SCEWERO

STRENGTHENING THE RESEARCH CAPACITIES FOR EXTREME WEATHER EVENTS IN ROMANIA

Data-driven forecast of intense precipitation

Locally tailored drivers and data-driven models for seasonal forecasting of intense precipitation in Europe

Eugenio Fella

eugenio.fella@cmcc.it

UNIVERSITATEA BABEȘ-BOLYAI

cmcc Centro Euro-Mediterraneo sui Cambiamenti Climatici

Universiteit Antwerpen

indeco soft

Funded by the European Union

Grant Agreement 101159497

Outline

1) Context
2) Seasonal prediction
3) Extreme precipitation events
4) Feature selection
5) Data-driven prediction
6) Results

Funded by the European Union
Grant Agreement 101159497

1) Context

Impacts

- Extreme precipitation events are among the most severe **weather hazards** in terms of loss of lives, socioeconomic impacts, and related disasters (one of the most common causes of floods and landslides).
- 2000–2019: **Floods** caused 100k+ deaths and $651B+ in losses worldwide (*United Nations Office for Disaster Risk Reduction, 2020*).



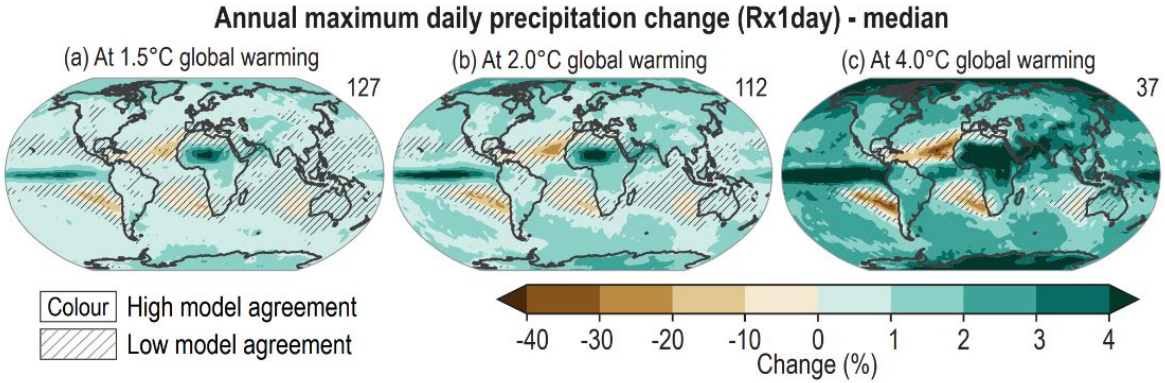
Emilia-Romagna, May 2023

Funded by the European Union
Grant Agreement 101159497

1) Context

Climate change

- **IPCC AR6**: Heavy precipitation expected to increase in intensity and frequency with further global warming, varying by region.
- This is consistent with the higher capacity of warmer air to hold **moisture**.



IPCC WG1, 2021

Funded by the European Union
Grant Agreement 101159497

2) Seasonal prediction

Weather forecast

- Aim at predicting actual specific evolution of the atmosphere at a **given place and time**.
- Support **short-term** decisions.

Seasonal prediction

- It gives probabilities or **aggregated statistics**.
- Help to manage risks and support **early decisions** (e.g., crop choice, water resource management).

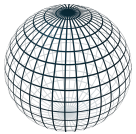


Funded by the European Union
Grant Agreement 101159497

2) Seasonal prediction

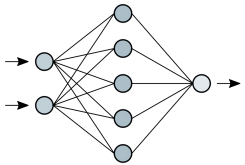
Dynamical predictions

- Prediction are produced by numerically solving the **dynamical and thermodynamic equations** governing the Earth-system components.
- They have long constituted the **state-of-the-art** in the field.



Statistical predictions

- Based on **statistical relationships** derived from observed data, linking predictor variables to the climatic quantity being forecast over specific regions.
- Span a wide range of methodological complexity, from simple regression to more advanced data-driven models (**machine learning** and Artificial Intelligence).



