oscillation (QBO, Gray 1984). Other effects that have a role in TC genesis can be the Rossby wave breaking (RWB, Zhang et al. 2016, 2017, Wang et al. 2020) or the baroclinic influence of the upper-level trough. The former may be represented by a driver that describes the extratropical influence (for example, upper-level layered-PV), the latter could be considered with 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; possible candidates are local SSTs and teleconnections (e.g., ENSO) (Gray 1979, Gray 1984, Song et al., 2022).

Special attention was given to the transition of TC to extra-tropical regions and the formation of TC outside of the tropics, and field variables such as MJO (whose phase can be determined by checking the EOFs of a set of zonal wind-related variables) were considered.

2.3 Long term horizon

2.3.1 Indices

Subtask 3.1.1 aims to improve the indices developed to detect the genesis of TCs. We considered the Genesis Potential Index of Emanuel-Nolan (2004) (EN-GPI), the Tropical Cyclone Genesis index (TCGI) of Tippett et al. (2011) (TC-GPI), and the Dynamical Genesis Potential Index (Wang & Murakami, 2020) (WM-GPI). EN-GPI is based on large-scale variables, such as absolute vorticity at 850 hPa, relative humidity at 600 hPa, wind shear between 200 and 850 hPa, and MPI (maximum potential intensity, Emanuel (1988), Bister & Emanuel (1998)). TC-GPI replaces MPI with relative SST (Ramsay & Sobel, 2011; Vecchi & Soden, 2007) and WM-GPI inserts additional dynamical variables instead of thermodynamic ones. Menkes et al. (2012) analysed the performance of these indices in reanalyses and found large differences.

2.3.2 Skills and performance of existing indices

The increased spatial resolution of HighResMIP (0.5° or finer in the atmosphere and 0.25° in the ocean) allowed the models to realistically reproduce TCs. Owing to this advance, it was possible to evaluate the performance of GPIs based on the number of generated TCs, using the ERA5 dataset as a benchmark.



The three GPIs were verified in several oceanic basins (North Atlantic NA, North East Pacific NEP, West North Pacific WNP, Northern Indian NI, Southern Indian SI, and West South Pacific WSP), focusing on the representation of interannual variability and multidecadal trends.



Figure 2.1: Left: Boxplots showing ensemble spread of correlation between yearly time series of number of detected TCs and number of TC detected with the three GPIs, calculated on historical simulations. Each box plot is related to a different ocean basin. The star indicates the same correlation, but calculated on ERA5 data. Right: Same as left column but for future projections. The star here represents the ensemble mean of correlations calculated on historical simulations.

Interannual variability was poorly represented by all GPIs in the historical simulations and future projections. In both cases, the indices showed different skills according to the basin considered (see Figure 2.1, left column). In general, a major decrease was found where indices calculated on reanalyses performed better, with a marginally poor performance of WM-GPI. The decrease appears to be more substantial for NA, probably due to overfitting, a consequence of the higher accuracy of historical observations in this area. Future projections result in an additional decrease of GPIs on NA, while for the other basins, they show a further improvement (Figure 2.1, right column).



The second step of the verification of index skill included analyses of how future changes in TC frequency and their multidecadal trends were described by GPIs. This was preceded by a comparison between the GPI projections on HighResMIP and the CMIP6 General Circulation Models ensemble over the 2015-2050 period and considering comparable emission scenarios. These experiments produced similar results, increasing the confidence in the robustness of the results obtained with HighResMIP.

GPIs appear to have low skill in reproducing the interannual variability and trends of future cyclones simulated by HighResMIP. However, the three indices showed similar patterns of change with minor regional differences. Figure 2.2 shows the trend of the predicted number of cyclones for each GPI compared to the simulated trends. This was performed to distinguish between the combinations of models and basins. Here, between half and two-thirds of the model-basin pairs show inconsistency of sign between simulated and GPI trends, with slightly better performance for EN-GPI. These results suggest the need for further research to solve this issue, indicating possible reasons for the poor choice of variables and their tuning of variable coefficients in the GPI formulas. The work presented in this section was published by Cavicchia et al. (2023).



Figure 2.2: Future trends of directly detected TCs and number of TC predicted with GPIs.

2.3.3 Algorithm

To improve the performance of GPIs, a genetic algorithm (NSGA-II, Deb et al., 2002) was used. This type of ML method can efficiently explore a large spectrum of solutions in multiple dimensions and is robust to local minima and maxima. Genetic algorithms are based on several iterations; in each of these, the best performing solutions (according to pre-defined objective functions) of the previous iteration are joined with permutations and mutations of their parameters. Within this joint group, the new best-performing ones are selected, forming a new generation of solutions that are passed-on to the next iteration.

The parameters to be optimised were the coefficients and exponents appearing in the EN-GPI formula (see Table 2.1). A multi-objective optimization was performed by choosing the spatial correlation and interannual correlation between the GPI and the observed cyclones.



Additionally, the algorithm was allowed to change the pressure level at which the variables were calculated.

2.3.4 Results: improved indices and relevant drivers

The best combination of parameters for the improved EN-GPI formula was identified by selecting those that showed a good improvement in spatial and temporal correlations. This evaluation was visualised with Pareto fronts (Figure 3), where the scores of the solutions are displayed on a scatter plot, and it is possible to compare them to the scores of the original EN-GPI. Solutions (b) and (e) show the best balance between temporal and spatial scores. When comparing them to the neighbouring ones, it was possible to identify recurring patterns, such as the use of absolute vorticity at 600 hPa instead of 850 hPa.



Figure 2.3: Pareto fronts obtained optimizing the algorithm on ERA5 (left) and MERRA (right)

Analysing solutions (a) and (d), which are the best performing in terms of temporal correlation, it was found that thermodynamic variables were not influential, due to the presence of low exponents or to the choice of levels where the variables have negligible values (e.g., relative humidity at 1 hPa). However, these solutions tended to perform significantly worse in terms of the representation of spatial patterns (Figure 2.4, top left). Furthermore, the interannual variability of optimised GPI and IBTrACS appeared to be highly correlated, but with low similarity in their peaks (Figure 2.4, top right). This last issue is much less evident for solution (b), evaluated on the two reanalysis datasets (Figure 2.4, bottom panels), indicating the superiority of the solution optimised on both spatial and temporal aspects.

Table 2.1 displays the shape of the solutions highlighted in Figure 2.3, and reports the verification results on different scores and on both reanalysis datasets. It is important to note that the exponents for the vertical wind shear (V_S) are always in the range [-3, -2.4], and those for the absolute vorticity (η) are around 2. Finally, as observed in the bottom panels of Figure 2.4, optimal solutions (b) and (e) performed with relatively high scores on the test reanalysis dataset (i.e., on which they have not been optimised). The work presented in this section is reported in Ascenso et al., 2023.



Table 2.1: Performance of the selected solution. Each row reports a solution with corresponding equations, and scores results.

		Evaluated on		
Solution	Optimised on	Evaluation metric	ERA5	MERRA2
(a)	ERA5	R _{spatial}	0.493	0.500
		R _{seasonal}	0.978	0.978
		R _{temporal}	0.570	0.310
oGPI(a) = (9)	$91.0 + 20.0 V_{S,850-250}$	$)^{-3.0} \frac{RH_{(1)}^{0.4}}{37.4} \frac{MPI^{0.3}}{54.0} \eta_{600} 10^5 ^1$	$^{.3}e^{2.5}$	
(b)	ERA5	R _{spatial}	0.696	0.700
		R _{seasonal}	0.990	0.985
		R _{temporal}	0.430	0.260
oGPI(b) = (9)	$98.9 + 23.5 V_{S,850-250}$	$(-2.9 \frac{RH_{700}^{2.3}}{43.7} \frac{MPl^{1.8}}{68.4} \eta_{600} 10^5 ^2$	$e^{2.0}e^{3.0}$	
(c)	ERA5	R _{spatial}	0.727	0.741
		R _{seasonal}	0.988	0.985
		R _{temporal}	0.219	0.055
oGPI(c) = (9)	$91.2 + 44.3 V_{S,600-250}$	$(10^{-2.9} \frac{RH_{600}^{23.0}}{70.6} \frac{MPI^{2.6}}{71.6} \eta_{500} ^2$	$^{1.1}e^{2.9}$	
(d)	MERRA2	R _{spatial}	0.046	-0.005
		R _{seasonal}	0.483	0.05
		R _{temporal}	0.391	0.617
oGPI(d) = ($14.7 + 99.2 V_{S,250-50}$	$\frac{0.6 \frac{RH_{1000}^{3.0}}{77.2} \frac{MPI^{0.6}}{99.3}}{ \eta_{1000} } \eta_{1000} ^2$	$^{.1}e^{2.1}$	
(e)	MERRA2	R _{spatial}	0.688	0.718
		Rseasonal	0.961	0.966
		R _{temporal}	0.279	0.393
oGPI(e) = (1)	$15.9 + 28.4 V_{S,500-200}$	$)^{-2.5} \frac{RH_{100}^{-1.9}}{75.5} \frac{MPI^{0.6}}{88.6} \eta_{600} 10^5 ^2$	$e^{1.9}e^{1.9}$	
(f)	MERRA2	R _{spatial}	0.713	0.752
		R _{seasonal}	0.989	0.984
		R _{temporal}	0.168	0.165
oGPI(f) = ($20.3 + 18.3 V_{S,500-250}$	$(10^{-2.4} \frac{RH_{600}^{3.0}}{75.8} \frac{MPI^{2.7}}{0.4.8} \eta_{600} 10^5 $	$^{2.1}e^{1.7}$	



Figure 2.4: Top-left: comparison between spatial distribution on IBTrACS and solution (a). Top-right: Interannual variability curves for solution (a). Bottom-left: Same as Top-Right but for solution (b) on ERA5. Bottom-right: Same as Bottom-left but on MERRA2.



2.4 Short term horizon

2.4.1 Target variable

In Subtask 3.1.2, we aimed to improve the TC activity forecasts in the medium range. The target variable was derived from the IBTrACS (Knapp et al., 2010, 2018) dataset version 4 by evaluating, at every grid point on a 2.5°x2.5° grid separately, whether at least one TC occurred within a 48 h time period and radius of 300 km, to be consistent with the definition of TC activity used at ECMWF. The evaluation of occurrence was based on the original 3-hourly temporal resolution and only considered cyclones that reached tropical storm intensity (i.e., \geq 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 was on the Southern Indian Ocean (defined here as 0°-30°S, 20°-90°E), where TCs occur over ocean grid points at a rather low mean relative frequency of 0.56% (Figure 2.5a), but are subject to a distinct annual variability (Figure 2.5b). As shown in many applications, such a ratio in the target variable should still be sufficient to train meaningful models.

2.4.2 Skills and performance of existing forecasts

Since the target variable considered in Section 2.4 is binary, all benchmarks and ML-based models were defined to output probability values, which convey more information than if the models predicted the binary labels directly. However, this implied that the verification of forecasts was more complex. To evaluate the predictive model performance on the test set, different tools were combined to address various verification aspects. ROC curves, which display the true positive rate as a function of the false positive rate, allowed the assessment of the potential predictive ability of a given model, with the best (no) skill indicated by an 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 not statistically consistent with observations. On the other hand, the BS (calculating the quadratic error averaged over all forecasts) carries this aspect in the evaluation by instances. 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).

Before the rise of ML models, TC activity forecasts were either based on dynamical or statistical models, with the former more heavily used on the medium range and the latter mostly on the seasonal range. Since subtask 3.1.2 targets the medium range, ECMWF's ensemble predictions (ENS) serve as the 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 for 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 was calculated in the same manner as for the target variable.



Figure 2.5: (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.

A second type of benchmark was generated from TC activity statistics over the training period from the climatology of the target variable, referred to as 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 2.5a). However, from the variance signal in Figure 2.5b and previous studies (e.g., Maier-Gerber et al., 2021) it seemed advantageous to calculate climatological probabilities separately for each day of the year to reflect seasonal variations. A 30-day window was applied to the day-of-year dimension to smooth out discontinuities resulting from undersampling. Because these statistics were calculated over a set of past realisations drawn from the observational distribution, climatological forecasts were 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 was constant, which made it a good choice as a reference for any skill score, as it allowed easy comparison of the predictive skill of models across lead times.

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

The dynamical forecasts turned out to perform better than the climatological forecasts by more than 40% in BSS at 0-1 days lead time (Figure 2.6a). With increasing lead time, skill decreased continuously and dropped below the climatological reference beyond day 9. The decomposition of the BS revealed that the DSC and MCB terms for CLIM were of the same order as the UNC term, but almost cancelled each other out so that the BS was slightly lower



(i.e., better) than the UNC (Figure 2.6b). In contrast, the ENS model resulted to be well calibrated overall, except over the first two lead days, but it exhibited a good discriminative ability that yielded good BSS values over the first week mentioned before. The generally low UNC term resulted from the high imbalance of the target variables, which means that the trivial approach of always predicting a zero probability (referred to as *ZEROS*) did not perform much worse than the CLIM reference model.



Figure 2.6: (a) Brier skill score (in %) of tropical storm strike 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.

2.4.3 Algorithm

Each of the following ML models was trained for each time-lag of one day from 0 to 13, including in the features the values of the identified candidate drivers lagged by the selected number of days.

2.4.3.1 Baselines

The first set of algorithms considered to perform this task were 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 were 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.

2.4.3.2 Convolutional-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 was designed such that the parameters number did not exceed the order of magnitude of the number of samples.

A U-Net (Ronneberger et al., 2015) was 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 2.7.





These convolutional-based approaches were 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 showed a slight improvement, with the cost of a much larger number of parameters. Therefore, both architectures were considered similarly meaningful.

Further analysis of these neural network architectures will be provided in the future, with the purpose of improving their prediction skills with additional information, such as climate indices, temporal patterns, or more precise features in terms of granularity.

2.4.4 Results: ML-based forecasts and relevant drivers

As shown in Figure 2.8a, the baseline models clearly performed worse than the dynamical model and were found to have the following descending order in BSS: *FFNN* performed best, followed by *extremely randomised trees, logistic regression,* and *AdaBoost*. Note that the BSS of the latter was so low that it was not shown for the sake of readability.





Figure 2.8: Same as in Figure 2.6a, but including (a) the baseline and (b) the CNN-based ML models, respectively.

Figure 2.8b presents the BSS over lead time for the best-performing versions of the CNNbased ML methods, comparing them with the dynamical ensemble benchmark and the FFNN baseline model. While the FFNN model only reached half of the dynamical ensemble model skill on day 0, the U-Net slightly exceeded the dynamical model skill. The benefit of using the advanced methods, however, is only found up to 2 days lead time. These results and others from additional experiments suggested the following descending ranking in BSS for the advanced methods: U-Net, CNN, and LSTM. Thus, the TC strike probability was most accurately predicted by the U-Net approach (brown) for lead times up to 2 days, followed by the FFNN (green) for lead days 3-4.



Figure 2.9: Qualitative summary of lessons learned: (a) the best-performing forecasting approaches for the lead times considered, and (b) evaluation of the impact of predictors when included in the predictor set.

Beyond this, all models trained so far could not outperform the climatological reference model.

More broadly speaking, the best-performing models were CNN-based models in the shortrange (Figure 2.9a), which were able to exploit spatial correlations in the input fields. For larger lead times, plain NNs were more useful as they still allow for modelling nonlinearities. However, these models were not able to identify spatiotemporal patterns. Finally, at very



large lead times the predictive signal resulted too weak, and climatological forecasts performed best.

Additional experiments revealed that using oversampling techniques to combat the imbalance in the target dataset, as well as training on global input fields, led to a strong deterioration in skill, and were therefore not considered for further development. As summarised in Figure 2.9b a neutral effect was found for including the climatological probability, geographical information (latitude and longitude), and temporal information (year and day of year) as additional predictors, as well as for adding predictors from previous days. Another experiment showed that the proxy variable for convection, top net thermal radiation, could be replaced by the total column water vapour. This is an instantaneous (i.e. referred to a specific point in time) parameter and hence simplifies preprocessing in any real-time application. Considerable improvements were achieved by expanding the predictor set using real-time observations (i.e., previous targets) and predictions for the previous day(s). Although operationally preferred, owing to the reduced data volume and lower pre-processing costs, the trial to replace the daily averaged (thus, non instantaneous) predictor data with only the 00-UTC values resulted in a non-desirable reduction in predictive skill and therefore will not be pursued further.

2.5 Summary and outlook

The detection of TCs on long time scales has for a long time suffered from the coarse resolution of dynamical models. Thanks to the introduction of high resolution models, it has become possible to realistically reproduce TCs. This allowed us to thoroughly evaluate the performance of a set of Genesis Potential Indices. It was found that GPIs perform differently according to the basin considered for both historical simulations and future projections. In particular, the decrease in skill (compared to ERA5) on the Northern Atlantic indicated the presence of an overfitting, due to the fact that GPIs were defined taking this basin as reference. Moreover, the indices showed low skill in the reproduction of future trends in all basins, highlighting the need for an improvement of the definition of such indices. This was performed enhancing the Emanuel-Nolan GPI formula with a generative algorithm, which identified better coefficients and exponents to employ and gave indication on the pressure levels to be used for the considered variables. The enhanced GPIS showed improved spatial and temporal correlation with the occurrence of TCs. Finally, the analysis of these performances gave an insight on the role that different variable, such as thermodynamic ones, can have in the optimal formulation of GPIs. Further future work could include including additional variables in the GPIs, or optimise the GPI separately for each ocean basin.

On the shorter range (up to two-week forecasts), ML models were developed for TC activity prediction and compared against predictive skill of dynamical ensemble forecasts and climatological probability predictions. Various model architectures were 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 CNN-based models turned out to perform best for days 0-2, followed by a transition at days 3-4, at which plain NNs



worked better. The reason why a skill comparable to the ECMWF ensemble is only reached at day 0 is partly due to the fact that a coarse resolution of 2.5°x2.5° was deliberately chosen to enable quick iterations during the model development stage. In a next step, we will re-train the best-performing ML models on 1°x1° data so that the CNN-based models can better unfold their potential in identifying spatially correlated structures. Furthermore, new sources of predictability will be included through adding information from climate indices and tropical wave activity, which are both known to modulate TC activity. In addition, a statistical-dynamical model will be implemented to extend the predictive skill, which has so far been limited to only a few days. In this hybrid approach, a statistical model is combined with predictors taken from dynamic model predictions to bring together the advantages of both individual approaches.



3 EXTRATROPICAL TRANSITION OF TROPICAL CYCLONES

3.1 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 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. 2-dimensional fields of TC activity in the northern part of the Atlantic can be predicted based on a set of predictors (either 2-dimensional fields or indices). The solution to this prediction problem is similar to that described in Section 2.4.
- 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 (2), as (1) was discussed previously in this deliverable, and (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 transferable to other ocean basins in which TCs undergo ET (e.g., Northwest Pacific and Southern Indian Ocean).



Figure 3.1: IBTrACS positions cyclones in the extratropical stage for April 2016-December 2022.

3.2 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 were defined as 1980-2015 for training and validation and 2016–October 2021 for the test period.

For the observation dataset for TCs, the alternatives were either based on estimations from the National Hurricanes Centre (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 were 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 sustained wind speed at 10 m height. We defined the target region to be north of 40N and between 98 W-OW, which agreed well with reported ET, as shown in Figure 3.1. 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 was 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 the table below, together with the fraction. As can be seen from these numbers, the proposed train-test split preserved the fraction of ET cases in both subsets.

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

Table 3.1: Total and train-test-split number statistics of all TCs and TCs reaching extratropical stage.

The predictors were 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 beforehand a temporal filter to the indices.

To benchmark the ML-based methods, we used two fundamentally distinct forecasting approaches. First, ECMWF ensemble forecasts (ENS) were used to compare the date-driven ML model with a physical model. Between March 2016 and July 2023, the ENS had a horizontal resolution of 18 km, but underwent several upgrades of the model and data assimilation. Based on automatic tracking (same as for ERA5 above), we examined the forecasts at the genesis time and counted 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 fulfilled the criteria determined the forecast probability. Second, the fraction of TCs undergoing ET in the training dataset (37.7%) could be considered as a climatological forecast (CLIM), assuming the training sample represented the underlying distribution of the target variable and that there are no trends.



3.3 Skills and performance of existing forecasts

Because the target variable is binary and forecasts are probability values, for the ET forecasting problem we used 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 3.2), still close to BS=0.25, which was the value that a random forecasting approach would have resulted without any prior (e.g., climatological) knowledge. This demonstrated 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 exhibited 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 3.2: 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.

3.4 Algorithm

3.4.1 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 scikitlearn package in Python. The only degree of freedom we explored was 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.

3.4.2 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 was applied to the predictor pool to determine a subset that maximises the predictive skill over the training dataset. In each step, the predictor was added from the remaining pool, for which a predefined score was optimised in a 5-fold cross-validation applied to the training data. We


tested 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.

3.5 Results: ML-based forecasts and relevant drivers

The results of the ROC analysis showed that the LOG model was the data-driven approach with the best potential predictive ability, followed by random forests (Figure 3.3a). While CLIM by definition follows the diagonal, indicating no skill, the use of a single decision tree did not perform much better. A similar ranking was also obtained when miscalibration was considered. All ML-based forecasts obtained 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 were better calibrated than the ENS, but were 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) could be seen by the highly reduced miscalibration and enhanced discrimination. However, the best BS among all data-driven models was achieved by the LOG model.



Figure 3.3: (a) ROC curves for all models with AUC scores in the legend. (b) As in Figure 3.2, but including the results for the ML models sorted by BS.

From the statistics of the predictor selection process (Figure 3.4), conclusions could 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 were a sign of overfitting, which the negative log loss is prone to. Using the BIC would have led to decrease the number of selected features so that the optimum would already have been 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 seemed to still lack relevant predictors.



CLINT - CLIMATE INTELLIGENCE Extreme events detection, attribution and adaptation design using machine learning EU H2020 Project Grant #101003876



Figure 3.4: 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.

3.6 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 found it difficult to improve the ECMWF ensemble forecast, despite a positive frequency bias in the ensemble. The strongest influence was 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 found 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 have been 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

4.1 Overview

Temperatures beyond the threshold of human comfort are known to induce excess mortality in a range of climates (Perkins-Kirkpatrick et al. 2015). Meanwhile, temperature extremes can drastically reduce or alter the timing of agricultural production, while abrupt changes in temperature cause similarly abrupt changes in energy demand (Thomas et al. 2020; García-Martínez et al. 2021; Zuo et al., 2015). Extreme temperatures are also precursors to other extreme climate events such as wildfires and droughts (e.g., Lesk et al., 2016). The timing of heat extremes, both in terms of seasons and time of day, plays a key role in the type of impacts. For example, above-normal nighttime temperatures have significant effects on human health, such as attenuated thermoregulation, exhaustion, and physiological effects favouring increased morbidity and premature deaths (Scoccimarro et al. 2017, Kendrovski et al. 2017; Garcia-Herrera et al., 2005). As a result, there are a growing number of indicators for extreme heat events, from heatwaves (HWs, prolonged periods of daily average or maximum temperature above a threshold) to warm nights (WNs, the equivalent for nighttime temperatures).

The detection of extreme temperatures is crucial for 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 variability, and therefore the potential predictability, of daytime and nighttime temperatures differs, with the latter being more sensitive to air humidity and cloud cover (Thomas et al. 2020; Luo et al. 2022). The skill of forecast systems, from short-term to seasonal, in detecting heatwaves has already been tested (e.g. Prodhomme et al., 2022). 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 systems, 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).

In Task 3.2, the aims are to extend the validation of dynamical seasonal forecast skill to a wider range of heatwave indices and to develop ML techniques which provide more reliable forecasts of HW/WNs and improved explainability of their drivers. Three main activities, in line with the task aims, have taken place since the previous deliverable:

- 1. Validation of warm night indicators in dynamical seasonal forecast systems (Section 4.3).
- 2. Development of a ML Feature Selection Framework for detecting HWs (Section 4.4).



3. Comparison of the variability and precursors of heat extremes via spatial clustering (Section 4.5).

4.2 Datasets, candidate drivers and indices

4.2.1 Datasets

ERA5 (Hersbach et al., 2020) is the principal source of data used here for the development of data-driven forecasts, driver detection, and validation of dynamical forecasts of heatwaves and warm nights. While we typically exploit the original resolution of 0.25°, for the spatial clustering of candidate drivers, a resolution of 0.5° was used to reduce the computational expense of predictor clustering in the Feature Selection framework (see Section 4.4, for details and a list of variables used).

The dynamical seasonal forecasting systems used for the comparison to data-driven methods (and for AI enhancement in future deliverables) are Meteo-France System 7, DWD System 2.1, CMCC SPS3.5, and ECMWF SEAS5.1. These systems are the only members of the C3S ensemble to provide 6-hourly fields necessary to calculate warm-nights, namely 2m-dew point temperature, sea level pressure, and t2m temperature.

A 2000-year paleo-simulation (MPI-ESM) has begun to be used in problems that benefit from an increased sample size (e.g., clustering of HWs in Section 4.5) (Jungclaus et al., 2017). This "past2k" simulation was performed with the MPI-ESM1.2-LR model, using ECHAM6.3 as its atmospheric and the MPIOM1.63 as its ocean component. A detailed description of the MPI-ESM model and past2k simulation can be found in Jungclaus et al. (2014). The temperature and variables used to represent the candidate drivers were processed from a spectral T63 grid, i.e. 192 × 96 grid points in longitude and latitude.

4.2.2 Candidate Drivers

In D3.1, a range of candidate drivers was identified from extensive literature to represent a range of timescales and local to regional phenomena that are known or expected to influence heat extremes over Europe. The approach to using candidate drivers in this Deliverable differed according to methodology. In the development of the Feature Selection framework (Section 4.4), a range of regional and global variables were clustered to reduce the dimensionality of the prediction problem and use the framework to identify those that contribute more to predictability. Indices of climate variability, such as NAO, were employed in the analysis of precursors to extreme heat events (Section 4.5).

4.2.3 Indices

A HW is defined as a persistent exceedance of the temperature over a threshold. In all activities reported here, the common definitions of persistence (three days or longer) and statistical threshold (90th percentile) (Barriopedro et al., 2023) are used. The reference period for the 90th percentile may vary depending on the data availability and application. As reported in D3.1, HWs are defined with a daily maximum 2m temperature, while WNs are defined using the average apparent nighttime temperature (between 23:00 and 06:00). The



apparent temperature is a function of relative humidity and therefore accounts for human discomfort.

The following definitions are applied in this Deliverable, and are applied equally to HWs and WNs:

- Daily Indicators:
 - HW occurrence. Truth value of whether there is a heatwave (1) or not (0).
 - HW intensity. The temperature anomaly of HW days relative to the 90th percentile.
- Seasonal Indicators:
 - Number of days above the 90th percentile in the target season (NDQ90).
 - Number of HW days in the target season.
 - HW Magnitude Index (HWMI): a measure of the strongest heatwave in a given season in terms of both duration and intensity (Russo et al., 2015).

4.3 Skills and performance of existing indices and forecasts

4.3.1 Heatwaves

Prodhomme et al. (2022) showed that for Europe C3S systems generally outperform statistical models based on climatology or warming trends, although there is a large amount of heterogeneity in the forecast reliability. Here, the Lake Como region was used as a case study for the impacts of extreme events on water supply and crop production. In the Lake Como area HW forecasts are only marginally better than climatological forecasts (Prodhomme et al., 2022). The most skilful member of the CMCC-35 ensemble was able to capture forecasts of summer HW days (initialised in May) with a correlation score of 0.68. However, the ensemble spread was large (Figure 4.1), and as a result, the correlation score of the median was insignificant (0.18). Even the long-lasting HWs in 2003 and 2015 were severely underestimated by the dynamical system median. Disagreement in the ensemble spread inhibits the use of this system by climate services and makes the Lake Como region a suitable target for ML-enhancing of HW forecast.

The measurement of seasonal forecast skill of extreme event detection is typically performed on seasonal indicators, such as the number of HW days, since these systems are not expected to provide accurate information on daily timescales. Skill of daily indicators is commonly tested with the F1-score, a measure of binary recall and precision of true positives (values range from 0 to 1, with 1 indicating perfect representation of the target data). To highlight this point, the CMCC-35 ensemble member with the highest seasonal skill score over Lake Como was found to have a very low F1-score for HW occurrence (0.12, Figure 4.2). Dynamical systems are generally unable to accurately capture the start and end times of HWs (Figure 4.2). Instead, their value manifests in their ability to capture propensity (Figure 4.1).



CLINT - CLIMATE INTELLIGENCE Extreme events detection, attribution and adaptation design using machine learning EU H2020 Project Grant #101003876



Figure 4.1: Seasonal forecasts of the number of days of summer HWs in the Lake Como region over the 1993-2016 period. Values shown here are the annual sums of HW occurrence shown in Figure 1. The CMCC-35 ensemble spread for each year is represented by the box plots, and the median denoted (orange line).



Figure 4.2: Seasonal forecasts of daily HW occurrence in the Lake Como region over the 1993-2016 period. Forecasts are taken from the CMCC-35 dynamical system for the summer period initialised on May 1st CMCC-35. The forecast shown (red) is the ensemble member with the highest correlation score of number of summer HW days (see Figure 4.1). ERA5 (black) data is used as the benchmark for validation. HWs are defined relative to the daily 90th-percentiles of the 1993-2016 period.

4.3.2 Warm Nights

Prodhomme et al. (2022) validated the dynamical seasonal forecast skill of seasonal HW indicators over Europe. Within CLINT, a similar study was performed on extreme nighttime extreme heat indicators (Torralba et al., in review). Here, we present a summary of the results for WNs. In addition, the study also compared HWs defined with only minimum daily temperature and average nighttime temperature.



CLINT - CLIMATE INTELLIGENCE Extreme events detection, attribution and adaptation design using machine learning EU H2020 Project Grant #101003876



Figure 4.3: Seasonal Forecast skill of warm nights over Europe. a) Domains: Northern Europe (NE), Western Europe (WE), Central Europe (CE), Eastern Europe (EE), Mediterranean (MED), Northern Africa (NAF), Middle East (ME), full European domain (ALL) b) Ensemble mean correlation values for the seasonal predictions of the b) HWMI and c) NDQ90 based on the ATn. Results for the seasonal predictions from individual C3S seasonal prediction systems (CMCC-35, DWD-21, ECMWF-SEAS5.1, MF-7) and their multi-model combination (MM) are shown for each region. The seasonal forecasts were issued on the 1st of May and the observational reference is ERA5. This assessment corresponds to the 15MJJA season in the 1993-2016 period. Significant correlations at the 95% confidence level are marked with an asterisk. From Torralba et al. (in review).

Similar to previous work on daytime HWs, we divided the European domain into seven regions and assessed the area-average WN correlation skill scores for a range of seasonal forecast systems (Figure 4.3). NDQ90 seasonal forecasts (Figure 4.3b) over the aggregated points in the Euro-Mediterranean domain were almost consistently better than the HWMI seasonal forecasts (Figure 4.3a) for the single regions, highlighting the decrease in predictability when applying the condition of temperature persistence to the HW definition. The predictions of HWMI showed potential in the Mediterranean region, where CMCC-35 and MM showed positive and significant correlations (Figure 4.3b). Similar results were obtained in Eastern Europe, where ECMWF-5 and MF-7 showed positive and significant correlations, and in the Middle East, where all systems considered showed the highest correlation values. As discussed in the previous section, limited skill was found in seasonal temperatures in Northern and Western Europe. The positive and significant correlations obtained for both HWMI and NDQ90 in several regions indicated the potential of seasonal predictions to provide useful information on WNs that can be used for the decision-making processes in different socio-economic applications. However, the potential application is regiondependent, and the MM shows the majority of regions display insignificant HWMI skill.

The seasonal prediction skill of HWMI and NDQ90 in specific regions such as Southern and Eastern Europe or the Middle East showed that these indices can be integrated into specific



climate services intended to reduce the impact of NHWs in vulnerable sectors such as public health or agriculture. Skill score patterns for WNs (defined with apparent temperature) across Europe were found to be greater than those of HWs and nighttime HWs (defined with standard temperature) only in specific locations across Europe (e.g. the Mediterranean coastal zone; Torralba et al; in review), indicating where variability in relative humidity provided greater sources of predictability. A more complete study of differences in predictability is the focus of Section 4.5.2.

4.4 Algorithm: feature selection framework

The data-driven forecasting framework is composed of two steps: dimensionality reduction of global variables, followed by feature selection to t identify the optimal combination of drivers. In the first step, the global clustering of specific variables (e.g., sea ice concentration and soil moisture) served to reduce the dimensionality of the problem and allowed the potential identification of local and regional dynamics crucial to the occurrence of European heat extremes (Figure 4.4). In the second step, a wrapper combining Extreme Learning Machines and the Coral Reef Optimization (CRO) algorithm (Salcedo-Sanz et al., 2014; 2017, Pérez-Aracil, J., 2023) was used to select the clusters of the different variables (features) that provided the optimal detection skill (F1-score) for the target time series: heatwave/warm night occurrence at Lake Como. The optimization is based on three targets: variable cluster, lead-time, and sequence length. The framework allowed the quantification of the relative importance of each variable and cluster, and crucially, to identify the time lag from short-term to seasonal time scales (up to 180 days). In Milestone 22, we provided a description of the general set-up. In Deliverable 2.4 (due Q1 2024) we will provide a fuller description of the setup, including a sensitivity analysis of hyperparameters. Here, we focus on the initial efforts to adapt the framework to create forecasts and provide an indication of the capabilities of the framework in detecting HWs.

The following variables were used as candidate drivers. MSLP and z500 represented atmospheric circulation, which can serve as both short-term predictors of surface atmosphere conditions or indications of teleconnections on longer spatial and temporal time scales, such as Rossby waves. SST represented the slow varying influence of ocean-atmospheric fluxes and their impacts on circulation, such as North Atlantic SSTs (Cassou et al. 2005; Duchez et al. 2016), ENSO (Zhu et al. 2015; Wulff et al. 2017), AMO (Della-Marta et al. 2007), and PDO (Kenyon and Hegerl 2008). SICreduction in late winter, which is known to influence springtime temperatures over Europe through changes in atmospheric heat flux and circulation (Zhang et al., 2020), was also included. Same was performed for precipitation (Stefanon et al., 2012) and soil moisture (Prodhomme et al., 2016), that are also precursors of European heatwaves. Outgoing Longwave Radiation was used as a proxy for blocking, MJO, and cloud cover (Rodrigues et al., 2022). In the case of T2M, ocean values were masked to avoid repeating the information provided by the SST clusters. The identification of relevant drivers using the framework is discussed in Section 4.5.

In the reported set-up, we used an arbitrary but low number of clusters (i.e., five per region/variable, Appendix) to perform preliminary tests. The time series of the average values



of the variables in these clusters served as the predictor data. The target data were Lake Como HW occurrence, from ERA5, over the period 1950-2010, using the 90th percentile over the 1981-2010 period as a threshold (see black squares in Figure 4.5). The CRO algorithm was run 15,000 times in 10 independent runs to provide a total of 150,000 solutions (see Milestone 11 for how validation scores increase with the number of runs). The concatenation of independent runs is important, given the inherent randomness of the initial solutions used in the evolutionary algorithm. A logistic regression model was used to train each solution, and the output was used to define the F1-score. Given that the algorithm tests lag times from 0 to 180 days, the forecasting set-up was effectively to "nowcast" i.e. use predictor data from the short-term. Thus, the following results cannot be fairly compared to the dynamical systems described in Section 4.3.





The best solution for Lake Como HW occurrence among the 150,000 solutions during the validation period had an F1-score of 0.54. The corresponding test period F1-score for the same solution was 0.36. The difference in the validation and test period scores is considered a measure of model overfitting to the training data. Although the optimization algorithm was trained to optimise with a Logistic Regression model, the solutions could be used to train other, more sophisticated, ML, or DL models, with the aim of finding more accurate forecasts (Table 4.1). Taking the best solution and using it to train the Gradient Booster Classifier (GBC) model increased the test period F1-score from 0.36 0.6. This particular model reproduced some of the longer HW episodes, such as those in 2015, but showed a tendency to overestimate some others (e.g., in 2022; Figure 4.5). Seasonal indicators could be constructed from the forecast HW occurrence (Figure 4.6). GBC had the highest F1-score and displayed one of the highest correlation scores. SVC had the highest correlation but drastically



underestimated the number of HW days each year. This was not optimal, since a reliable ML model should display high skill for both daily and seasonal indicators. Furthermore, as shown in Table 4.1, the relationship between skills for daily and seasonal indicators was not linear.

In the implemented preliminary set-up, the F1-scores shown here resembled those of the seasonal forecasts for the 1993-2016 period. The scores for the CMCC-35 ensemble median and the worst performing ML model (SVC) were 0.18 and 0.10 respectively, while the scores for the best-performing CMCC-35 ensemble member and the GBC model were 0.678 and 0.60, respectively. Two important differences were found, in addition to the time period of the study. First, the ML models were in a nowcasting mode, benefitting from driver information on short timescales. Dynamical systems, on the other hand, have initial conditions for the 1st May and are provided no extra observation-based information during the summer period. Ideally, ML models should have had a considerably higher F1 score, reflecting near-perfect reconstruction of the HW record. However, the predictor data were based on averages over regional-scale areas. Thus, the (spatial) dimensionality reduction was considerably larger than the grid cell size of the seasonal forecasts (1°). It is therefore encouraging that the ML models could recreate HW occurrence with F1-scores of up to 0.6.

Method	F1-score (daily)	Correlation (Seasonal)
Logistic Regression	0.36	-0.11
SVC	0.10	0.51
Decision Tree Classifier	0.44	0.30
Random Forest Classifier	0.12	0.68
K Neighbours Classifier	0.11	0.32
AdaBoost Classifier	0.49	0.08
MLP Classifier	0.2	0.26
Gradient Boosting Classifier	0.6	0.68

Table 4.1: F1 scores of various ML models using the best solution from the evolutionary algorithm run.