Funded by the European Union
Grant Agreement 101159497

2) Seasonal prediction: dynamical predictions

- A **coupled climate model** represents the atmosphere, ocean, land, and sea-ice components together with their interactions.
- It is **integrated** forward in time, starting from initial conditions that describe the observed state of the whole system.

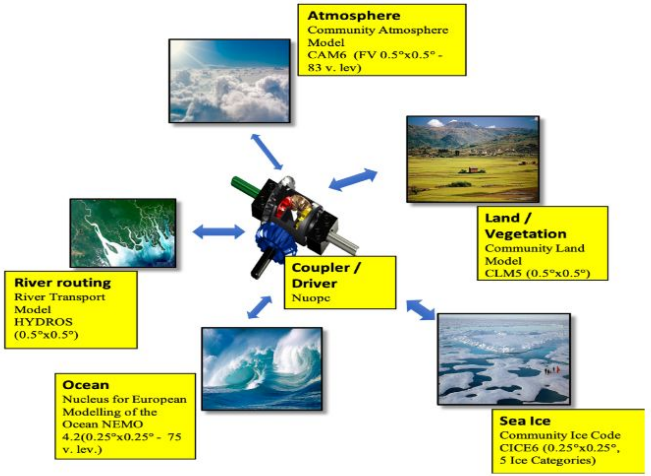


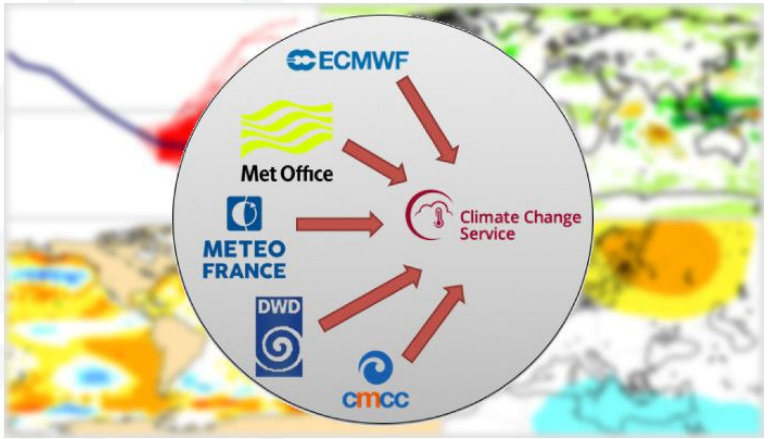
Figure 2. General scheme of the CMCC-CM3 fully coupled Global Model

The new **CMCC Seasonal Prediction System SPS4**, *Sanna et al., 2025*

Funded by the European Union
Grant Agreement 101159497

2) Seasonal prediction: dynamical predictions

- Today, dynamical models are used **operationally** to produce seasonal predictions.
- They are also combined into **multi-model systems** to account for structural uncertainty.
- Each system provides **hindcasts** (re-forecasts of past years) used to calibrate the forecasts and assess their skill.



Copernicus C3S seasonal forecast

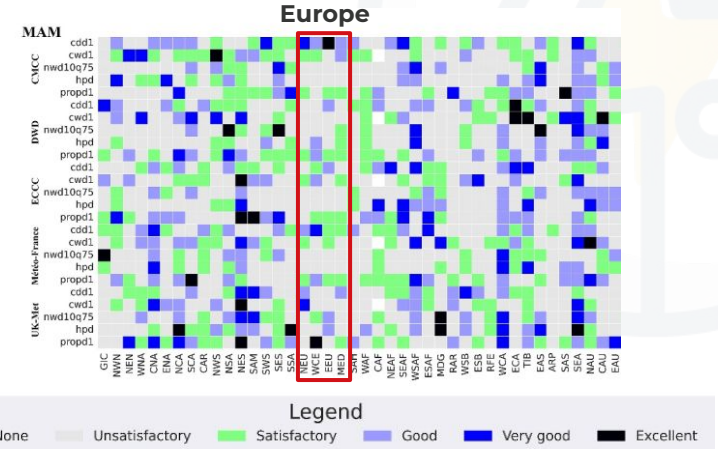
Funded by the European Union
Grant Agreement 101159497

2) Seasonal prediction: dynamical predictions

Issues:

Despite advances in weather and climate systems, skill in predicting **severe precipitation events** remains limited, particularly in extratropical regions (*Nikraftar et al., 2024*).

Computational expense of dynamical forecasting systems.



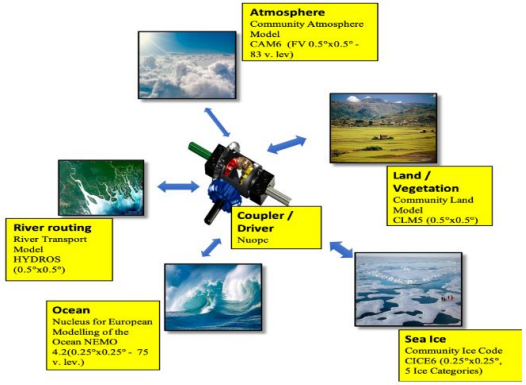
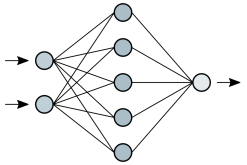


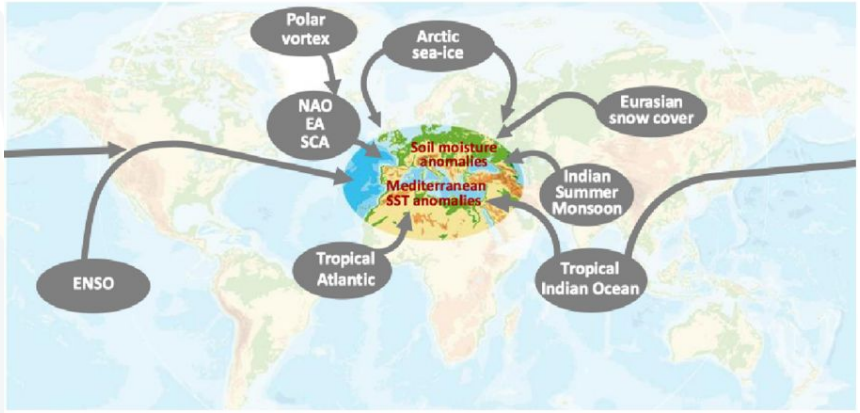
Figure 2. General scheme of the CMCC-CM3 fully coupled Global Model

Funded by the European Union
Grant Agreement 101159497

2) Seasonal prediction: statistical predictions



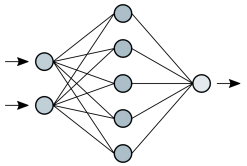
- Advances in **Earth Observation** have generated vast datasets, enabling the growth of Artificial Intelligence (AI) applications in Earth science.
- AI can detect and make explicit the **spatial and temporal patterns** among Earth-system variables, potentially revealing processes not yet represented in dynamical models.
- It does not require complete prior knowledge of the underlying Earth-system processes and dynamics, it learns the relationships directly from data.