CLINT - CLIMATE INTELLIGENCE Extreme events detection, attribution and adaptation design using machine learning EU H2020 Project Grant #101003876



Figure 4.5: Gradient Booster classifier forecast of daily HW occurrence in the Lake Como region over the 2011-2022 test period. The forecast shown (red) corresponds to the solution with the highest validation score from the evolutionary algorithm. ERA5 (black) data was used as the benchmark for validation. HWs were defined relative to the daily 90th-percentiles of the 1981-2010 period.







Using identical hyperparameters, set-up, and number of evaluations, equivalent forecasts were made for different extreme heat indicators for Lake Como to understand the capabilities and limitations of this method. Different indicators presented different levels of data imbalance, defined as the proportion of extreme event days in the target data. For example, using NDQ90 instead of HW occurrence increased the number of extreme events by approximately 40%, because this does not filter heat extremes by duration. In terms of data imbalance, the percentage of extreme event days in the dataset changed from 8.3% to 11.7%. Although the validation score was equal for both experiments, the test score increased to 0.48 for NDQ90; similar increases were seen in the average of the top 10% of solutions. Overfitting seemed to have been reduced and test scores increased, implying that NDQ90 detection is an easier problem to solve. The difference in skill quantified the loss of predictability when considering persistent heat extremes.

There are fewer WN on records in Lake Como compared to HWs (Table 4.2), and this imbalance appeared to impact ML detection capabilities. Although the best validation scores were higher than those for HWs, the corresponding test scores were lower, indicating a higher degree of overfitting. Even the GBC method, which considerably boosted scores for HWs, performed similarly to the Logistic Regression. It should be noted that the lower number of WN events might not be the only reason for the reduction in skill; this could indicate that the predictor data used were insufficient to detect apparent temperature extremes. Temperature was used as a predictor and, as shown in Section 4.5, was unsurprisingly a key predictor for HWs in the short term. We did not include relative humidity as a predictor, which is a variable that may similarly impact WNs in the short term (especially given that WNs are defined with RH).

Table 4.2: Solutions to experiments on diverse heat extremes in the Lake Como region. Lake Como HW occurrence is depicted in Figure 4.6. NDQ90 refers to the number of days above the 90th percentile (1981-2010). WNs are the equivalent to HWs but with the average apparent temperature at night. S2S removes the first 20 days lag time as possible solutions to the evolutionary algorithm.

	Total HW occurrence (total days = 7380)	Best F1-score (CV/Test)	Average F1-score of top 10% (CV/Test)	Highest F1-score (Method)
Lake Como HWs	612	0.54/0.36	0.49/0.39	0.6 (GBC)
Lake Como NDQ90	861	0.54/0.48	0.52/0.48	0.64 (GBC)
Lake Como WNs	517	0.60/0.32	0.53/0.33	0.37 (GBC)
Lake Como S2S	612	0.3/0.24	0.25/0.21	0.28 (DTC)

A crucial aim of this Task was to develop a seasonal forecast mode for this framework as opposed to the nowcasting setup. As explored in Section 4.5, the strongest sources of ML



model predictability come from variable clusters that impact HWs in the short term, as opposed to those on sub-seasonal to seasonal (S2S) timescales. We re-ran the optimization algorithm with the Lake Como HW data but forced it to ignore lag times of less than 20 days to create a preliminary seasonal-style forecast. This set-up was not yet comparable to dynamical systems initialised in May, as lag times of 20 days or longer for HWs later in the summer (i.e., August) still correspond to the summer period. In both the Logistic Regression and highest-performing models, the F1-score of the validation and tests was halved. The fact that a large portion of the skill comes from short lag times was not surprising, but it is also encouraging that half of the skill is derived from S2S drivers.

4.5 Results: relevant drivers

Two approaches for identifying extreme heat drivers were explored. The first was the continuation of the feature selection framework used to perform HW detection on the lake Como case study. Here, the clusters chosen to make optimal detection are reported. The second was a more "traditional" analysis of composites of atmospheric conditions, a spatial clustering of HW patterns was used to define regional drivers.

4.5.1 Feature Selection Framework: Results for Lake Como

As described in the previous section, the feature selection framework optimised the ML forecast F1-scores based on three dimensions of the solution-variable cluster, lead time, and sequence length. These dimensions were interpreted as potential drivers of summertime Lake Como HWs. In early tests (Milestone 22), an optimal solution (from a separate run of the optimization algorithm) was found to choose local and short-term drivers of HWs, such as T2m and Z500, in the clusters that contained the Lake Como region. Here, using a wider sample of data (the top 10% of the 150,000 evolutions), optimised solutions for Lake Como HW forecasts (corresponding to an average F1-score of 0.49) consistently chose the nearby T2m and Z500 clusters, as well as OLR from the Eastern Pacific and MSLP over the subpolar North Atlantic (Figure 4.7). The local dependence on temperature and atmospheric circulation is intuitive, while the OLR and MSLP clusters, given their locations, most likely correspond to representations of ENSO and NAO, respectively. However, these interpretations still need to be corroborated. Applying a further step of filtering, by taking the solutions that appear in more than 90% of the optimal solutions shown in Figure 4.7, we found that only these four solutions appear. The implications were that the majority of skill came from these clusters (hereon denoted as "VIP" clusters) and that the other potential predictors were essentially a source of noise.

Prior studies, however, have shown that extreme temperature predictability is not limited to the three VIP clusters (e.g., Stefanon et al., 2012). To explore this seemingly contradictory result, we repeated the experiment with the VIP clusters removed (NOVIP). The optimal solution for logistic regression was found to have only a slightly reduced F1-score, indicating that reconstruction of the target data was nearly as skillful without the VIP clusters (ALL vs. NOVIP; Figure 4.8). Moreover, ML methods using the optimal solutions of the NOVIP experiments were able to match or even outperform the ALL experiments (see GNB in Figure



4.8, for example). The clusters that appeared most in the NOVIP optimal solutions included local precipitation (tpEurope cluster3) and European soil moisture (sm1Europe cluster1 and sm Europe cluster2). Both could be linked to summer HWs in Europe (Stefanon et al., 2012, Prodhomme et al., 2016). We drew three conclusions from the NOVIP-ALL comparison. Firstly, the ability of clusters to provide predictive skill depends greatly on the models used. By using AdaBoost, for example, the NOVIP experiment was able to provide more skillful reconstruction than the ALL experiments. Secondly, NOVIP experiments showed that predictors with longer lag times can permit a similar quality of HW detection compared to short-term predictors; the peak skill from precipitation comes from 22 days prior to HW (in NOVIP), compared to the < 10 days for local temperature in ALL. This is a potentially promising result for the development of S2S forecasting applications, which cannot make use of shortterm information. Lastly, the equivalent skills between NOVIP and ALL imply that equivalent information was provided by the least discarded clusters in both experiments. For example, local temperature (the least discarded in ALL) and precipitation (the last discarded in NOVIP) may behave in a highly correlated way prior to HWs (e.g., Stefanon et al., 2012). Thus, the optimization algorithm must choose between clusters which effectively represent the same driving process. Further tests are necessary to interpret the algorithm's capability to differentiate between similar clusters.

Figure 4.7: Solutions to optimisation of Lake Come HW occurrence predictors. The data shown corresponds to the top 10% of solutions by validation period skill (corresponding to an average F1-score of 0.49; Table 4.2). The colorbar represents the number of solutions in which the cluster and lead time is used. Maps of the clusters are shown in Appendix 4.





CLINT - CLIMATE INTELLIGENCE Extreme events detection, attribution and adaptation design using machine learning EU H2020 Project Grant #101003876



Figure 4.8: ML forecast scores using the optimal solution from two independent runs of the optimisation algorithm: using all clusters (ALL - grey) and excluding the most-frequently picked variables from the ALL run (NOVIP - green).

4.5.2 Spatial Clustering of Heat Extremes: common drivers on regional-scales

HWs at different locations are expected to be influenced by different drivers; however, certain combinations of driver states may lead to HWs in a larger area. To reduce the task of identifying drivers for a potentially infinite number of data points to a reasonably generalised one, an approach was chosen to find the dominant, recurrent, spatially coherent HW patterns over Europe. Drivers can then be identified for each of these larger-scale HW patterns, thus generalising the results obtained for local case studies such as Lake Como. To detect the dominant HW patterns over Europe, a clustering method was applied to cluster warm days, that is, exceedances of the 90th percentile of daily maximum 2m-temperatures, in ERA5 (1950-2022) and the MPI-ESM past2k simulation. As in Prodhomme et al. (2022), a procedure by Mahlstein et al. (2015) was followed, where a polynomial regression was applied to smoothen the 90th percentiles computed for each day in summer (MJJA).

While the clustering used in Section 4 was designed for feature selection and, hence, to identify potential local dynamical HW drivers, the clustering presented here was to reduce the dimensionality of the HWs themselves, in order to identify drivers at a regional scale. The clustering method of choice was the Simulated Annealing and Diversified Randomization (SANDRA) scheme (Philipp et al., 2007). SANDRA is based on conventional k-means clustering but implements two additional concepts: the process of simulated annealing allows fields to be temporarily assigned to certain clusters, even though this assignment might result in a temporary decrease in the overall data partitioning quality. In SANDRA, simulated annealing is repeated a large number of times under the concept of diversified randomization, which means that the starting clusters as well as the ordering of input fields were randomised throughout the iterative process of checking and reassigning. Therefore, SANDRA can



overcome many of the limitations of other methods, leading to clusters closer to the global optimum. SANDRA was first developed to classify large-scale atmospheric circulation fields (represented by sea-level pressure or geopotential height maps), and, to our knowledge, its CLINT application in the context of HWs is novel.

Sensitivity tests were performed to determine an ideal number of clusters with both high intra-cluster similarity and low inter-cluster similarity. Finally, five clusters computed separately for the northern and central/southern European domain fulfilled these conditions best (Figure 4.9). The northern domain resulted to comprehend HWs over northern, southern, and whole Scandinavia as well as the Baltic regions, respectively. On the other hand, the southern domain showed HWs clusters centred over western Europe (France/Spain), central Europe, northeastern central and southeastern Europe. In both domains, more than 15% of all summer days (MJJA) could be assigned to one of these clusters, whereas the remaining days were classified in the no-heatwave cluster, respectively. For the following analysis, based on heatwave clusters, a heatwave is defined as a period of at least three days being assigned to the same HW cluster.



Figure 4.9: Clusters of daytime and nighttime HWs over Scandinavia and Europe. Using ERA5 over 1950-2022. Rows 1 and 3: Daytime HWs (Tmax). Rows 2 and 4: Nighttime HWs (Tmin).

The patterns of HW clusters obtained from the past2k simulation were generally very similar to those computed from ERA5, although the order of the clusters (determined by their occurrence frequencies; the most frequent cluster is ranked first) varied slightly. The sensitivity to the climatology period used to compute the 90th percentile of daily maximum temperatures was tested in ERA5. In the southern domain, the effect of the underlying climatology resulted to be very small, and the clusters computed from anomalies with respect to percentiles based on either 1950-2022 or 1981-2010 were found to be very similar in terms of spatial pattern, maximum intensity, and the number of days assigned to them. In the northern domain, the order of the first two HW clusters was affected by underlying climatology. The northern domain clusters computed from the 2000-year past2k simulation showed the same order as that in ERA5. In the southern domain, the two eastern European clusters were centred slightly off the ones in ERA5, while the central European cluster occurred more frequently than in ERA5.



To identify potential physical drivers for the individual HW clusters, a classical composite analysis was performed, where a large set of different variables and derived indices were tested for their connection to the occurrence of the HWs. Maps, as shown in Figure 4.10 (upper left), for geopotential height at 500hPa (GPH500) averaged over the one-week period before the onset of all northern Scandinavian HWs were produced for all HW clusters using mean sea level pressure, geopotential height at different pressure levels, sea surface temperatures, sea ice, and soil moisture over the North Atlantic/Europe area. Time lags of up to several months before the onset of HWs were analysed.



Figure 4.10: Composites of variables prior to HWs. First row: GPH500 one week before and NAO index around northern Scandinavian heatwaves. Second row: PCs of EOF4 and EOF5 of GPH500 anomalies over the North Atlantic around western European (left) and Baltic (middle) HWs; frequency of southern Scandinavian HWs dependent of ENSO state in March.

From the example shown in Figure 4.10 (upper left), the well-known connection between a high-pressure system and an HW occurrence in the same region was confirmed. Furthermore, a connection between HWs over Scandinavia and the North Atlantic Oscillation (NAO) was indicated by the negative GPH500 anomalies close to Iceland, together with the positive anomalies south of them. This connection could also be seen when looking at the NAO index prior to HW onset (Figure 4.5, upper right), which showed a clear upward trend from around two weeks before the onset of Scandinavian HWs onwards. This upward trend appeared even earlier (~3.5 weeks) before HWs over southern Scandinavia (cluster 3 in the northern domain) and around the same time before HWs over western Europe (cluster 3 in the southern domain).

The NAO index was computed as the principal component (PC) of the first empirical orthogonal function (EOF) of the North Atlantic GPH500 anomalies. PCs with higher EOFs, which represent variabilities such as those stemming from the East Atlantic pattern, were also



tested, and some of them indicated a potential connection to individual HW patterns, such as PC4 for Western European HWs and PC5 for Baltic HWs (see Figure 4.10, second row).

Composite maps and time series, such as those shown in Figure 4.10, upper left, were computed for both ERA5 and the past2k simulation. While there are strong similarities between the datasets, for example the importance of the NAO for HWs in specific regions, other variables differ in their signals prior to the HW onset. These differences, which include different SST signals in the North Atlantic, are yet to be understood. Differences are also seen in a potential connection between the El Niño Southern Oscillation (ENSO) in spring and the occurrence of certain HW clusters, which is suggested only from the long model simulation, but not from the shorter reanalysis dataset (see Figure 4.10, lower right, as an example). By understanding the reasons for these differences, deeper insight into the underlying physical mechanisms in the development of HWs at different locations can be obtained.

We also began to analyse potential drivers not only of daytime HWs, but also to conduct a systematic comparison of warm nights and nighttime HWs (Section 4.2). The differences between extremes in the daily temperature maxima (HWs) and minima were explored. Cluster patterns of daytime and nighttime HWs were generally very similar in the northern domain but, as expected, were less intense during the night (Figure 4.9). The same was found to be true for nighttime HW intensity in the southern domain, where not all daytime HW patterns had equivalents at nighttime. Similarities and differences could be inferred for the role of the NAO in daytime and nighttime HWs (Figure 4.11). More nighttime HWs occurred in summers after positive NAO phases in March. This effect was observed throughout summer and resulted to be similarly consistent after positive NAO phases in February for nighttime HWs, but slightly weaker and especially not as strong in February for daytime HWs.



Figure 4.11: NAO phase prior to summer HW/WNs. Left: Frequency of nighttime HWs during the summer based on all years (green), years in which NAO in March (mon3) was positive (red) or negative (blue). Nighttime (middle) and daytime (right) HW frequency anomalies after positive (respective lower left triangles in each square; red-blue color scale) and negative (respective upper right triangles in each square; brown-green color scale) NAO phase in spring and winter months before HW onset.

4.6 Summary and outlook

Here we report demonstrate progress on the main three aims of Task 3.2. Firstly, validation of warm night indicators in dynamical seasonal forecast systems was performed in a



consistent way to equivalent works on (daytime) HWs (Section 4.3.2) Prodhomme et al., 2022). Secondly, a ML feature selection framework (Section 4.4) was developed to detect drivers of extremes and reconstruct the daily HW record. Lastly, we performed a comparison of the variability and precursors of heat extremes via spatial clustering of the events (Section 4.5). The work on dynamical system validation provided an indication as to where forecast enhancement efforts, using the feature selection framework, should be made. Meanwhile, the physical explainability of both dynamic and ML-based predictability was explored with the spatial clustering of the events.

Following the validation of dynamical seasonal forecasts of both daytime (Prodhomme et al., 2022) and nighttime (Torralba et al., in review) temperature extremes for C3S systems, we obtained enough information regarding target identification for data-driven forecasts or ML enhancements to dynamical systems. The large ensemble spread in the Lake Como region, typical of the heterogeneity of dynamical forecast reliability across Europe, is representative of the current obstacles to the mid-latitude predictability of heat extremes.

The two-step feature selection framework presented here provided forecasts of heat extremes and a level of explainability (relevant clusters and timings). Preliminary results showed the capacity to recreate HW records with reasonable skill, considering the dimensionality reduction. Future work will explore changes in skill with dimensionality (i.e., the number of clusters used). In principle, this framework can be extended to other target data and extreme events. Other means to reduce dimensionality include target and predefined candidate drivers (e.g. Zhang et al., 2022). The optimization algorithm can select from a range of candidate drivers, thereby uncovering new drivers. Manually defined drivers of HWs will need to change depending on the season. To adapt to other target variables or times, no adjustment to the method described here needs to be made. However, much continued development of this method is underway: refining the number of clusters to explore the effect of dimensionality reduction, applicability to different heat extremes (e.g., WNs), representation of trends linked to global warming, representation of teleconnections, and propagating wave patterns (e.g. complex EOFs to represent Rossby waves; Majumder et al., 2019).

A crucial application of this method will be the creation of data-driven seasonal forecasts. The results shown here mostly correspond to a nowcasting mode of the framework, in the sense that predictor information from the summer is not omitted (unlike in a dynamical forecast initialised in May). Thus, we cannot yet provide a comparison to dynamical system skill. While we demonstrated that removing the short-term (<20 days) information from the algorithm drastically reduces skill, there is still room for improvement over dynamical systems in regions with very poor dynamical skills, such as Lake Como. We will explore changes in the set-up to provide a seasonal forecast version of the framework. First, a data-driven approach that learns from predictors prior to April can be compared with dynamical forecasts initialised in May. Second, the predictor output of dynamical systems will itself be used to make enhanced dynamical forecasts, with the aim of reducing the ensemble spread or choosing the optimal ensemble members. Moreover, advances in the parallelization of the framework will speed



up the optimization algorithm (to be described in Deliverable 2.4) and make a pan-European application of the method more feasible.

The exploration of HW and WN drivers using more "traditional" methods, such as composite analysis, can provide us with an indication of which drivers/indicators the data-driven approach could benefit from. However, the suggested impacts found so far, such as the potentially greater impact of NAO on WNs compared to HWs, will require further investigation.



5 EXTREME DROUGHTS

5.1 Overview

Drought is one of the costliest natural hazards, causing extensive damage and affecting a significant number of people (Wilhite, 2000). For this reason, and because of the everincreasing severity and frequency of this phenomenon in recent decades (Chiang et al., 2021), the interest in drought monitoring, prediction and risk analysis is growing (AghaKouchak et al., 2023). The first crucial step in reducing the damage caused by droughts is to detect them.

Drought detection is based on the analysis of a series of drought indices that represent the conditions of different components of the hydrological cycle (e.g., precipitation, soil moisture, and river flow) that are associated with a particular type of drought. Drought indices are quantitative measures that characterise drought in terms of intensity, onset, termination, duration, and severity by assimilating data from one or several variables into a single numerical value. These indices generally represent statistical anomalies of the current situation with respect to the long-term climatology at a given location and period and thus provide a measure of the probabilistic severity of a given event.