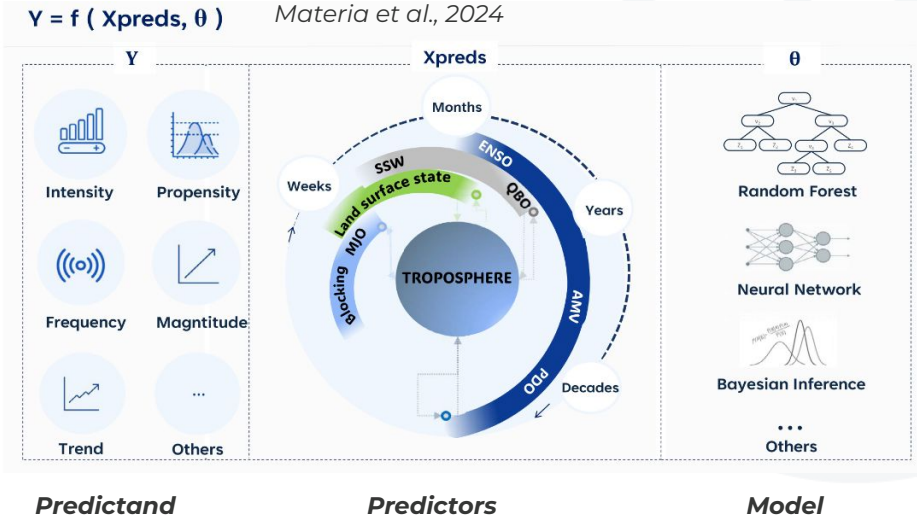


Gualdi et al., 2026

Funded by the European Union
Grant Agreement 101159497

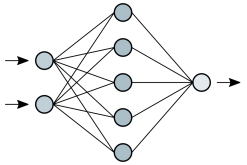
2) Seasonal prediction: statistical predictions





- **Model** *f(Xpreds, θ)*: a parametric function that takes input variables (predictors *Xpreds*) from times before the target (predictand *Y)*, and produces an output further ahead in time.
- **Training**: tune the model's parameters *θ* so it best performs the task (i.e. minimises a loss function measuring forecast error).
- **Inference**: apply the trained model to new inputs to generate predictions.
- Good performance typically requires a large number of training **samples**.

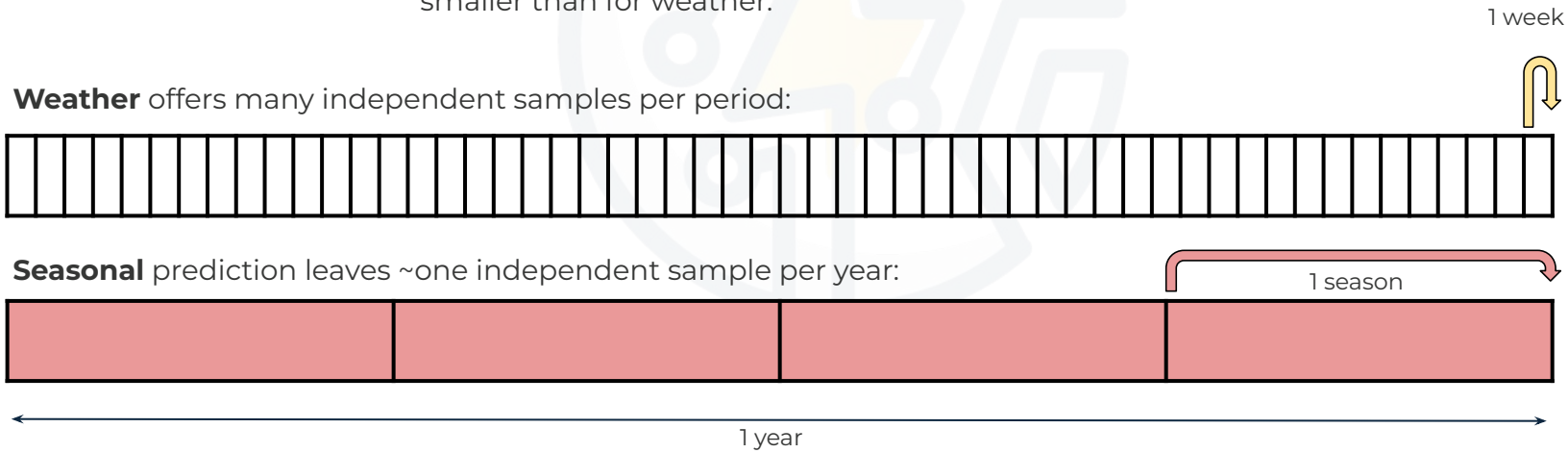
Funded by the European Union
Grant Agreement 101159497



2) Seasonal prediction: statistical predictions

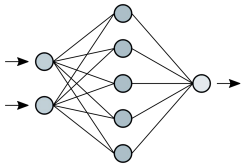
Issues:

- **AI** advances have been striking for weather timescales, but less so on subseasonal-to-decadal timescales.
- One of the main challenges is the limited amount of independent **training data**, up to two orders of magnitude smaller than for weather.