5.2 Datasets and indices

The computation of drought indices requires a time series of different hydroclimatic variables over the Pan-European domain. The European Hydrological Predictions for the Environment model (E-HYPE; Hundecha et al 2016), a semi-distributed hydrological model, combined with HydroGFD2.0 (Berg et al 2018) reanalysis data represents a suitable dataset for this analysis, as it provides the following data:

- Precipitation and temperature obtained from HydroGFD2.0 reanalysis.
- Simulations of evapotranspiration, streamflow, and soil moisture were produced as outputs of the E-HYPE model forced with HydroGDF2.0.

HydroGFD is a merged dataset of historical precipitation and temperature from meteorological reanalysis and global observations. The reanalysis system from ECMWF (European Centre for Medium Range Weather Forecasts) uses atmospheric and surface observations to reproduce the observed weather and climate as closely as possible on a global scale. However, the reanalysis product has systematic errors (biases) that prevent its direct use in hydrological models. In fact, reanalysis data are known to contain biases due to errors in the underlying weather forecast model. For this reason, there is the need for bias correction, in order to bring simulated capacity factors in line with reality. The bias-adjustment for ECMWF seasonal forecasts was performed using a modified version of the distribution based scaling (DBS) method (Yang et al 2010) to HydroGFD as the reference.

The E-HYPE model (the HYPE model version for European basins) developed by the Swedish Meteorological and Hydrological Institute, is based on a semi-distributed, process-based approach where the hydrological system is represented by a network of sub-basins. It



simulates components of the water cycle (i.e., snow accumulation and melting, evapotranspiration, soil moisture, streamflow generation, groundwater recharge, and routing through rivers and lakes) using a daily time step.

Statistical indices are currently the most commonly used tools for the detection of drought events. CLINT aims to advance traditional drought detection by defining AI-enhanced, impactbased drought indices that link the observed impacts of extreme droughts (e.g., reduction in electricity production or crop failures) with the candidate drivers of the event, including climatic, meteorological, and hydrological variables over different spatial and temporal scales.

However, traditional indices do not consider some relevant factors that may actually lead to impacts, which are ultimately determined by a complex set of social, economic, and environmental factors. Therefore, it is important to explore the link between drought indices and their impact.

In total, the analysis focuses on the calculation of eight statistical indices focusing on the 1, 3 and 6 months scale:

- Standardised Precipitation Index at 1-month scale (SPI-1) and 3-months scales (SPI-3).
- Standardised Precipitation and Evapotranspiration Index at 1-month scale (SPEI-1) and 3-months scale (SPEI-3).
- Soil moisture anomalies at the 1-month (SMA-1), 3-months (SMA-3), and 6-months (SMA-6) scales.
- Standardised Streamflow Index at 6-months scale (SSI-6).

The E-HYPE combined with the HydroGFD2.0 reanalysis were used for the computation of these indices for each month between 1993 and 2018 and for every sub-basin (35,408 in total), as reported in Figure 5.1.

To test the skill of the indices, the drought events detected were compared with the Fraction of Absorbed Photosynthetically Active Radiation (FAPAR) anomaly index. FAPAR is a biophysical variable derived from satellite observations and monitored by the European Drought Observatory (EDO), which measures the proportion of incoming solar radiation in the photosynthetically active radiation (PAR) range (400-700 nm) that is absorbed by the vegetation canopy of a particular area. Hence, the higher the FAPAR, the higher the photosynthetic activity, indicating a better state of vegetation. In this sense, the FAPAR anomaly (FAPAN) can be considered a proxy for drought impact.

5.3 Skills and performance of existing indices

Considering the large gradient of topography and climate conditions in the pan-European domain, a quantitative analysis was performed on predefined hydrological clusters among the sub-basins. Building on the work of Pechlivanidis et al. (2020), the E-HYPE sub-basins were divided according to their hydrological behaviours, and 15 streamflow signatures from the E-HYPE hydrological model setup were categorised, resulting in 11 clusters of different sizes and



variable distributions in the signatures. Figure 5.2 shows their spatial distribution, which did not necessarily include sub-basins that are geographically close. The properties that characterise the clusters are as follows: (i) key streamflow signatures, (ii) geographical domains, and (iii) dominant hydrological processes.



Figure 5.1: Occurrence of droughts (left) and mean duration of droughts measured in months (right) in 1993-2018, according to the different statistical drought indices.

A qualitative comparison between the statistical drought indices and the FAPAN index was produced in the form of a heatmap (Figure 5.3), which can also be used to investigate whether the drought events detected by these indices are aligned with each other. The example reported here refers to cluster 10 (see Appendix A5.1 for the other clusters). Each row in the figure represents a specific index, whereas on the x-axis, 312 months between 1993 and 2018 were reported. Each pixel was filled with a colour according to the result obtained in that specific month by each index. Red pixels denote identified drought events, and dark red pixels refer to extreme drought events. Not Available values (NA) are marked in grey; we have a long series of NA at the beginning of the time period for FAPAN since this index has been implemented in EDO since 2001.



CLINT - CLIMATE INTELLIGENCE Extreme events detection, attribution and adaptation design using machine learning EU H2020 Project Grant #101003876



Figure 5.2: Spatial distribution of Clusters. The 11 clusters are identified by ascending numbers from 1 to 11.



Figure 5.3: Heatmap: Cluster 10 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).





Figure 5.4: Scatterplot matrix: Cluster 10.

The FAPAN trajectory showed less intense drought events than those identified by statistical indices. This result is consistent with the analysis of the other ten clusters.

Quantitatively, these discrepancies between the statistical drought indices and FAPAN were confirmed using the scatterplot matrix (Figure 5.4). The figure shows both scatterplots (below the main diagonal) and correlation values (above the main diagonal) between the different indices, and the last column reports the correlation values between the FAPAN and the statistical drought indices. Considering the last column, the maximum correlation did not exceed 0.2 (0.17 between SPEI-1 and FAPAN), suggesting that the statistical drought indices failed in detecting drought impacts in these sub-basins. Similarly, a low correlation between the statistical drought indices and FAPAN indices was identified for the other clusters. These low correlation values are probably due to the discrepancies between the processes captured by the drought indices that refer to drought drivers and the impacts. Indeed, it is possible that during a meteorological drought captured by the SPI, the impacts on vegetation could be mitigated by the presence of water storage or groundwater irrigation (Zaniolo et al., 2018). In contrast, relevant impacts may have been registered in months when no meteorological drought was detected because of the challenges in irrigation supply.

5.4 Algorithms

The traditional indices analysed in the previous section failed to detect the drought impacts represented by the FAPAN values. In this context, the FRamework for Index-based Drought Analysis (FRIDA) was used to support the construction of new composite drought indices that could be tailored to the unique hydrological and meteorological conditions of a particular region, thus better capturing the impacts of a drought event.



Table 4.1: Set of candidate input features for W-QEISS.

Feature type	Feature name	Type of aggregation	Time aggregation [months]
	Precipitation (sumCPRC)	sum	1-3
Variables	Precipitation minus Pot. Evap.	sum	1-3
	(sumCPRC-sumEPOT)		
	Soil Moisture (meanSRFF)	mean	1-3-6
	Streamflow (meanCOUT)	mean	1-3-6
	SPI	-	1-3
Indexes	SPEI	-	1-3
	SMA	-	1-3-6
	SSI	-	6



Figure 5.5: FRamework for Index-based Drought Analysis (FRIDA).



FRIDA was structured in three steps (see Figure 5.5):

1. Identification of basin characteristics

The first step consisted of the selection of a target variable and collection of candidate predictors. The target variable is an appropriately chosen proxy for drought impacts in the basin, which was the FAPAN index in our study. The dataset of predictors contains candidate features to reproduce the target variable and consists of observed hydro-meteorological variables and composite drought indices computed over different spatiotemporal scales.

In this study, the resulting dataset was composed of 18 features, as listed in Table 4.1. Because the FAPAN index (target variable) is available only from the middle of 2001, the length of the time series is equal to 210 samples.

2. Feature extraction

The second step was devoted to the selection of more relevant variable subsets that better explained the selected target variable. This was done by employing an advanced input variable selection (IVS) algorithm, namely, the Wrapper for Quasi-Equally Informative Subset Selection (W-QEISS). A multi-objective evolutionary algorithm (in this work Borg MOEA (Hadka and Reed, 2013)) recursively explored the input space of candidate predictors to select Pareto-efficient subsets of predictors with respect to four objectives: (i) accuracy, (ii) cardinality, (iii) relevance and (iv) redundancy. To compute the predictive accuracy of each set, a calibrated Extreme Learning Machine (ELM, (Huang et al., 2006)) was used. The iterative search stopped when the termination criterion was met (i.e., the number of iterations).

Low-accuracy solutions were discarded from the pool of Pareto-efficient subsets, and the output of the W-QEISS algorithm was by consequence a set of quasi-equally accurate subsets.

In this work, The FRIDA experimental setting was the following:

- The number of function evaluations of the Borg MOEA was equal to 50,000.
- The number of hidden neurons in the ELM presenting a sigmoidal activation function was equal to 30.
- A k-fold cross-validation process (with k=10) was repeated 10 times, and the average of the Symmetric Uncertainty values over ten 10 experiments was used to estimate the predictive accuracy of each model.
- The W-QEISS experiment was run 10 times to filter out the random component of the search process. The final results were obtained by merging the Pareto fronts obtained by each repetition into a final set of non-dominated solutions.
- *3.* Drought index modelling



After selecting the preferred efficient solution within the set of quasi-equally accurate subsets, an appropriate regressor is fitted to the sample of inputs and target variable. The choice of the model class was determined by the application of interest.

5.5 Results

5.5.1 Improved indices and relevant drivers

New impact-based drought indices were formulated for each cluster using FRIDA to better capture drought impacts with respect to traditional drought indices.

Taking Cluster 10 as an example, the Selection Matrix obtained by running the input variable selection of W-QEISS is shown in Figure 5.6. It includes four subsets of predictors with accuracy values within a 20% range with respect to the highest one. The alternative subsets were sorted in ascending order of cardinality (from top to bottom) and accuracy (within each cardinality level). A coloured marker is placed at the intersection between the row that identifies a given subset and columns corresponding to the selected predictors. The marker colour varies with the cardinality of the subset, with lighter shades of grey indicating the smaller subsets. In this case, cardinality spans three to five features. The highest accuracy is reported in red. Moreover, the vertical bars traced by joining markers across multiple rows provide information regarding the relevance of a predictor.

Among the quasi-equally informative subsets reported in the selection matrix, we selected subset number 2 as the preferred one for Cluster 10: beside being the most accurate one, it resulted to include the two most relevant predictors and its cardinality is sufficiently small. In summary, the selected predictors were as follows:

- Mean soil moisture aggregated over 3 months (meanSRFF3);
- Mean runoff aggregated over one month (meanCOUT1);
- Standardised Precipitation Index aggregated over three months (SPI-3);
- Standardised Precipitation-Evapotranspiration Index aggregated over three months (SPEI-3).

The basins of this cluster, characterised by the typical streamflow responses of Mediterranean river systems, are located at low elevations and experience low flows and relatively low runoff coefficients due to high evapotranspiration. This is consistent with the selected predictors that were detected: meanCOUT1 reflects the low runoff coefficients and the possibility of worsening an already compromised situation even considering a short cumulation time, whereas SPEI-3 is related to the important role that temperature plays in this cluster. This result demonstrates the advantages of using FRIDA to support the automatic identification of the main drivers of drought conditions and related impacts.



CLINT - CLIMATE INTELLIGENCE Extreme events detection, attribution and adaptation design using machine learning EU H2020 Project Grant #101003876



Figure 5.6: Selection matrix: Cluster 10.

The same selection procedure was applied to the other clusters (see Appendix A5.2 for selection matrices).

Starting from Cluster 1, the most relevant predictors in the most accurate subset were identified as SMA-1 and meanSRFF6, while the latter reflected the slow dynamics of soil moisture in this cluster, where the streamflow characteristics are controlled by baseflow.

Cluster 2 is characterised by precipitation-driven river systems with frequent peaks and long recessions. Consistently, the most relevant predictors from the selection matrix were: meanSRFF3, SMA-6, and SPI-3. The first two regard the long recession aspect, whereas SPI-3 is related to the influence of precipitation in the streamflow signature.

For Cluster 3, the most relevant predictors were the meanCOUT1, SPI-3, and SPEI-3. This cluster is in fact marked by high interannual variability, particularly between the low and high



streamflow segments (confirmed by the presence of meanCOUT1) and snow-dominated streamflow regimes. Moreover, the dampening of streamflow due to the presence of lakes and wetlands and low actual evapotranspiration is coherent with the other two predictors associated with precipitation and temperature conditions.

In Cluster 4, regions receive high precipitation and are highly responsive. Occasionally, flow is regulated for hydropower production during winter (snow and ice melt), but still has some spring streamflow tendency. This was again confirmed by the most relevant predictors: meanCOUT1 reflects the significant role of streamflow, while sumCPRC-sumEPOT3 and SPI-1 are related to responsiveness to precipitation.

In Cluster 5, basins are characterised by a highly variable streamflow regime. The response is sometimes driven by snow melting, and this was confirmed by the most relevant predictor SSI-6 at longer terms; the streamflow is also precipitation-driven, which explains the high relevance of SPI-3.

Looking at the selection matrix of Cluster 6, the most relevant predictors were sumCPRCsumEPOT3, SPI-3, and SPEI-3. These variables are consistent with the hydrological description of the cluster: a highly variable streamflow regime quickly responds to precipitation, yet with long recessions. In particular, a long recession is evident by the seasonal cumulation time (three months) characterising all the selected variables.

Cluster 7 has elevated basins with a high variability in the streamflow regime. Precipitation causes flashy streamflow responses owing to low actual evapotranspiration, which results in high runoff coefficients. The most relevant predictors were SPI-1, SPI-3, and SMA-1, which confirm the fast dynamics given their short accumulation times. In particular, the first two indices are related to the important role of precipitation. In addition, the high runoff coefficients may lead to fast saturation of the soil, which led to the selection of SMA-1 in the matrix.

The selection matrix related to Cluster 8 reported SMA-1 as the most relevant predictor, with a relevance reaching 100%. In this cluster, the hydrographs are baseflow-dominated with a streamflow characterised by a small annual variability. The variables related to the soil component, such as the SMA, are mainly linked to a slow dynamic, regardless of the time scale for which they are computed.

Similar interpretation of Cluster 8 was also applied to Cluster 9. The basins in this cluster are characterised by highly baseflow-dominated streamflow with very little response to precipitation; therefore, the most relevant variables are SMA-3, SMA-6, and SSI-6.

In Cluster 11, basins are characterised by low runoff coefficients, yet they experience relatively high annual variability, that is, a fast response to precipitation and fast hydrograph recession. Streamflow can also be influenced by human practices (i.e., irrigation). According to the selection matrix, the most relevant predictors were SMA-3 and the meanSRFF1. The low runoff coefficient obtained indicates that what happens in the soil is sufficient to consider streamflow drought conditions.



5.5.2 Results: AI-enhanced drought indices

Concerning model class choice, on one hand Artificial Neural Networks (ANNs) provide a good balance between accuracy and flexibility. However, since they are black-box models, their interpretation is not intuitive. On the other hand, interpretability of linear models is easy and immediate to be understood in its physical meaning, though at the price of poorer approximation skills. Therefore, different model structures were considered: ANNs, ELMs, and linear models. The new impact-based drought index was represented by the best model structure that used the variables selected by W-QEISS to reproduce the FAPAN index trajectory as closely as possible to the observed one. The skill of the new impact-based drought index was assessed with the correlation between predicted and observed FAPAN. For Cluster 10, the selected model structure was a Deep ANN with 15 nodes subdivided into three layers, being able to reach a correlation value of 0.70. On average, considering all the clusters (see Table 5.1), the FRIDA indices increased the correlation with FAPAN by 0.35. The improvements obtained via FRIDA are shown in Figure 5.7.



Figure 5.7: Qualitative improvement obtained via FRIDA: on the left side, the catchments are filled with light colours, representing low initial correlation values between the best traditional statistical drought indices for each cluster (i.e., the one with the highest correlation) and the corresponding FAPAN; on the right side, the catchments are filled with darker colours, indicating higher correlation values between the FRIDA indices and FAPAN.

5.6 Summary and outlook

This chapter reports the results of Task T3.3 focused on the design of impact-based drought indices, with an application at the pan-European scale. Specifically, we first computed different standardised drought indices and assessed their skill in reproducing the drought impacts, represented here by the Fraction of Absorbed Photosynthetically Active Radiation (FAPAR) anomaly index. We then leveraged feature extraction algorithms using the FRamework for Index-based Drought Analysis (FRIDA) to identify new, impact-based drought indices via Machine Learning. Finally, we quantified the improved representations of drought impacts provided by the FRIDA indices and discussed the role of the selected variables involved in the index definition with respect to the characteristics of river basins.



Our results show that the FRIDA indices substantially advanced the detection of drought impacts related to the agricultural sector, as represented by FAPAN, with respect to traditional statistical drought indices. This improved detection method will then be used to advance climate services in WP6 and produce AI-enhanced projections of drought impacts, which will be reported in Deliverable D6.2. Moreover, the detection of drought impacts on other sectors, such as energy, will also be investigated. Finally, FRIDA will be tested at the local scale in selected Climate Change Hotspots in WP7, where stakeholders are interested in advancing drought detection.

Table 5.1: Numerical improvement obtained via FRIDA. 2nd column: best model structure selected; 3rd column: selected predictors for each cluster, 4th column: best initial correlation value between traditional drought indices and FAPAN; 5th column: final correlation value between FRIDA index and FAPAN; 6th column: improvement obtained for each cluster via FRIDA.

Cluster	Impact-based drought	Selected	Best initial	Final	Improvement
	index structure	predictors	correlation value	correlation value	
1	Deep ANN (15 nodes)	meanSRFF6	0.36	0.58	0.22
		SMA-1			
2	Shallow ANN (20 nodes)	meanSRFF3	0.28	0.77	0.49
		meanCOUT1			
		SPI-3			
		SPEI-1			
		SPEI-3			
3	Deep ANN (15 nodes)	sumCPRC-sumEPOT1	0.31	0.63	0.32
		meanSRFF1			
		meanCOUT1			
		SPI-3			
		SPEI-3			
		SMA-1	0.10	0.00	
4	Shallow ANN (20 nodes)	sumCPRC-sumEPOT1	0.12	0.69	0.57
		sumCPRC-sumEPOT3			
		meanSRFF1			
		meanSRFF3			
		SDEL 1			
	$\mathbf{D}_{\text{res}} = \mathbf{A} \mathbf{N} \mathbf{N} \left(20 - 1 - 1 - 1 \right)$	SPEI-1	0.10	0.57	0.99
9	Deep ANN (30 hodes)	SP1-3 SMA 2	0.19	0.57	0.38
		SMA-5			
		SNIA-0 SSL6			
6	Deep ANN (15 nodes)	sumCPBC sumEPOT3	0.20	0.70	0.41
0	Deep Mill (10 houes)	SPI-3	0.25	0.10	0.41
		SPEI-3			
		SMA-1			
7	Shallow ANN (15 nodes)	meanSRFF6	0.27	0.57	0.30
		SPI-1			
		SPI-3			
		SMA-1			
8	Deep ANN (30 nodes)	SPI-1	0.37	0.54	0.17
	- , , ,	SMA-1			
9	Shallow ANN (20 nodes)	sumCPRC3	0.33	0.50	0.17
		meanSRFF3			
		SMA-3			
		SSI-6			
10	Deep ANN (15 nodes)	meanSRFF3	0.17	0.70	0.53
		meanCOUT1			
		SPI-3			
		SPEI-3			
11	Shallow ANN (30 nodes)	meanSRFF3	0.12	0.44	0.32
		SMA-3			



6 COMPOUND EVENTS AND CONCURRENT EXTREMES

6.1 Overview

The objective of T3.4 is to study compound events and their impacts on the food, water, and energy sectors in Europe. For such temporally or spatially simultaneous or lagged climate extremes, the interconnectedness of droughts and heatwaves is often of interest, as these types of events are strongly linked through their physical nature (MS24 and D4.1) and have high impact on socio-economic sectors, such as crop yields, water, vegetation and energy (Hao et al., 2022). In the food sector, for example, droughts and heatwaves can reduce cereal yields by 9-10% at national level, and these types of phenomena can explain 40% of the yield interannual variability (Lesk et al., 2016; Zampieri et al., 2017). Extreme events being dependent in space and/or time are often called concurrent extremes (Toreti et al., 2019b). Furthermore, multiple climate events that are not individually extreme can also be associated with disproportional socioeconomic impacts, as ecosystems may not be directly adapted to the covariability of temperature and precipitation, so that bivariate anomalies can have large effects without these variables being univariately extreme (Mahony and Cannon, 2018). Such events are called compound events (Zscheischler et al., 2018; Zscheischler et al., 2020). An example of these types of events are the so-called *false-spring events*. These events occur when above-average wet and warm conditions prevail in winter, leading to early plant growth, followed by severe drought or frost in the following spring, resulting in significant crop losses (Allstadt et al., 2015; Chamberlain et al., 2019). These events are expected to occur even more frequently in the future due to climate change (Ault et al., 2013; IPCC, 2021) with record hot and dry summers like 2018 potentially becoming the norm by the mid-century (Toreti et al., 2019a). This emphasises the need to study these types of events to adequately assess the risk, which can be underestimated by considering only single extreme events (Wahl et al., 2015; Zscheischler and Seneviratne, 2017). We studied interconnected drought and heatwave events at the global level and wet and warm later winters together with dry and warm springs, focusing on the corresponding impact on winter wheat yields in Europe.

We first review the datasets, drivers and known climate indices in sub-chapters 6.2 and 6.3, followed by a discussion on the AI and hybrid approaches employed for the analysis in chapter 6.4. Finally, chapter 6.5 discusses the current results and future directions of this work.

6.2 Datasets and candidate drivers

6.2.1 Datasets

The analysis of concurrent extreme events focused on global interconnectivities of droughts and heatwaves, while the analysis of compound events dealt with multiple climate events and their impact on the water, energy, and food sectors in Europe. For all meteorological variables, the ERA5 (Hersbach et al., 2020) reanalysis was used, except for Sea Surface Temperatures (SSTs), which stemmed from the DOISST (Huang et al., 2021). For the impact-related datasets, quality-controlled data from AGRI4CAST and EUROSTAT were utilised for the



food sector, while data from the European Network of Transmission System Operator of Electricity were obtained for the energy sector. For impacts in the water sector, model data from the E-Hype hydrological model operated by the Swedish Meteorological and Hydrological Institute were used. A detailed description of the datasets can be found in D3.1. For this deliverable, we focused on compound events impacting the food sector. Other sectors will be covered in future research.

6.2.2 Candidate drivers

While the drivers of compound events and concurrent extremes are intensively discussed in the frame of MS24 and D4.1, a summary is provided here for a comprehensive report. Drivers of compound events and concurrent extremes can be divided into the following subcategories: atmospheric circulation patterns, land-atmosphere interactions, ocean dynamics and coupled ocean-atmosphere patterns, and the climate change signal due to anthropogenic forcing. Multiple hot and dry climate extremes are frequently associated with persistent blocking highs, subtropical highs, and stagnation events (Zhang et al., 2021). The prevalence of high-pressure systems is related to increased shortwave radiation, reduced moist air inflow, and can impose a shift in storm track paths (Dong et al., 2018; Schumacher et al., 2019; Kautz et al., 2022), leading to stable dry conditions and clear skies with high air temperature conditions. Relevant and connected circulation patterns for such events are stationary jetstream positions (Duchez et al., 2016), persistent circulation modes such as the positive-phase North Atlantic Oscillation (NAO; Hao et al., 2019b; Mukherjee et al., 2020), the Scandinavian Pattern (Bueh and Nakamura, 2007) and Rossby waves (Kornhuber et al., 2020; Ionita et al., 2021). Another important modulator for warm and dry conditions is soil moisture (Miralles et al., 2019). Soil moisture deficits reduce evapotranspiration, resulting in reduced latent heat, thus leading to enhanced local heat (Barriopedro et al., 2023; Domeisen et al., 2023). During droughts, soil moisture affects the atmospheric evaporative demand (AED) as the combination of low relative humidity (RH), high air temperatures and low cloud cover increases the AED and triggers soil evaporation and plant water consumption through transpiration (Miralles et al., 2019). Plants also react to these conditions by closing their stomata to prevent water loss, which further reduces evapotranspiration (Massmann et al., 2019). Increased AED can exacerbate agricultural and environmental droughts by further stressing crops and increasing water use in irrigated areas (García-Garizábal et al., 2014), thereby contributing to hydrological droughts (Vicente-Serrano et al., 2017). SST anomalies can also force persistent atmospheric circulation patterns connected with heatwaves and drought conditions (Domeisen et al., 2023). Furthermore, teleconnection patterns related to SSTs have been shown to positively influence droughts and heatwaves such as El Niño Southern Oscillation (Hoerling et al., 2013; Hao et al., 2019a; Cai et al., 2020), Pacific Decadal Oscillation (Nguyen et al., 2021), Indian Ocean Dipole, as well as combinations of these patterns when they are in phase (Zanchettin et al., 2008; Steptoe et al., 2018; Nguyen et al., 2021). Finally, anthropogenic influence is an individual driver of drought and heatwave events, as the global temperature increase leads to a higher frequency of heatwaves. Together with the interconnectivity of heatwaves and droughts this implies that hot and dry conditions will become more frequent in the future (IPCC 2021).



Warm and wet conditions are modulated when temperature increases over open water bodies, thereby increasing surface humidity (Zhang et al., 2021). However, this effect can be limited over land (Fischer and Knutti, 2013). Intense heat can also reduce sensible heat flux and moisture convergence, resulting in extreme precipitation events (IPCC, 2021). Studies have further linked warm and wet conditions to the advection of warm air originating in tropical areas (Katsafados et al., 2014; Freychet et al., 2017). However, recent reviews have noted that the drivers of wet and warm events are still less well understood (Raymond et al., 2021; Zhang et al., 2021). Future projections show that these types of events are expected to increase (Russo et al., 2017; Meng et al., 2022) with potentially very critical implications for human health (Davis et al., 2016). For example, it is being discussed that the 6-hour wet bulb temperature in tropical and subtropical Asia could rise to over 30 °C by 2100 and possibly even exceed 35 °C, which is considered a critical threshold for the survival of humans (Pal and Eltahir, 2016). This further highlights the importance of studying such events.

6.3 Existing indices

This section briefly describes the climate indices and algorithms used to analyse compound events and concurrent extremes. For a more detailed description, we refer to D3.1 and the literature.

6.3.1 Climate indices

To describe wet and dry conditions, the standardised precipitation and evapotranspiration index (SPEI; Vicente-Serrano et al., 2010) was utilised, which is essentially a normalised water balance index. The derivation of the water balance requires an approximation of the evapotranspiration, for which we employed the Hargreaves-based approach by Droogers and Allen (2002). To describe the impact of warm temperature-related impacts on agriculture, we employed the Active Temperature Sum (ATS), which is the aggregated daily temperature above 0 °C and a canonical metric to describe the consequent plant phenological phases within the growing season (Ceglar et al., 2019). The Heat Magnitude Day (HMD) is defined as the cumulative maximum temperature exceedances during a heatwave, where a heatwave is commonly defined as days on which the maximum temperature exceeds the long-term 90th percentile for at least three consecutive days (Perkins and Alexander, 2013). The HMD was employed to detect heatwave events, as it has been shown to capture well the covariability between heatwaves and droughts and the resulting impacts on agriculture (Zampieri et al., 2017; Toreti et al., 2019b).

6.3.2 Statistical measures for dependence

A statistical tool for detecting the dependencies of large-scale heatwaves and droughts is the inhomogeneous J-function (Cronie and van Lieshout, 2016; Toreti et al., 2019b). The main advantage of these types of functions is the ability to take into account the non-stationarity of the occurrence of extreme events, which must be assumed as the frequency of heatwaves and droughts is expected to alter in the course of climate change (IPCC, 2021). For instance, if we examine connectivities of heatwaves in two different regions, these types of



dependencies might be confused with the shared forcing or the co-trending of the two phenomena. Furthermore, the trend can be modelled nonparametrically through an inhomogeneous intensity function (e.g., Diggle, 2014), which can offer useful insights into the temporal evolution(s) of these types of events without making strict assumptions. The J-function models three types of dependency structures: Independence, Clustering and Inhibition (see e.g., Baddeley et al., 2016). However, the final decision as to which dependency structure the observed phenomenon belongs to is still determined graphically by the user and is therefore characterised by subjectivity. Furthermore, visual analysis of the output does not allow the function to be applied to large datasets and/or ensembles. In CLINT, we propose an AI-based automation tool based on Monte Carlo simulations to circumvent this problem (see chapter 6.4.3).

6.4 Algorithms

The following section describes the basic concepts of the implemented algorithms. More technical details are provided in WP2 deliverables.

6.4.1 Time series clustering algorithms

Clustering algorithms offer a flexible framework for defining subgroups in high-dimensional datasets. However, many classical clustering methods, such as k-means (e.g., Hastie et al., 2017) calculate clusters based on constant centroids, which are potentially unable to learn the temporal evolution of the time series, including nonstationarities imposed by climate change (IPCC, 2021). A well-known framework to employ time distance-based measures and avoid the above problematics is the Dynamic Time Warping (DTW), which has also been intensively used for clusters (e.g., Aghabozorgi et al., 2015; Sardá-Espinosa, 2019). We employed a recently developed approach for DTW called soft-dynamic time warping (SDTW; Cuturi and Blondel, 2017), which was seen to significantly outperform baseline clustering algorithms. In addition, the individual time series can have different time lengths and/or time intervals and can be multivariate. Considering the covariability of climate components (such as droughts, heatwaves, and agro-climatic regions), the multivariate approach increases the sample size and leads to more robust statistics. These features made this method particularly useful for our purposes.

6.4.2 Regularised generalised canonical correlation analysis (RGCCA)

Canonical correlation analysis (CCA) can identify large-scale predictors and relationships between climate variables. However, to make this algorithm feasible for high-dimensional settings, the feature space needs preprocessing using, for instance, principal component analysis (e.g., Wilks, 2011) to handle the spatial correlations in climate data. A promising approach called Regularized Generalised Canonical Correlation Analysis (RGCCA; Tenenhaus and Tenenhaus, 2011; Garali et al., 2018) introduces adequate regularisation schemes and can deal with high-dimensional predictors and high collinearity in the features. Furthermore, the approach can handle multiple variables and can be augmented by a priori graph structure for preliminary hypotheses, thus setting a stage for a very flexible framework covering many




well-known multi-block methods as special cases (Tenenhaus et al., 2017). However, this approach is based on maximising the covariances between variables such that only linear relationships can be studied. Tenenhaus et al., (2015) showed that RGCCA can be extended to nonlinear relationships by making use of kernels (e.g., Efron and Hastie, 2016). Rahimi and Recht (2007) proposed an elegant mapping method making use of random features, which are designed to have an approximately equal inner product to those in the feature space of a shift-invariant kernel. Loosely speaking, the idea behind this is that applying linear algorithms to the random features leads to a similar result as using kernel-based approaches, but since the algorithm is based on linear methods, it is faster and potentially numerically more efficient. Indeed, they show that these methods often outperform kernel-based approaches on benchmark datasets. Further theoretical and simulation evidence was presented by Sutherland and Schneider (2015) and we used their methods to approximate the Gaussian kernel, which has desirable characteristics for climate variables (Hannachi, 2021). We combine the random feature-based approach with RGCCA to obtain a non-linear RGCCA, which, to our knowledge, has not yet been implemented in climate science.

6.4.3 Imbalanced random forests

As substantiated in the introduction, multiple climate events that lead to high socio-economic impacts can also be a combination of non-extreme climate events, the impacts of which may be negligible individually but may be harmful when combined (Mahony and Cannon, 2018; Zscheischler et al., 2018). An additional difficulty arises with respect to the thresholds for identifying such events. In other words, what thresholds are "warm enough" or "wet enough" to be associated with high socio-economic impacts? Within CLINT, we analysed these questions using Random Forests (RF) as they perform forecasting and classifying based on partitioning the dataset. Exploiting these splitting bounds can be useful for identifying the thresholds that lead to a high impact, since they have a certain degree of optimality in predicting the desired outcome and are objectively calculated by the algorithm. An additional difficulty, however, is that multiple climate events seldom occur, by definition, and are thus constrained to be highly imbalanced. Recently, O'Brien and Ishwaran (2019) proposed an imbalanced RF approach based on the q-classifier to mitigate this issue. If the number of events is still too small to be adequately trained, they show that their algorithm can be augmented with oversampling methods, for which they suggest the Majority Weighted Minority Oversampling Technique (MWMOTE) algorithm (Barua et al., 2014).

6.4.4 AI-enhanced inhomogeneous J-function

The inhomogeneous J-function allows the user to explore whether the obtained extreme events showed clustering, inhibition, or independence while considering their non-stationarity (Toreti et al., 2019b). However, the choice of which dependency structure the given dataset belongs to is still subjective. The main goal in CLINT is to circumvent this subjectivity by introducing an AI-based interpretation tool that makes the method applicable to large datasets, such as ensembles. The method is applied to automatically identify the relationships between large-scale droughts and heatwaves on a global scale. For this purpose, well-known point process models are implemented to simulate surrogate data that mimic the



CLINT - CLIMATE INTELLIGENCE



desired dependency structures so that they can be labelled, and a classification problem can be trained. We used Monte Carlo simulations based on numerical experiments by van Lieshout (2011) and Cronie and van Lieshout (2015) to simulate a large dataset of surrogate data, and we estimated the (marked) inhomogeneous J-function for each simulated dataset following Cronie and van Lieshout (2016). Specifically, we used the inhomogeneous Poisson process for simulating independent samples, the Log-Gaussian Cox Process for clustering, and the Mattern-Type-II process for inhibition (see e.g., González et al., 2016). Furthermore, we estimated the intensity function non-parametrically using a recently proposed resamplesmoothed Voronoi estimator, which has been shown to outperform kernel-based approaches (Cronie and van Lieshout, 2018; Moradi et al., 2019). With this setup, we simulated an arbitrarily large training set and used deep learning methods, as they can process multivariate data well. Furthermore, we were interested in the highest possible accuracy of the predictions such that there are no "black box" problems (McGovern et al., 2019; Kashinath et al., 2021;

Schultz et al., 2021). We used a combination of Convolutional Neural networks and Gated Recurrent Units (e.g., Chollet et al., 2022) to train the classifier showing already desirable performance with relatively simple networks.

6.4.5 Non-parametric SPEI

As described above, the SPEI is based on the probability integral transformation and is generated by estimating a seasonally dependent distribution function of the water balance, which is then transformed using Inverse Transform Sampling to obtain a standard normal distributed time series. This requires an appropriate choice of distribution function for the variable, most commonly the log-logistic in the case of water balances (Vicente-Serrano et al., 2010; Beguería et al., 2014; Vicente-Serrano and Beguería, 2016). Some studies have indicated that the approach works well on the monthly scale, but might not be optimal on the daily scale, for which the Generalised Extreme Value Distribution can be superior (Stagge et al., 2015). More importantly, the log-logistic distribution is bounded, and the support depends on the associated parameters, such that values outside the support are assigned a zero probability and cannot be mapped through Inverse Transform Sampling. This can be critical when the distribution is calibrated on reference periods, as values outside the support or unforeseen values cannot be extrapolated. An example of this is given in Appendix A6.1 for "wichita" the time series used in SPEI demo (https://cran.rproject.org/web/packages/SPEI/index.html).

Reference periods are important tools in climate science, and we were additionally interested in the link between droughts and heatwaves. Heatwaves are, however, commonly defined based on reference periods (Barriopedro et al., 2023) and when studying their links with droughts, this problem must be taken into account, as the comparison of two time series calibrated on different reference periods is inappropriate. We used a nonparametric kernelbased estimator of the distribution function with an unbounded kernel function instead of the log-logistic distribution. By construction, the estimator approximates the true (unknown) distribution function and can circumvent the extrapolation problem as the estimator is a superimposition of unbounded kernel functions (thus unbounded). We employed a localpolynomial likelihood kernel estimation as it can correct also for higher moments (Loader,



1996; Loader, 1999). Originally, these types of estimators were only feasible for unbounded continuous (random) variables, but recent studies (Geenens, 2014; Geenens and Wang, 2018; Nagler, 2018a; Nagler, 2018b) have shown that they can be extended to discrete variables, mixtures of discrete and continuous variables, and bounded variables. This allowed the mapping trick used to construct the SPEI to be extended to other climate variables, such as temperature, total precipitation, relative humidity, and cumulative intensities, to obtain standardised non-seasonal climate indices. For instance, we used this approach for the non-Gaussian ATS to obtain the standardised non-seasonal active temperature sum (SATS) used in the study of compound events. Finally, the proposed estimators could be combined with vine copulas (e.g., Czado and Nagler, 2022) to construct multivariate standardised climate indices.

6.5 Results

6.5.1 Compound events

The investigation of compound events focused mainly on relatively wet and warm late winters, followed by dry and warm springs, with the corresponding (agricultural) impacts on winter wheat yields in France, the largest winter wheat producer in Europe.

6.5.1.1 Agro-climatic sub-regions

France can be divided into climate subregions, among which the contribution of meteorological drivers to crop yield varies remarkably (Ceglar et al., 2016). As described above, clustering was employed for SATS, the non-parametric SPEI (NP-SPEI), and the observed winter wheat yields. To study the impacts of climate on crop yields, the adaptation effect defined by the improvement in the agricultural practices must be taken into consideration. A commonly made assumption is that the adaptation effect is mainly captured by the multi-annual trend, such that non-linear detrending can be applied to remove this effect (Ceglar et al., 2016; Zampieri et al., 2017). We used local polynomial smoothing as the detrending approach, with the bandwidth chosen as in Feng *et al.* (2020). For ease of interpretation, we multiplied the derived crop yield anomalies with -1 such that positive anomalies are associated with high impacts. After the trend was removed, the SDTW clustering algorithm (Section 6.4.1) was applied to all time series as a multivariate clustering problem. The outputs of the clusters are shown in Figure 6.1.