Funded by the European Union
Grant Agreement 101159497

2) Seasonal prediction: statistical predictions



Training on dynamical model simulations

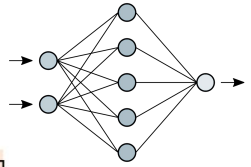
- **Data abundance**: model output provides huge training sets, enabling more advanced deep-learning schemes and the possibility of generalising to future climate conditions.
- **Inherited errors**: dynamical models have biases, especially for extreme precipitation, so the AI risks inheriting those errors.

Training on reanalysis/observational data

- **New insights:** learns from the real observed system, which can reveal new insights into physical processes (including ones not captured by dynamical models).
- **Data scarcity**: limited data availability, forcing the use of Machine-Learning methods suited to small datasets.

Funded by the European Union
Grant Agreement 101159497

2) Seasonal prediction: statistical predictions



Other approaches:

- **Hybrid**: various combinations of dynamical forecasts with AI methods.
- **Autoregressive**: the model learns the atmospheric evolution over a short time step (e.g. 6 hours), then produces longer forecasts by feeding its own predictions back in as inputs. Drawback: error accumulation can cause instabilities or blurring (over-smoothing) at longer lead times.

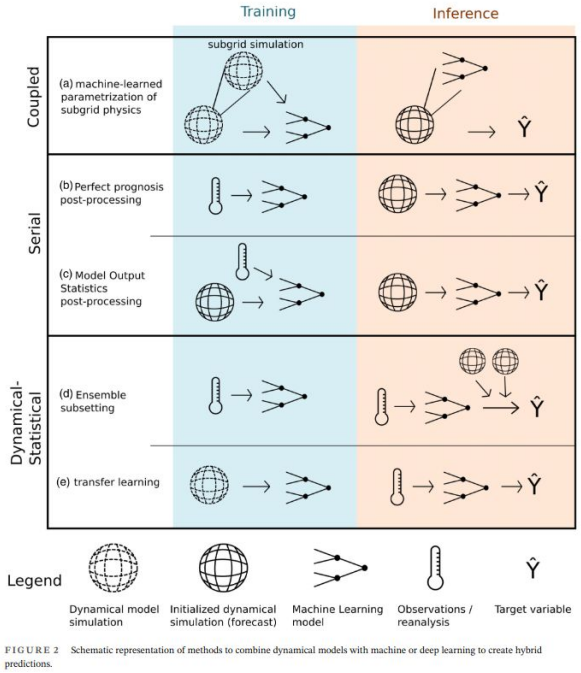
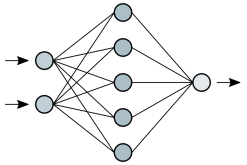


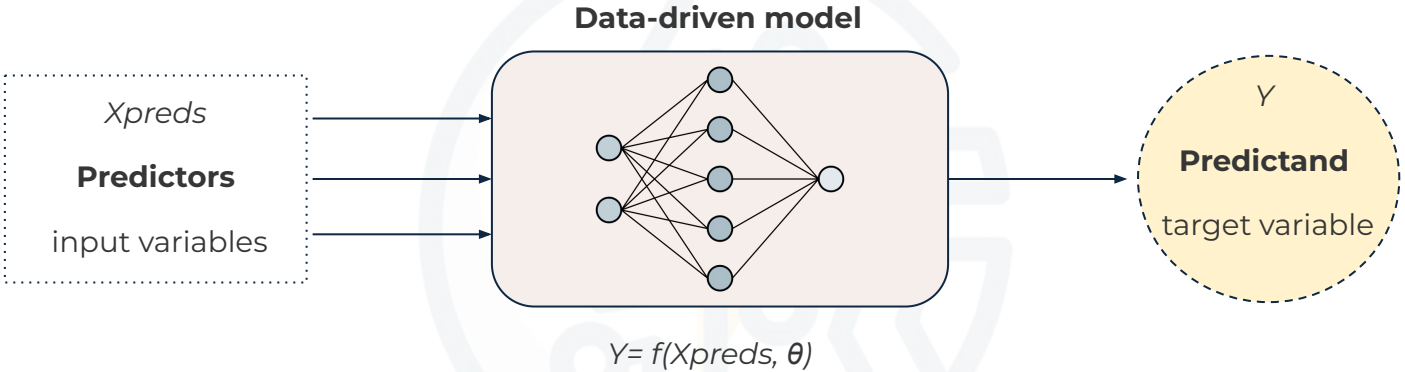
FIGURE 2 Schematic representation of methods to combine dynamical models with machine or deep learning to create hybrid predictions.