The clusters (similar to those of other studies, e.g., Ceglar et al., 2016) resulted in reflecting climate conditions well and identifying the Mediterranean regions in cluster one, central continental France in two, northern oceanic France in three, southwestern-Pyrenees France in four, and eastern mountainous France in five. The SDTW approach produced a time series as the centroid, as shown in Figure 6.2. These captured high-impact events, such as 1998, 2003, and 2016 differently throughout the sub-regions. Table 6.1 gives an overview of the total wheat produced in the five clusters.

Even though the northern regions (clusters two, three and five) were found to generate approximately 80% of the total wheat and might already be appropriate for studying the





agricultural impacts, we found significant improvements when all regions were studied, as the local effect of climate could be better modelled.



Figure 6.1: Obtained agro-climate regions from multivariate clustering approach.

Table 6.1: Total and relative contribution of each cluster to the total winter wheat yield in France.

Yield	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Total (t/ha)	1658.96	4000.80	5671.78	1885.91	3382.27
Relative	10.00 %	24.13 %	34.89 %	11.37 %	20.40 %

6.5.1.2 Relevant drivers

For the analysis, we first generate the random features for all climate variables following the recommendations of Sutherland and Schneider (2015). For winter wheat such a transformation was not employed to enhance the interpretability of the analysis, as the predicted variable of the experiment could be better understood. The analysis could be augmented using an a priori graph of the connections. We connected all variables to winter wheat to identify its predictors, and we connected winter wheat to itself to identify the dominant patterns. Finally, we interconnected all climate variables to model the interdependencies of the climate variables separately for each season. In addition, as the clusters are intercorrelated, we connected the yields in different clusters. The local climate resulted to be represented by NP-SPEI and SATS, while large-scale patterns were represented by SSTs, geopotential height at the 500 hPa level, and RH at the 700 hPa level. We estimated the final model jointly for all the clusters.

Winter wheat basic function values were all positive for all clusters, except for cluster four, where 75% of the weights were positive. This allowed us to interpret the retained winter



wheat time series for each cluster as positive weighted averages so that positive values could be associated with impacts or crop yield losses. Figure 6.3 shows the time series of the retained components, where for the northern clusters, the well-known 2016 crop failure is visible, while for the southern regions 1998 can be identified.



Figure 6.2: Centroids corresponding to the Clusters displayed in Figure 6.1.

A shortcoming of non-linear kernel-based methods is the lack of one-to-one mapping from the kernel input space to the original feature space (Scholkopf et al., 1999), implying that the eigenvectors cannot be displayed in the usual manner. To identify monotone relationships, we correlated each grid point with the retained projected component using the Spearman correlation coefficient. We tested the significance of the correlation using a Student's t-test and considered the multiple testing problem by applying the false discovery method (Benjamini and Hochberg, 1995), which was also shown to effectively deal with spatial correlations (Wilks, 2006).

For the NP-SPEI displayed in Figure 6.4, January and February were significantly associated with wet conditions in almost all regions and a drought pattern emerged in April. March and May exhibited weak and inconsistent signals, respectively. The SATS values in Figure 6.5 show warm patterns in January, February, and April throughout France, while in March and May, they partially show dominant warm patterns. Hence, it was found that non-linear and multivariate analyses can effectively filter out events similar to the desired compound events. Figure 6.3 marks in orange lines the events when the projected patterns of NP-SPEI and SATS were in a positive phase, demonstrating that they were mainly associated with yield losses for most of the clusters.







Figure 6.3: Obtained projected components from the non-linear RGCCA for the winter wheat series of individual clusters. Orange lines indicate time points at which the associated SATS and NP-SPEI time series were in a positive phase.

Next, we explored the associated large-scale circulation of these patterns. Figure 6.6 shows the 500 hPa geopotential height, where, except for cluster one, a pattern similar to a positive NAO is observed in January and especially in February. In April, a tripole is visible in the different clusters, associated with strong blocking and consequently with dry and warm conditions (see Chapter 6.2), consistent with the NP-SPEI and SATS patterns (Figures 6.4 and 6.5). Dry conditions in April also emerge in the 700 hPa RH patterns (Figure 6.7). However, positive correlations over Central Europe in February indicate wet conditions.

In Figure 6.8, warmer SSTs along the French Atlantic coast are observed for January and February. Considering the cooler atmosphere during winter compared to the ocean, these stronger differences might trigger enhanced heat exchange towards the atmosphere, increased evaporation, and higher air humidity. The westerly circulation in Figure 6.6 at 500 hPa may thus be connected to increased precipitation over land and wet conditions, suggesting a potential mechanism connecting the three different large-scale patterns observed in the winter months.





Figure 6.4: Spearman correlation coefficients for obtained projected time series of non-linear RGCCA. Only statistically significant values at the 90% confidence level are displayed.



SATS patterns

Figure 6.5: Same as figure 6.4, but for SATS.





500hPa Geopotential Height - Patterns

Figure 6.6: Same as Figure 6.4, but for 500 hPa geopotential height. Contour lines indicate statistical significance at the 90 % confidence level.





700 hPa Relative Humidity - Patterns

Figure 6.7: Same as Figure 6.6, but for relative humidity on the 700 hPa level.

To further understand the non-linear relationship between crop yield and climate variables, we modelled the non-linear impact of these derived components on winter yield using D-vine copula-based regression (Kraus and Czado, 2017; Schallhorn et al., 2017). Because including too many variables might lead to overfitting, and estimation can be hindered due to the small sample size, a good balance between the model's complexity and predictability is required. Kraus and Czado (2017) recommend using the Akaike Information Criterion to determine the best subset of predictors. The results for cluster three (associated with the highest winter wheat yield) are shown in Figure 6.9, and the marginal effect of each component on crop yield was estimated following Schallhorn et al. (2017).





Sea Surface Temperature - Patterns

Figure 6.8: Same as Figure 6.6, but for SSTs.



Figure 6.9: Estimated Marginal effects of D-Vine-Copula based quantile regression model for predicting yield variability in cluster three. Values of alpha represent the conditional quantile for which the model was estimated.



The chosen variables or statistical predictors for this cluster are the winter NP-SPEI, spring RH, SST, and 500 hPa geopotential height. The algorithm tended to choose more large-scale predictors (Figure 6.9), which is interesting because they modulate the local climate variables and are apparently also better statistical predictors. Hence, higher crop yield losses were observed during the positive phases of the variable (xk > 0). For the negative phase, only moderate impacts were found, except for RH-spring, where a non-linear increase in the impact could be observed with the strengthening magnitude of the pattern (i.e., high RH-spring values). For the SST-spring, for instance, the effect resulted to be approximately linear for the positive phase. Nevertheless, Figure 6.9 shows that the relationship between climate variables and impacts on winter wheat were found to be non-linear functions, underlining the added value of the non-linear analysis.

6.5.1.3 Learning compound event definition

Compound events are generally regarded as not necessarily extreme multivariate climate events that lead to high socioeconomic impacts (e.g., Zscheischler et al., 2018). This "nonextremeness" can pose a challenge for their definition and selection, as the thresholds to be chosen for their characterization cannot be chosen beforehand. For example, it is not clear which temperatures are sufficiently high, or water balances sufficiently low, to be associated with high impacts. These thresholds can be indicated by using decision trees, as they split the variables for prediction and classification and the corresponding internal nodes (James et al., 2021) gives a suggestion for "warm enough" or "dry enough" from a statistical or predictionbased point of view. Multivariate climate events seldom occur by definition, so we employed imbalanced RFs based on the q-classifier (O'Brien and Ishwaran, 2019). To find the splitting points, we extracted the first internal node to distinguish between compound and noncompound events. This was the safest choice, because deeper nodes become more uncertain and unimportant for predictions (Ishwaran et al., 2010). We defined the compound event of interest (wet and warm late winter, together with dry and warm springs) as classes for the classification. Based on the insights of Section 6.5.1.2, we could define warm conditions by setting SATS-2 > 0 for February and NP-SPEI-2 > 0, both values above average for the corresponding month. Similarly, we set -NP-SPEI-2 > 0 and SATS-2>0 in April to obtain the overall warm and dry conditions from March to April. We used the negative NP-SPEI-2 in April so that the RFs could search for strictly positive bounds in all variables and associate negative values with the absence of compound events. At this point, we defined two types of compound events: meteorological compound events (MCE) and pure compound events (PCE). An MCE takes place when all the meteorological conditions (i.e. NP-SPEI and SATS being in the state described above) are fulfilled and the PCE event is defined when, additionally, the 70th percentile of the wheat yield reduction is exceeded, indicating "high agricultural impact". Thus, for the latter, we defined a multivariate climate event with corresponding socio-economic impacts, thus falling into the class of compound events (Zscheischler et al., 2018).

To recognize the desired events (MCE or PCE) starting only from the climate variables, these are used as predictors in the training of the classification problem. For the MCE, only the state of the climate needs to be recognized, whereas for the PCE, events with high impact also need



to be predicted. The MCE is considered a control experiment to show that, for compound events, the decision to estimate the thresholds with RF works. If successful, the RFs should be able to recognize that for the MCE, the input variables are in a positive state such that the extracted decision bounds cluster around zero. On the other hand, when including high impacts in the class, the RF would theoretically need to adjust the bounds to capture agricultural impacts. In other words, it is necessary to determine which temperature or water balance anomalies are sufficiently warm or wet/dry to be associated with a high impact, which is our question of interest. We expect the extracted limits from the PCE to be higher than those of the MCE, as it takes a significant but potentially non-extreme anomaly to damage the crop.

However, for the entire experiment, we first had to check the capability of the RF to predict the desired events. For this preliminary assessment, we trained the RF for each of the clusters shown in Figure 6.1, considering the agro-climatic zones and climate sub-regions in France. We considered accuracy as a classical intuitive metric and the geometrical mean (G-Mean; Kubat et al., 1997) as a measure for imbalanced categorization, which is the recommended metric for the q-classifier (O'Brien and Ishwaran, 2019). Table 6.2 shows the results for the test set, defined as the most recent one-third of the observations. We also included "cluster O" in the experiment, which is simply the full dataset without sub-regions, to evaluate whether using the clusters adds value to the RF performance.

Cluster	MCE-Accuracy	MCE – G-Mean	PCE - Accuracy	PCE – G-Mean
1	0.88	0.88	0.95	0.94
2	0.99	0.99	0.98	0.93
3	0.97	0.97	0.92	0.94
4	0.97	0.98	0.96	0.98
5	0.98	0.93	0.97	0.83
0	0.99	0.92	0.99	0.91

Table 6.2: Performances of RF for each cluster. Cluster one to five correspond to the clusters in Figure 6.1 and cluster zero means that the training is done for the full region.



We observed that the performance of the MCE appeared to be very good, with the metrics being mostly higher than 0.9 (being 1 the optimal value) except for cluster one with an accuracy and G-mean of approximately 0.88. In comparison, the performance of the PCE was slightly worse, except for cluster five with a potentially undesirable G-mean of 0.83. Furthermore, if we ignored the local climate and performed the RF for the entire region (denoted as cluster zero), we obtained the second-worst performance in terms of G-mean for PCE and MCE. This proved that the prediction can be improved by considering the local climate. In summary, the performance of the RFs was satisfactory.



Figure 6.10: Retained splitting bounds from RF for the SATS. Light blue lines indicate the PCE and orange lines the MCE. Points are displayed below together with a kernel density estimator for graphical orientation.

The report now focuses on the results related to the extracted internal nodes. Figures 6.10 and 6.11 show that, except for SATS-winter in cluster two to four and SATS-spring in cluster four, the MCE results are grouped around zero. This indicates that the extraction of bounds worked for all variables except SATS-winter. This finding requires further investigation and cannot be explained at this stage. As expected, the bounds of PCE resulted to be larger than or equal to those for MCE because small variations could be compensated by the crop's resistance, but serious damage can occur when these types of conditions intensify. We statistically verified whether the bounds for the MCE were stochastically smaller than those of the PCE using the Kolmogorov–Smirnov test. We also computed the mean of the reported statistics as an additional intuitive measure. It was observed that the KS-test was indeed



significant for most of the panels and the mean of the PCE was larger. The latter is potentially a first suggestion for an objective bound for defining the thresholds "wet enough", "dry enough" and "warm enough" to be associated with high impacts on the crop yields.

Furthermore, all the means were found to be smaller than one (i.e. below one standard deviation of the standardised input variables), suggesting that they could hardly be considered extremes and that non-extreme events could actually be associated with high impacts. Our idea of extracting bounds showed plausible results and could be a promising step towards the definition of thresholds for non-necessarily-extreme events that lead to large socio-economic (in this case, agricultural) impacts. However, because we only extracted the first internal node of the tree, it is likely that we have not yet optimally processed the results of the classification trees. This will be the focus of future applications.



Figure 6.11: Same as Figure 6.10, but for NP-SPEI.

Finally, because the performance of the RFs in the first experiment was undesirable, the data were augmented with oversampling approaches. For this, we used the MWMOTE approach, simulating so much surrogate data that the minority class had approximately 20% of the data. Higher values were not found to increase RF performance.



6.5.2 Concurrent extremes

In contrast to the analysis of compound events, the focus of the concurrent extreme events analysis was on dependent extreme events that are spatially and/or temporally linked, and the report will focus on the dependencies of large-scale droughts and heatwaves.

6.5.2.1 Non-parametric SPEI

To validate our proposed method with the NP-SPEI, we compared the SPEI for different accumulation schemes, namely 1, 3, 6 and 12. First, we investigated whether the NP-SPEI could better extrapolate unforeseen events associated with infinite SPEI (obtained when the quantile function of the standard normal distribution is applied to 0 or 1). For this study we restricted the analysis to the Northern Hemisphere, as it has more land and higher data quality. Figure 6.12 shows the results for SPEI calibrated for the full time period (1940-2022) and figure A.6.2.1, when they are calibrated on the reference period 1961-1990.

Number of non-extrapolatable points with



Figure 6.12: Number of non-extrapolatable points, when the log-logistic distribution is used as a mapping function for the SPEI.

The SPEI calibrated during the full period performed well and only showed small regions where non extrapolatable points occurred. Beguería et al. (2014) stated that this issue occurs mainly in very dry climates or high altitudes. However, when focusing on the SPEI calibrated for the 1961-1990 reference period, non-extrapolatable points appeared quite frequently across the globe. This suggests that the extrapolation issue is a major problem for the SPEI,



and that the log-logistic distribution may not be appropriate if the SPEI is calibrated on reference periods. For NP-SPEI (figure A.6.2.1), this phenomenon almost vanished. Hence, the NP-SPEI was quite successful in performing extrapolation compared to the original SPEI.

This finding sets the stage for our next experiment, in which we compared the NP-SPEI to the SPEI and judged whether the NP-SPEI can be an adequate drought index. For this purpose, we correlated the two SPEI versions and computed the upper and lower tail dependence (see e.g., Coles, 2004). We set the quantile to exceed to -1 for drought detection and to 1 for wet event detection, which are the typical thresholds for detecting these events. In other words, we checked the likelihood that the NP-SPEI had the same central tendencies (indicated by high correlation) and detected the same drought and wet events as the already established SPEI. Figure 6.13 shows the results when the indices were calibrated on the full period and figure A.6.4.1, in the appendix, when it was calibrated on the reference period 1961-1990.



Figure 6.13: Pearson Correlation coefficient, upper and lower tail dependence of SPEI and NP-SPEI, when the indices are calibrated on the full period. Pearson correlation values are unitless, while tail dependence is expressed in probability. Since both indices have results between 0 and 1, they are plotted with the same scale.

The NP-SPEI and SPEI calibrated over the entire time period (figure 6.13) resulted to be highly correlated and tail-dependent over the entire globe. For the versions calibrated to the reference period, this phenomenon was somewhat attenuated but still very strong, with values mostly greater than 0.8. For the reference period, the sample size was smaller (30 years) than that for the full period, resulting in more uncertain statistics that could explain



the observed difference. However, the NP-SPEI was found to be an adequate drought index because it captured (almost) the same phenomenon as the SPEI.

Finally, we examined whether the NP-SPEI could be considered a superior index to the SPEI. To answer this question in a statistical manner, we checked whether the NP-SPEI reproduced values that could be better matched by a standard normal distribution than the SPEI (Vicente-Serrano et al., 2010; Beguería et al., 2014). For this purpose, we computed the Cramer-von-Mises (CVM) statistics with the standard normal distribution as a reference for the obtained samples from the SPEI and NP-SPEI. The statistic decreased if the mapping to the standard normal distribution was better preserved. At this stage, we performed this analysis only for the full period, as indices calculated on the reference period were expected to be standard normally distributed only on the reference period itself, and it was not clear which distribution function should have been used as a reference check for the full period. We focused on the latter for future applications.



Figure 6.14: Difference of Cramer-von-Mises Statistics for the SPEI and NP-SPEI. Positive values indicate that the Cramer-von-Mises statistic is smaller for the NP-SPEI, suggesting a better mapping to the standard normal distribution.

As shown in Figure 6.14, the NP-SPEI produced overall lower CVM statistics for all scales, hence a better fit. This phenomenon increased with the increasing scale of the SPEIs. This result suggests that the NP-SPEI is superior to the SPEI. Furthermore, our approach can be used for any climate variable to obtain a non seasonal and normalised climate index, as the



distribution function is estimated nonparametrically. On the other hand, the classical SPEI shows even worse performance.

This clearly proved the added value of our proposed method, and demonstrated that vine copulas can be used to extend the approach to multiple climate variables. For instance, one can construct a joint heatwave and drought index by estimating the distribution function nonparametrically and then estimating the joint distribution function with a vine copula. However, an assumption underlying the methods (for both SPEI versions) is that the distribution is stationary over time. Recently, there have been approaches to consider non-stationary reference distribution functions (Masanta and Srinivas, 2022) and we will focus on this in our future studies.

6.5.2.2 Clustering of droughts and heatwave

To identify the areas of interest, the objective grouping of the drought- and heatwave-related indices was performed by clustering. A time series-based clustering approach based on dynamic time warping was implemented to allow centroids to vary in time. Owing to the presence of climate change, this is a more reasonable assumption compared to methods that use fixed centroids such as k-means or k-medoids (e.g., Hastie et al., 2017). Additionally, as HMD and SPEI are highly correlated, we adopted a multivariate clustering approach that processed the information of the datasets more efficiently. Clustering was performed on a seasonal basis to consider the different atmospheric circulation patterns or seasonal characteristics of the two phenomena. Figure 6.15 shows the first results using the SDTW (Chapter 6.4.1) with ten clusters.



Figure 6.15: Obtained clusters from the multivariate clustering of heatwaves and droughts based on soft-dynamic time warping for k=10 clusters.

The number of clusters was chosen using the Silhouette Coefficient (Rousseeuw, 1987), which works well in a wide variety of clustering setups, particularly in high dimensions (Arbelaitz et al., 2013). Figure 6.16 shows the (bivariate) centroid of one example cluster (six). The centroids were found to represent the peaks between heatwaves (red) and droughts (blue),



indicating that multivariate clustering is well-suited to capture the covariability between these two components.



Figure 6.16: Bivariate centroids of cluster six from the soft-dynamic time warping (figure 6.15) approach.

6.5.2.3 Deep Learning based interpreter of J-functions

The training of the J-function-based classifier was performed for the full ensemble of Jfunctions (see, Toreti et al., 2019b) to better account for the uncertainty, such that the problem was comparable to a multivariate time series classification problem. We solved this problem by using a deep neural network based on convolutional operations and gated recurrent units. Figure A.6.4.1 and Table 6.3 present preliminary results for the network. The test and validation accuracies were found to be very similar (see A.6.4.1), and from Table 6.3 the performance on the test set is approximately 97.4%. Thus test, validation, and testing sets performed with very similar performance and appeared to not overfit the problem (Chollet et al., 2022). Table 6.3 further reveals that the network showed very favourable performance for all different types of dependence structures, with the accuracy of inhibition (Matterntype-II process) being the highest (99.7%), clustering (Log-Gaussian Cox process) in the middle (98.5%), and independence (inhomogeneous Poisson process) the lowest (94.5%). These simulations were performed using deterministic intensity and mean functions. Future applications will focus on the implementation of stochastic versions to generalise the datagenerating process. Furthermore, we explore the possibility of including the simulated point process in addition to the estimation J-Function ensemble to potentially further enhance the performance of the neural network as additional information is included.

Independence	Clustering	Inhibition
0.945	0.054	0.001
0.015	0.985	0.000
0.003	0.000	0.997

Table	6.3:	Test	performance	of deep	neural	network.
, abic	0.0.	1000	perjormanee	oj accp	neuru	



6.5.2.4 Example application

To demonstrate the detection of large-scale heatwaves and droughts, the regions in northeastern USA (USA-NE; 263°–285° E, 36°–52° N), southeastern USA (USA-SE; 263°–285° E, 26°–36° N), and Central Europe (3°–20° E, 46°–56° N) were chosen following the Köppen-Geiger climate classification by Beck et al., (2018). We performed the analysis for the summer (June, July, and August). SPEI and HMD indices were computed for each of these regions, and we performed first grid point-wise drought (SPEI < -1) and heatwave detection (HMD exceeds the 90th percentile). A large-scale heatwave event or drought event was then defined when at least 20% of grid points fulfil these conditions for the stated regions of interest above. For these large-scale events, the J-functions were calculated and classified using the neural network described above. The results are shown in Figure 6.17. The concurrent extreme events can be seen as the type of events for which clustering of events is identified, such as heatwaves in the USA-NE and Central Europe (Figure 6.17, top row, right column).

The case study demonstrated that the analysis of concurrent extreme events can be performed in a fast and automatized way because of the developed J-function tool, while taking into account the non-stationarity of the climate imposed by climate change.



Figure 6.17: Results of the J-Function based interpreter applied for the three regions of interest.



6.6 Summary and outlook

The analysis showed that machine learning can provide added value in analysing compound events and concurrent extremes. AI can be used in recognising relationships between largescale heatwaves and droughts by improving known non-stationary statistical methods from point process theory. Kernel-based methods were able to detect physically reliable large-scale drivers and local climate patterns that dominate winter wheat variability in Europe's largest winter wheat producer, France. In addition, RFs were used to derive definitions for multiple climate events leading to socio-economic impact assessments based on more objective climatic thresholds. Moreover, RFs were able to efficiently predict the desired climate events at local level. Finally, the use of multivariate methods improved the performance of clustering algorithms used to identify homogeneous drought and heatwave regions both at the global level and in the agroclimatic sub-regions in France, demonstrating the added value of multivariate analysis.

Future applications will focus on improving the derived methods and better understanding their physical causes. For example, the J-function based interpreter will be improved by implementing a second pipeline that processes both the point process and the J-function, allowing to fully utilise the available data. We will also analyse how these dependencies relate to teleconnections such as NAO and ENSO and will identify the joint modulators of the dependencies using causal methods (WP4). In addition, non-stationary distribution functions will be used to develop the non-parametric climate indices, for which methods from AI can potentially be adopted.

The analysis of relatively wet and warm winters will be expanded to larger regions of Europe. The RF-based approach to derive objective boundaries will be extended to extract decision boundaries deeper in the individual classification trees. Causal algorithms developed within WP4 will be used to further identify the interdependencies of these phenomena as they are currently built with covariances that cannot distinguish indirect and direct relationships. Finally, further types of compound events will be analysed. For instance, the analysis of dry winters followed by hot summers will focus on representative case studies for recordbreaking summers in Europe and their impact on agriculture and energy.



7 GENERAL SUMMARY AND OUTLOOK

This deliverable reports the preliminary results in the improvement of extreme events detection with ML algorithms.

The study has focused on different categories of extreme events:

- Tropical cyclones: 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)

Each part has focused on pre-existing detection indices (e.g. GPI, SPI) or methods evaluating their scores and capabilities to identify the corresponding extreme events. Generally, these were taken as a starting point for the research. However, in some cases, specifically when evaluating the prediction skill, the goal was to outperform dynamical forecasts (for example, ECMWF).

The employed ML algorithms (developed in WP2) are diverse and have been chosen according to the peculiarities of each problem. In some cases, they were used to improve the definition of existing indices (e.g. EN-GPI for TC) or to identify thresholds which are useful to detect the event themselves (e.g. RF for compound events). In other cases (e.g. ET, HW drought detection) ML algorithms were employed to highlight the drivers to consider for the implementation of data driven detection systems, opening discussion on the physical meaning of such selection. Finally, ML algorithms were trained to perform predictions of the events (short term TC prediction).

In all chapters the results were assessed with scores, evaluating eventual improvements compared to the state of the art and identifying where further work can be performed to increase the advantages of ML-enhances extreme event detection.

The implementation of ML prediction systems, together with the testing of the methods on different spatial scales (either locally or regionally, according to the needs) and with different ML settings, will be the subject of the next studies and upcoming Deliverables.



REFERENCES

AghaKouchak, A., Huning, L.S., Sadegh, M. et al. Toward impact-based monitoring of drought and its cascading hazards. Nat Rev Earth Environ 4, 582–595 (2023). https://doi.org/10.1038/s43017-023-00457-2

Aghabozorgi, S., Seyed Shirkhorshidi, A. and Ying Wah, T. (2015), "Time-series clustering – A decade review", Information Systems, Vol. 53, pp. 16–38.

Allstadt, A.J., Vavrus, S.J., Heglund, P.J., Pidgeon, A.M., Thogmartin, W.E. and Radeloff, V.C. (2015), "Spring plant phenology and false springs in the conterminous US during the 21st century", *Environmental Research Letters*, Vol. 10 No. 10, p. 104008.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J.M. and Perona, I. (2013), "An extensive comparative study of cluster validity indices", *Pattern Recognition*, Vol. 46 No. 1, pp. 243–256

Ascenso, G., Cavicchia, L., Scoccimarro, E., & Castelletti, A. (2023). Optimisation-based refinement of genesis indices for tropical cyclones. Environmental Research Communications, 5(2), 021001.

Ault, T.R., Henebry, G.M., Beurs, K.M. de, Schwartz, M.D., Betancourt, J.L. and Moore, D. (2013), "The False Spring of 2012, Earliest in North American Record", *Eos, Transactions American Geophysical Union*, Vol. 94 No. 20, pp. 181–182.

Baddeley, A., Rubak, E. and Turner, R. (2016), *Spatial point patterns: Methodology and applications with R, Chapman & Hall / CRC Interdisciplinary Statistics*, CRC Press Taylor & Francis Group, Boka Raton, London, New York.

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, doi: 10.1029/2020JD033924

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

Barriopedro, D., García–Herrera, R., Ordóñez, C., Miralles, D. G., & Salcedo–Sanz, S. (2023). Heat waves: Physical understanding and scientific challenges. *Reviews of Geophysics*, e2022RG000780.

Barua, S., Islam, M.M., Yao, X. and Murase, K. (2014), "MWMOTE--Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning", IEEE Transactions on Knowledge and Data Engineering, Vol. 26 No. 2, pp. 405–425.

Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A. and Wood, E.F. (2018), "Present and future Köppen-Geiger climate classification maps at 1-km resolution", *Scientific Data*, Vol. 5 No. 1, p. 180214.



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, doi: 10.1002/joc.3887

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing", *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 57 No. 1, pp. 289–300.

Berg, P., Donnelly, C., and Gustafsson, D. (2018), Near-real-time adjusted reanalysis forcing data for hydrology, Hydrology and Earth System Sciences, 22, 989–1000, doi: 10.5194/hess-22-989-2018.

Bister, M., & Emanuel, K. A. (1998). Dissipative heating and hurricane intensity. *Meteorology* and Atmospheric Physics, 65(3), 233-240, doi: 10.1007/BF01030791

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, doi: 10.1175/1520-0493(1995)123<3268:PTCRIA>2.0.CO;2.

Bueh, C. and Nakamura, H. (2007), "Scandinavian pattern and its climatic impact", *Quarterly Journal of the Royal Meteorological Society*, Vol. 133 No. 629, pp. 2117–2131.

Cai, W., McPhaden, M.J., Grimm, A.M., Rodrigues, R.R., Taschetto, A.S., Garreaud, R.D., Dewitte, B., Poveda, G., Ham, Y.-G., Santoso, A., Ng, B., Anderson, W., Wang, G., Geng, T., Jo, H.-S., Marengo, J.A., Alves, L.M., Osman, M., Li, S., Wu, L., Karamperidou, C., Takahashi, K. and Vera, C. (2020). Climate impacts of the El Niño–Southern Oscillation on South America. *Nature Reviews Earth & Environment*, *1*(4), 215-231, doi: 10.1038/s43017-020-0040-3

Cassou, C., Terray, L., & Phillips, A. S. (2005). Tropical Atlantic influence on European heat waves. *Journal of climate*, *18*(15), 2805-2811.

Cavicchia, L., Scoccimarro, E., Ascenso, G., Castelletti, A., Giuliani, M., & Gualdi, S. (2023). Tropical Cyclone Genesis Potential Indices in a New High-Resolution Climate Models Ensemble: Limitations and Way Forward. *Geophysical Research Letters*, 50(11), e2023GL103001.

Ceglar, A., Toreti, A., Lecerf, R., van der Velde, M. and Dentener, F. (2016), "Impact of meteorological drivers on regional inter-annual crop yield variability in France", *Agricultural and Forest Meteorology*, Vol. 216, pp. 58–67.

Ceglar, A., Zampieri, M., Toreti, A. and Dentener, F. (2019), "Observed Northward Migration of Agro-Climate Zones in Europe Will Further Accelerate Under Climate Change", *Earth's Future*, Vol. 7 No. 9, pp. 1088–1101.

Ceglar, A., van der Wijngaart, R., Wit, A. de, Lecerf, R., Boogaard, H., Seguini, L., van den Berg, M., Toreti, A., Zampieri, M., Fumagalli, D. and Baruth, B. (2019). Improving WOFOST model to



simulate winter wheat phenology in Europe: Evaluation and effects on yield. *Agricultural Systems*, *168*, 168-180, doi: 10.1016/j.agsy.2018.05.002

Chamberlain, C.J., Cook, B.I., García de Cortázar-Atauri, I. and Wolkovich, E.M. (2019), "Rethinking false spring risk", *Global Change Biology*, Vol. 25 No. 7, pp. 2209–2220.

Chollet, F., Kalinowski, T. and Allaire, J.J. (2022), Deep learning with R, Second edition, Manning Publications Co, Shelter Island, NY.

Loader, C.R. (1996), "Local likelihood density estimation", *The Annals of Statistics*, Vol. 24 No. 4, pp. 1602–1618., S. (2004), An introduction to statistical modeling of extreme values, Springer Series in Statistics, 4. printing, Springer, London, Berlin, Heidelberg.

Chiang, F., Mazdiyasni, O. & AghaKouchak, A. Evidence of anthropogenic impacts on global drought frequency, duration, and intensity. Nat Commun 12, 2754 (2021). https://doi.org/10.1038/s41467-021-22314-w

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). AJ-function for inhomogeneous spatio-temporal point processes. *Scandinavian Journal of Statistics*, *42*(2), 562-57, doi: 10.1111/sjos.12123

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, doi: 10.1007/s10463-015-0515-z

Cronie, O. and van Lieshout, M.N.M. (2018), "A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions", *Biometrika*, Vol. 105 No. 2, pp. 455–462

Cuturi, M. and Blondel, M. (2017), Soft-DTW: a Differentiable Loss Function for Time-Series.

Czado, C., & Nagler, T. (2022). Vine copula based modeling. *Annual Review of Statistics and Its Application*, *9*(1), 453-477, doi:

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

Deb, K., Pratap, A., Agarwal S. and Meyarivan T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II IEEE Trans. Evol. Comput. 6 182–97

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



Diggle, P.J. (2014), Statistical analysis of spatial and spatio-temporal point patterns, Monographs on statistics and applied probability, Vol. 128, 3. ed., CRC Press, Boca Raton, Fla.

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.

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, doi: 10.3390/atmos9010033

Droogers, P. and Allen, R.G. (2002), "Estimating Reference Evapotranspiration Under Inaccurate Data Conditions", *Irrigation and Drainage Systems*, Vol. 16 No. 1, pp. 33–45.

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. doi:10.1088/1748-9326/11/7/074004

Efron, B. and T. Hastie, 2016: Computer age statistical inference. Algorithms, evidence, and data science, Cambridge University Press.

Emanuel, K., & Nolan, D. S. (2004, July). Tropical cyclone activity and the global climate system. In *26th conference on hurricanes and tropical meteorology*.

Emanuel, K. A. (1988). The maximum intensity of hurricanes. J. Atmos. Sci, 45(7), 1143-1155.

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

Feng, Y., Gries, T. and Fritz, M. (2020), "Data-driven local polynomial for the trend and its derivatives in economic time series", *Journal of Nonparametric Statistics*, Vol. 32 No. 2, pp. 510–533.

Fischer, E. M., & Knutti, R. (2013). Robust projections of combined humidity and temperature extremes. *Nature Climate Change*, *3*(2), 126-130, doi: 10.1038/nclimate1682

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

Garali, I., Adanyeguh, I.M., Ichou, F., Perlbarg, V., Seyer, A., Colsch, B., Moszer, I., Guillemot, V., Durr, A., Mochel, F. and Tenenhaus, A. (2018), "A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia", *Briefings in bioinformatics*, Vol. 19 No. 6, pp. 1356–1369.

García-Herrera, R., Díaz, J., Trigo, R. M., & Hernández, E. (2005, February). Extreme summer temperatures in Iberia: health impacts and associated synoptic conditions. In *Annales Geophysicae* (Vol. 23, No. 2, pp. 239-251). Göttingen, Germany: Copernicus Publications.

García-Garizábal, I., Causapé, J., Abrahao, R. and Merchan, D. (2014), "Impact of Climate Change on Mediterranean Irrigation Demand: Historical Dynamics of Climate and Future Projections", *Water Resources Management*, Vol. 28 No. 5, pp. 1449–1462.

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*, Vol. 109 No. 505, pp. 346–358.

Geenens, G. and Wang, C. (2018), "Local-Likelihood Transformation Kernel Density Estimation for Positive Random Variables", *Journal of Computational and Graphical Statistics*, Vol. 27 No. 4, pp. 822–835.

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

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.

González, J. A., Rodríguez-Cortés, F. J., Cronie, O., & Mateu, J. (2016). Spatio-temporal point process statistics: a review. *Spatial Statistics*, *18*, 505-544, doi: 10.1016/j.spasta.2016.10.002

D. Hadka and P. Reed. Borg: An auto-adaptive many-objective evolutionary computing framework. *Evolutionary computation*, 21(2):231–259, 2013.

Hannachi, A. (2021), Patterns Identification and Data Mining in Weather and Climate, Springer Atmospheric Sciences Ser, Springer International Publishing AG, Cham.



Hao, Z., Hao, F., Singh, V. P., & Zhang, X. (2019a). Statistical prediction of the severity of compound dry-hot events based on El Niño-Southern Oscillation. *Journal of Hydrology*, *572*, 243-250, doi: 10.1016/j.jhydrol.2019.03.001

Hao, Z., Hao, F., Xia, Y., Singh, V.P. and Zhang, X. (2019b), "A monitoring and prediction system for compound dry and hot events", *Environmental Research Letters*, Vol. 14 No. 11, p. 114034.

Hao, Z., Hao, F., Xia, Y., Feng, S., Sun, C., Zhang, X., Fu, Y., Hao, Y., Zhang, Y. and Meng, Y. (2022), "Compound droughts and hot extremes: Characteristics, drivers, changes, and impacts", *Earth-Science Reviews*, Vol. 235, p. 104241.

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.

Hastie, T., Tibshirani, R. and Friedman, J.H. (2017), *The elements of statistical learning: Data mining, inference, and prediction, Springer Series in Statistics,* Second edition, Springer, New York, NY.

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., ... Thépaut, J. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049.doi: 10.1002/qj.3803

Hoerling, M., Kumar, A., Dole, R., Nielsen-Gammon, J.W., Eischeid, J., Perlwitz, J., Quan, X.-W., Zhang, T., Pegion, P. and Chen, M. (2013), "Anatomy of an Extreme Event", *Journal of Climate*, Vol. 26 No. 9, pp. 2811–2832

G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

Huang, B., Liu, C., Banzon, V., Freeman, E., Graham, G., Hankins, B., Smith, T. and Zhang, H.-M. (2021), "Improvements of the Daily Optimum Interpolation Sea Surface Temperature (DOISST) Version 2.1", *Journal of Climate*, Vol. 34 No. 8, pp. 2923–2939.

Hundecha, Y., Arheimer, B., Donnelly, C., and Pechlivanidis, I. (2016), A regional parameter estimation scheme for a pan-European multi-basin model, *Journal of Hydrology Regional Studies*, 6, 90–111, doi: 10.1016/j.ejrh.2016.04.002

Ionita, M., Caldarescu, D. E., & Nagavciuc, V. (2021). Compound Hot and Dry Events in Europe: Variability and Large-Scale Drivers. *Frontiers in Climate*, *3*, 688991, doi: 10.3389/fclim.2021.688991



IPCC (2021), Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, In Press, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Ishwaran, H., Kogalur, U.B., Gorodeski, E.Z., Minn, A.J. and Lauer, M.S. (2010), "High-Dimensional Variable Selection for Survival Data", Journal of the American Statistical Association, Vol. 105 No. 489, pp. 205–217.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021), An introduction to statistical learning: With applications in R, Springer texts in statistics, Second edition, Springer, New York.

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. and 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, doi: 10.5194/gmd-10-4005-2017

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.

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.

Kashinath, K., and Coauthors, 2021: Physics-informed machine learning: case studies for weather and climate modelling. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 379, 20200093, doi: 10.1098/rsta.2020.0093.

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. and Woollings, T. (2022), "Atmospheric blocking and weather extremes over the Euro-Atlantic sector – a review", *Weather and Climate Dynamics*, Vol. 3 No. 1, pp. 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, doi: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, doi: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.

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, doi:10.1175/2009BAMS2755.1.

Knapp, K. R., Diamond, H. J., Kossin, J. P., Kruk, M. C., & Schreck, C. J. (2018). International Best Track Archive for Climate Stewardship (IBTrACS) Project, Version 4. NOAA National Centers for Environmental Information, doi:10.25921/82ty-9e16.

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*, Vol. 110, pp. 1–18.