Materia et al., 2024

Funded by the European Union
Grant Agreement 101159497

2) Seasonal prediction: statistical predictions

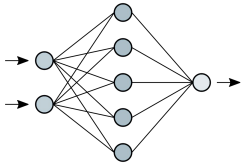


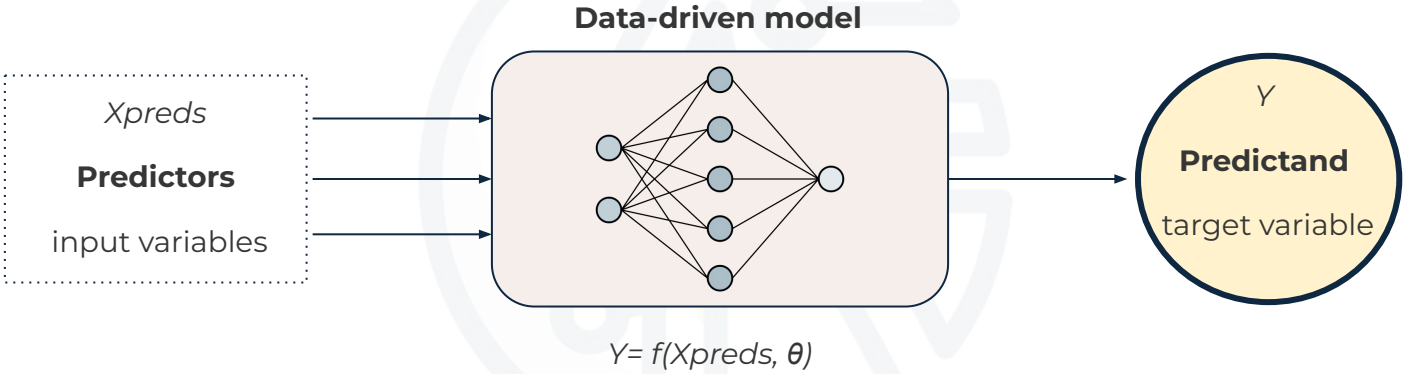


- We present an application of data-driven seasonal prediction of intense precipitation, **trained on reanalysis data**.
- Our choices of predictors, data-driven model, and predictand are each designed to tackle the **data-scarcity** issue.

Funded by the European Union
Grant Agreement 101159497

2) Seasonal prediction: statistical predictions





Funded by the European Union
Grant Agreement 101159497

3) Extreme precipitation events

Index	Health	Agriculture and food security	Water resources and food security	Coasts	Disaster Risk Reduction	Energy	Fisheries	Forestry/ GHGs	Cryosphere
CWD Consecutive Wet Days				✅					
R10mm Number of very heavy rain days				✅					
Rnnmm Number of customised rain days				✅					
SDII Daily PR intensity			✅					✅	
R95p Total annual PR from heavy rain days				✅					
R99p Total annual PR from very heavy rain days				✅					
Rx1day Max 1-day PR		✅		✅	✅			✅	
Rx5day Max 5-day PR		✅		✅	✅			✅	
CDD Consecutive Dry Days	✅	✅	✅		✅			✅	
R20mm Number of very heavy rain days		✅	✅					✅	
PRCPTOT Annual total wet-day PR		✅	✅					✅	
R95pTOT Contribution from very wet days		✅	✅					✅	
R99pTOT Contribution from extremely wet days		✅	✅					✅	
RXdday User-defined consecutive days PR amount	✅	✅	✅		✅			✅	
SPI Standardised Precipitation Index	✅	✅	✅		✅			✅	

Y

Predictand

target variable

- Multiple definitions exist in the scientific literature, there is no single standard.
- Most are based on the amount of water reaching the Earth's surface over a given **accumulation period**.
- The choice of **definition** (emphasising frequency, intensity, or duration) depends on the **application** and highlights different **physical processes** (e.g. convective events: short and intense; large-scale atmospheric circulation: longer-lasting).

Climpact (https://climpact-sci.org/indices/)

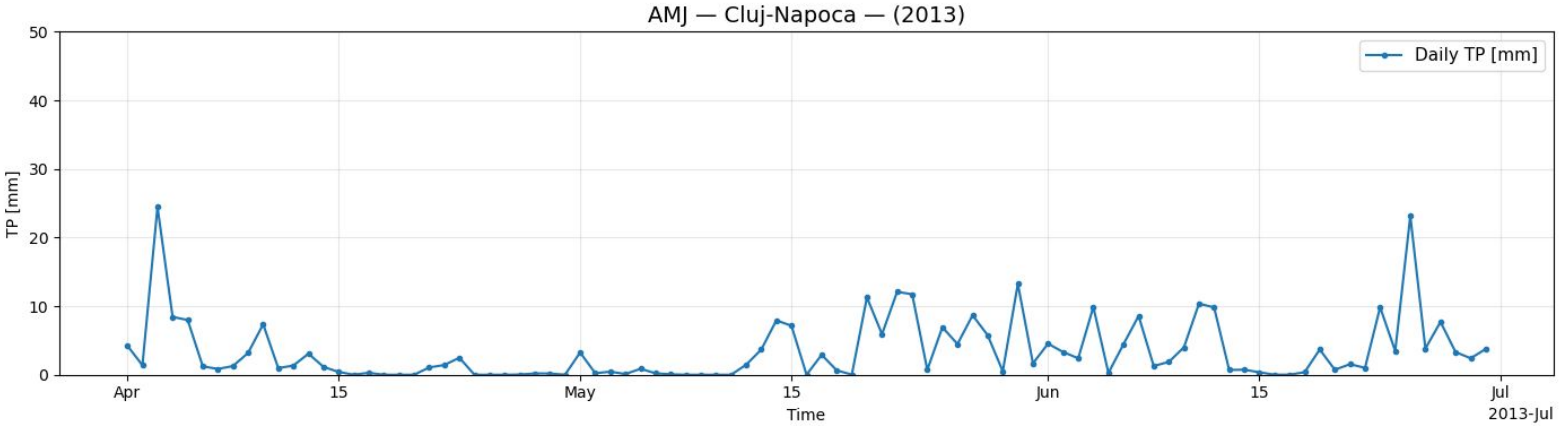
Funded by the European Union
Grant Agreement 101159497

Y
Predictand
target variable

3) Extreme precipitation events

N90p: The number of events per season in which 3-day accumulated precipitation exceeds the 90th-percentile threshold, where the threshold is computed over wet days during the reference period 1993–2016.