Kubat, M., Holte, R. and Matwin, S. (1997), "Learning when negative examples abound", in Someren, M.v. and Widmer, G. (Eds.), *Machine Learning: ECML'97: 9th European Conference on Machine Learning, Prague, Czech Republic, April 23 - 25, 1997, Proceedings, Springer Berlin Heidelberg; Imprint: Springer, Berlin, Heidelberg, pp. 146–153.*

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., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*, 2278–2324.

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.



Lesk, C., Rowhani, P., & Ramankutty, N. (2016). Influence of extreme weather disasters on global crop production. Nature, 529(7584), 84–87. doi:10.1038/nature16467

Loader, C.R. (1996), "Local likelihood density estimation", *The Annals of Statistics*, Vol. 24 No. 4, pp. 1602–1618.

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

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.

Mahlstein I, Spirig C, Liniger MA, Appenzeller C (2015). Estimating daily climatologies for climate indices derived from climate model data and observations. J Geophys Res Atmos 120(7):2808–2818. <u>https://doi.org/10.1002/2014JD022327</u>.

Mahony, C.R. and Cannon, A.J. (2018), "Wetter summers can intensify departures from natural variability in a warming climate", *Nature Communications*, Vol. 9 No. 1, p. 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

Majumder, S., Goes, M., Polito, P. S., Lumpkin, R., Schmid, C., & Lopez, H. (2019). Propagating modes of variability and their impact on the western boundary current in the South Atlantic. Journal of Geophysical Research: Oceans, 124(5), 3168-3185.



Masanta S.K. and Srinivas, V.V. (2022), "Proposal and evaluation of nonstationary versions of SPEI and SDDI based on climate covariates for regional drought analysis", *Journal of Hydrology*, Vol. 610, p. 127808.

Massmann, A., Gentine, P. and Lin, C. (2019), "When Does Vapor Pressure Deficit Drive or Reduce Evapotranspiration?", *Journal of advances in modeling earth systems*, Vol. 11 No. 10, pp. 3305–3320.

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

McGovern, A., Lagerquist, R., John Gagne, D., Jergensen, G.E., Elmore, K.L., Homeyer, C.R. and Smith, T. (2019), "Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning", Bulletin of the American Meteorological Society, Vol. 100 No. 11, pp. 2175–2199.

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.

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, doi: 10.1016/j.gloplacha.2022.103773

Menkes, C. E., Lengaigne, M., Marchesiello, P., Jourdain, N. C., Vincent, E. M., Lefèvre, J., et al. (2012). Comparison of tropical cyclogenesis indices on seasonal to interannual timescales. Climate Dynamics, 38(1), 301–321. https://doi.org/10.1007/s00382-011-1126-x

Miralles, D. G., Gentine, P., Seneviratne, S. I., & Teuling, A. J. (2019). Land–atmospheric feedbacks during droughts and heat waves: state of the science and current challenges. Annals of the New York Academy of Sciences, 1436(1), 19–35. doi:10.1111/nyas.13912

Moradi, M. M., Cronie, O., Rubak, E., Lachieze-Rey, R., Mateu, J., & Baddeley, A. (2019). Resample-smoothing of Voronoi intensity estimators. *Statistics and computing*, *29*(5), 995-1010, doi: 10.1007/s11222-018-09850-0



Mukherjee, S., Ashfaq, M., & Mishra, A. K. (2020). Compound drought and heatwaves at a global scale: The role of natural climate variability-associated synoptic patterns and land-surface energy budget anomalies. *Journal of Geophysical Research: Atmospheres*, 125(11), e2019JD031943.

Nagler, T. (2018a), "A generic approach to nonparametric function estimation with mixed data", *Statistics & Probability Letters*, Vol. 137, pp. 326–330.

Nagler, T. (2018b), "Asymptotic Analysis of the Jittering Kernel Density Estimator", *Mathematical Methods of Statistics*, Vol. 27 No. 1, pp. 32–46.

Nguyen, P.-L., Min, S.-K. and Kim, Y.-H. (2021), "Combined impacts of the El Niño-Southern Oscillation and Pacific Decadal Oscillation on global droughts assessed using the standardized precipitation evapotranspiration index", *International Journal of Climatology*, Vol. 41 S1, E1645-E1662.

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

Pechlivanidis, I., Crochemore, L., Rosberg, J. and Bosshard, T.. What are the key drivers controlling the quality of seasonal streamflow forecasts? *Water Resources Research*, 56 (6):e2019WR026987, 2020.

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.

Perkins, S.E. and Alexander, L.V. (2013), "On the Measurement of Heat Waves", Journal of Climate, Vol. 26 No. 13, pp. 4500–4517.

Perkins-Kirkpatrick, S. E., & Lewis, S. C. (2020). Increasing trends in regional heatwaves. *Nature Communications*, 11(1). doi:10.1038/s41467-020-16970-7

Philipp, A., Della-Marta, P.M., Jacobeit, J., Fereday, D.R., Jones, P. D., Moberg, A. and Wanner, H. (2007). Long-term variability of daily North Atlantic-European pressure patterns since 1850 classified by simulated annealing clustering. Journal of Climate, 20(16), 4065–4095. https://doi.org/10.1175/JCLI4175.1.

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



Rahimi A. and Recht B. (2007), "Random Features for Large-Scale Kernel Machines", Neural Information Processing Systems.

Ramsay, H. A., & Sobel, A. H. (2011). Effects of relative and absolute sea surface temperature on tropical cyclone potential intensity using a single-column model. Journal of Climate, 24(1), 183–193. https://doi.org/10.1175/2010jcli3690.1

Raymond, C., Matthews, T., Horton, R.M., Fischer, E.M., Fueglistaler, S., Ivanovich, C., Suarez-Gutierrez, L. and Zhang, Y. (2021), "On the Controlling Factors for Globally Extreme Humid Heat", *Geophysical Research Letters*, Vol. 48 No. 23.

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

Rodrigues, R. R., Taschetto, A. S., Sen Gupta, A., & Foltz, G. R. (2019). Common cause for severe droughts in South America and marine heatwaves in the South Atlantic. *Nature Geoscience*, *12*(8), 620-626.

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*, (p. 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., and 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

Rousseeuw, P.J. (1987), "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics*, Vol. 20, pp. 53–65.

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.

Sardá-Espinosa, A. (2019), "Time-Series Clustering in R Using the dtwclust Package", The R Journal, Vol. 11, p. 22.



Schäfler, A., Craig, G., Wernli, H., Arbogast, P., Doyle, J.D., McTaggart-Cowan, R., Methven, J., Rivière, G., Ament, F., Boettcher, M. and 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

Schallhorn, N., Kraus, D., Nagler, T. and Czado, C. (2017), *D-vine quantile regression with discrete variables*.

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.

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

Scholkopf B., Mika S., Burges C., Knirsch P., Müller K, Rätsch G. and Smola A. (1999), "Input space versus feature space in kernel-based methods", *IEEE transactions on neural networks*.

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

Schultz, M. G., C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L. H. Leufen, A. Mozaffari, S. Stadtler, 2021: Can deep learning beat numerical weather prediction? Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 379, 20200097

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

Serifi, Agon & Günther, Tobias & Ban, Nikolina. (2021). Spatio-Temporal Downscaling of Climate Data Using Convolutional and Error-Predicting Neural Networks. Frontiers in Climate.

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

Stagge, J.H., Tallaksen, L.M., Gudmundsson, L., van Loon, A.F. and Stahl, K. (2015), "Candidate Distributions for Climatological Drought Indices (SPI and SPEI)", International Journal of Climatology, Vol. 35 No. 13, pp. 4027–4040.



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

Steptoe, H., Jones, S.E.O. and Fox, H. (2018), "Correlations Between Extreme Atmospheric Hazards and Global Teleconnections: Implications for Multihazard Resilience", *Reviews of Geophysics*, Vol. 56 No. 1, pp. 50–78.

Sutherland, D.J. and Schneider, J. (2015), On the Error of Random Fourier Features.

Tenenhaus, A. and Tenenhaus, M. (2011), "Regularized Generalized Canonical Correlation Analysis", *Psychometrika*, Vol. 76 No. 2, pp. 257–284.

Tenenhaus, A., Philippe, C. and Frouin, V. (2015), "Kernel Generalized Canonical Correlation Analysis", Computational Statistics & Data Analysis, Vol. 90, pp. 114–131.

Tenenhaus, M., Tenenhaus, A. and Groenen, P.J.F. (2017), "Regularized Generalized Canonical Correlation Analysis: A Framework for Sequential Multiblock Component Methods", *Psychometrika*, Vol. 82 No. 3, pp. 737–777.

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

Tippett, M. K., Camargo, S. J., & Sobel, A. H. (2011). A Poisson regression index for tropical cyclone genesis and the role of large-scale vorticity in genesis. *Journal of Climate*, *24*(9), 2335-2357, doi: 10.1175/2010JCLI3811.1

Toreti, A., Belward, A., Perez-Dominguez, I., Naumann, G., Luterbacher, J., Cronie, O., Seguini, L., Manfron, G., Lopez-Lozano, R., Baruth, B., Berg, M., Dentener, F., Ceglar, A., Chatzopoulos, T. and Zampieri, M. (2019a), The exceptional 2018 European water seesaw calls for action on adaptation. *Earth's Future*, *7*(6), 652-663, doi: 10.1029/2019EF001170

Toreti, A., Cronie, O., & Zampieri, M. (2019b). Concurrent climate extremes in the key wheat producing regions of the world. *Scientific reports*, *9*(1), 1-8, doi: 10.1038/s41598-019-41932-5

Torralba, V., Materia, S., Cavicchia, L., Alvarez-Castro, C., Prodhomme, C., McAdam, R., Scoccimarro, E., Gualdi., S. Nighttime heat waves in the Euro-Mediterranean region: definition, characterisation, and seasonal prediction. In review.

Van Lieshout, M. N. M. (2011). AJ-function for inhomogeneous point processes. *Statistica Neerlandica*, *65*(2), 183-201, doi: 10.1111/j.1467-9574.2011.00482.x

Vecchi, G. A., & Soden, B. J. (2007). Global warming and the weakening of the tropical circulation. Journal of Climate, 20(17), 4316–4340. https://doi.org/10.1175/jcli4258.1


Vicente-Serrano, S.M., Beguería, S., and López-Moreno, J.I. (2010), A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index, *Journal of Climate* 23 (7), 1696–1718, doi: 10.1175/2009JCLI2909.1

Vicente-Serrano, S.M. and Beguería, S. (2016), "Comment on 'Candidate distributions for climatological drought indices (SPI and SPEI)' by James H. Stagge et al", *International Journal of Climatology*, Vol. 36 No. 4, pp. 2120–2131.

Vicente-Serrano, S.M., Zabalza-Martínez, J., Borràs, G., López-Moreno, J.I., Pla, E., Pascual, D., Savé, R., Biel, C., Funes, I., Martín-Hernández, N., Peña-Gallardo, M., Beguería, S. and Tomas-Burguera, M. (2017), "Effect of reservoirs on streamflow and river regimes in a heavily regulated river basin of Northeast Spain", *CATENA*, Vol. 149, pp. 727–741.

Vitart, F., 2009: Impact of the MJO on tropical storms and risk of landfall in the ECMWF forecast system. *Geophysical Research Letters*, 36, L15802, doi:10.1029/2009GL039089

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, doi: 10.1038/nclimate2736

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.

Wang, B., & Murakami, H. (2020). Dynamic genesis potential index for diagnosing presentday and future global tropical cyclone genesis. Envi-ronmental Research Letters, 15(11), 114008. https://doi.org/10.1088/1748-9326/abbb01

D. Wilhite. Chapter 1 drought as a natural hazard: Concepts and definitions drought as a natural hazard: Concepts and definitions. *Drought A Glob. Assessment, Drought Mitig. Cent. Fac. Publ. Pap. 69Available at http://digitalcommons. unl. edu/cgi/viewcontent. cgi,* 2000.

Wilks, D.S. (2006), "On "Field Significance" and the False Discovery Rate", *Journal of Applied Meteorology and Climatology*, Vol. 45 No. 9, pp. 1181–1189.

Wilks, D.S. (2011), *Statistical methods in the atmospheric sciences, International geophysics series*, v. 100, 3rd ed., Academic Press, Oxford, Waltham, MA.

Wulff, C. O., Greatbatch, R. J., Domeisen, D. I., Gollan, G., & Hansen, F. (2017). Tropical forcing of the summer East Atlantic pattern. *Geophysical Research Letters*, 44(21), 11-166, doi:10.1002/2017GL075493

Yang, W., Andréasson, J., Phil Graham, L., Olsson, J., Rosberg, J., and Wetterhall, F. (2010), Distribution-based scaling to improve usability of regional climate model projections for hydrological climate change impacts studies, *Hydrology Research*, 41, 211–229



Zampieri, M., Ceglar, A., Dentener, F., and 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. doi:10.1088/1748-9326/aa723b

Zanchettin, D., Franks, S.W., Traverso, P. and Tomasino, M. (2008), "On ENSO impacts on European wintertime rainfalls and their modulation by the NAO and the Pacific multi-decadal variability described through the PDO index", *International Journal of Climatology*, Vol. 28 No. 8, pp. 995-1006.

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, W., Luo, M., Gao, S., Chen, W., Hari, V., & Khouakhi, A. (2021). Compound hydrometeorological extremes: drivers, mechanisms and methods. *Frontiers in Earth Science*, *9*, 673495, doi: 10.3389/feart.2021.67349

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

Zhu, J., Huang, B., Cash, B., Kinter, J. L., Manganello, J., Barimalala, R., ... & Towers, P. (2015). ENSO prediction in Project Minerva: Sensitivity to atmospheric horizontal resolution and ensemble size. *Journal of Climate*, *28*(5), 2080-2095. doi:10.1175/JCLI-D-14-00302.1.

Zscheischler, J., & Seneviratne, S. I. (2017). Dependence of drivers affects risks associated with compound events. *Science advances*, *3*(6), e1700263, doi: 10.1126/sciadv.1700263

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. and Zhang, X. (2018). Future climate risk from compound events. *Nature Climate Change*, *8*(6), 469-477

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. and Vignotto, E. (2020), A typology of compound weather and climate events. *Nature reviews earth & environment*, *1*(7), 333-347, doi: 10.1038/s43017-020-0060-z

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

In this Appendix, the clusters of predictor variables used for the heatwave driver feature selection are reported.



Figure A4.1: Clusters of European predictor variables for the heatwave driver Feature Selection Framework (Section 4). *K*-means clustering is applied to ERA5 daily data over the period 1951-2010. The domain covers [30N,70N], [-15E,46E]. Values over the ocean are removed for 2m temperature and Soil Moisture.



Figure A4.2: Clusters of North Atlantic predictor variables for the heatwave driver Feature Selection Framework (Section 4). K-means clustering is applied to ERA5 daily data over the period 1951-2010. The domain covers [0N,70N], [90W,46E].





Figure A4.3: Clusters of Arctic Sea Ice Concentration for the heatwave driver Feature Selection Framework (Section 4). Kmeans clustering is applied to ERA5 daily data over the period 1951-2010. The domain covers the northern polar region; parts of the domain which have never experienced sea ice are removed from the clustering.



Figure A4.24 Clusters of global predictor variables for the heatwave driver Feature Selection Framework (Section 4). Kmeans clustering is applied to ERA5 daily data over the period 1951-2010. The domain covers all latitudes and longitudes.



APPENDIX A5

A5.1

In this Appendix, the heatmaps for the Clusters not discussed in the Chapter 5 are reported.



Figure A5.1: Heatmap: Cluster 1 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).





Figure A5.2: Heatmap: Cluster 2 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).



Cluster 3

Figure A5.3: Heatmap: Cluster 3 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).





Figure A5.4: Heatmap: Cluster 4 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).



Cluster 5

Figure A5.5: Heatmap: Cluster 5 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).





Figure A5.6: Heatmap: Cluster 6 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018)



Figure A5.7: Heatmap: Cluster 7 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).





Figure A5.8: Heatmap: Cluster 8 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).









Figure A5.10: Heatmap: Cluster 11 (on the x-axis the 312 months between 1993 and 2018 are reported, e.g. month 100 is May 2001, month 200 is September 2009, month 300 is January 2018).

A5.2

In this Appendix, the selection matrices for the Clusters not shown in the Chapter 5 are reported.







Figure A5.2.2: Selection matrix: Cluster 2.





Figure A5.2.3: Selection matrix: Cluster 3.



Figure A5.2.5: Selection matrix: Cluster 5.



Figure A5.2.4: Selection matrix: Cluster 4.



Figure A5.2.6: Selection matrix: Cluster 6.

62.50% 50.00% 37.50% 12.50% 12.50% 25.00% 50.00% 12.50% 12.50% 25.00% 50.00% 87.50% 12.50% 87.50% 0.00% 25.00% 0.00% 0.00%





Figure A5.2.7: Selection matrix: Cluster 7.



Figure A5.2.9: Selection matrix: Cluster 9. 11.



Figure A5.2.8: Selection matrix: Cluster 8.



Figure A5.2.10: Selection matrix: Cluster



APPENDIX A6

A6.1 Nonparametric SPEI for the wichita time series

As an example for the extrapolation problem described in chapter 6.4.5, figure A.6.1 displays a time series called "wichita" which is used in the demo software of the SPEI. It can be seen that values outside the reference period cannot be extrapolated by the algorithm indicated by the red dots in the figure for which the value infinity is obtained, whereas NP-SPEI is able to map these time points. Furthermore, by visual inspection we observed that the NP-SPEI and SPEI were highly correlated. For instance, the Pearson correlation coefficient of the two time series calculated for the full period was approximately 0.996.



Figure A.6.1: Example calibration of SPEI and NP-SPEI for the wichita of the SPEI-demo package (see text for more details). Upper panels show the calculated SPEI based on the log-logistic distribution, while the lower panels display the calculated SPEI based on the upper-right panel show values which cannot be mapped.



A6.2 Number of non-extrapolatable points of the nonparametric SPEI

Figure A.6.2 shows the number of non-extrapolatable points for the non-parametric SPEI. These points cannot be identified by the eye. Indeed, all maps shown in figure A.6.2 have approximately 31 million grid points, and we have found 1000 grid points, where non-mappable values occur and the latter only when the calibration was performed on the reference period. When the full time period was used, this phenomenon vanished, in contrast to the regular SPEI.



Figure A.6.2: Same as Figure 6.13, but for calibration on the reference period.



A6.3 Comparison of nonparametric SPEI and SPEI on the reference period

Figure A.6.3 compares the nonparametric SPEI to the SPEI, which are both calibrated on the reference period 1961-1990 with the same statistics as in figure 6.13.



Figure A.6.3: Same as Figure 6.15, but for calibration on the reference period.

A6.4 Deep Learning based training.

Figure A.6.4 shows the evaluation plot of the deep neural network used for the AI-based J-Function-based interpreter in Section 6.5.2.3. The accuracy of neural networks converged fast with similar performances to those obtained on the validation set, thus not indicating overfitting (Chollet *et al.*, 2022).



Figure A.6.4: Evaluation plot of the deep neural network used. The blue line represents the accuracy of the test set, while the green line represents the accuracy of the validation set.







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