- ➛ **Intense** precipitation
- ➛ **Frequency**-oriented definition



Funded by the European Union
Grant Agreement 101159497

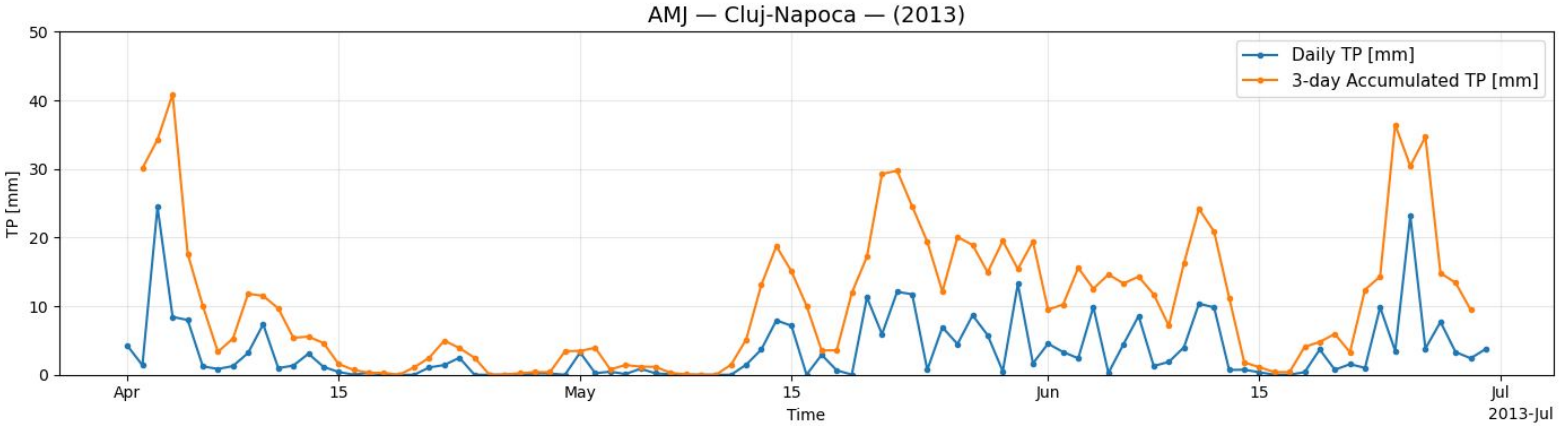
Y

Predictand

target variable

3) Extreme precipitation events

N90p: The number of events per season in which **3-day accumulated** precipitation exceeds the 90th-percentile threshold, where the threshold is computed over wet days during the reference period 1993–2016.



Funded by the European Union
Grant Agreement 101159497

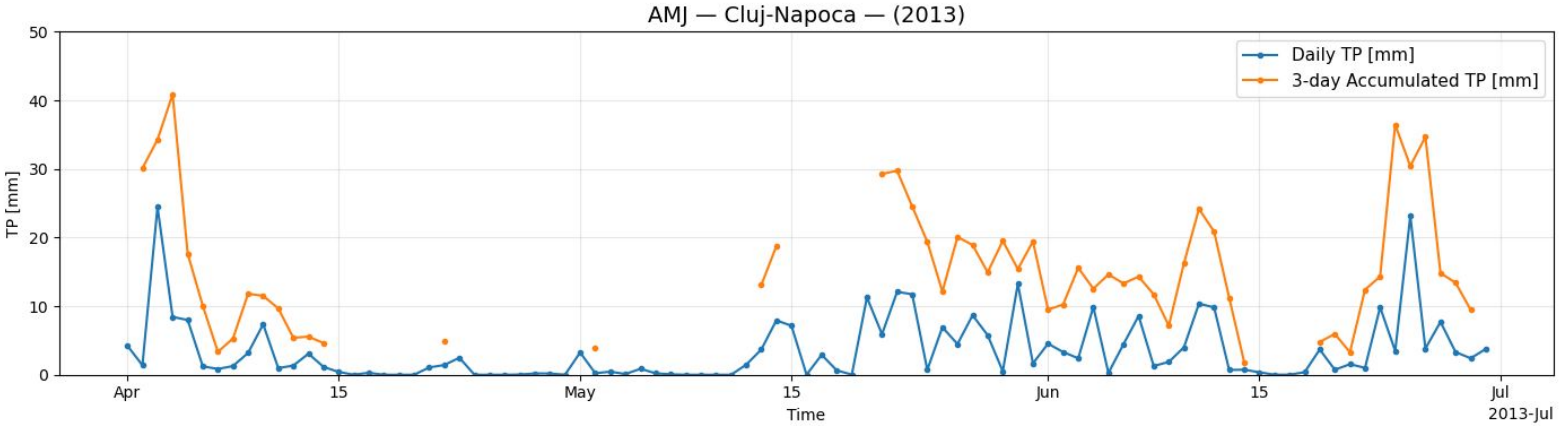
Y

Predictand

target variable

3) Extreme precipitation events

N90p: The number of events per season in which 3-day accumulated precipitation exceeds the 90th-percentile threshold, where the threshold is computed over **wet days** during the reference period 1993–2016.



Funded by the European Union
Grant Agreement 101159497

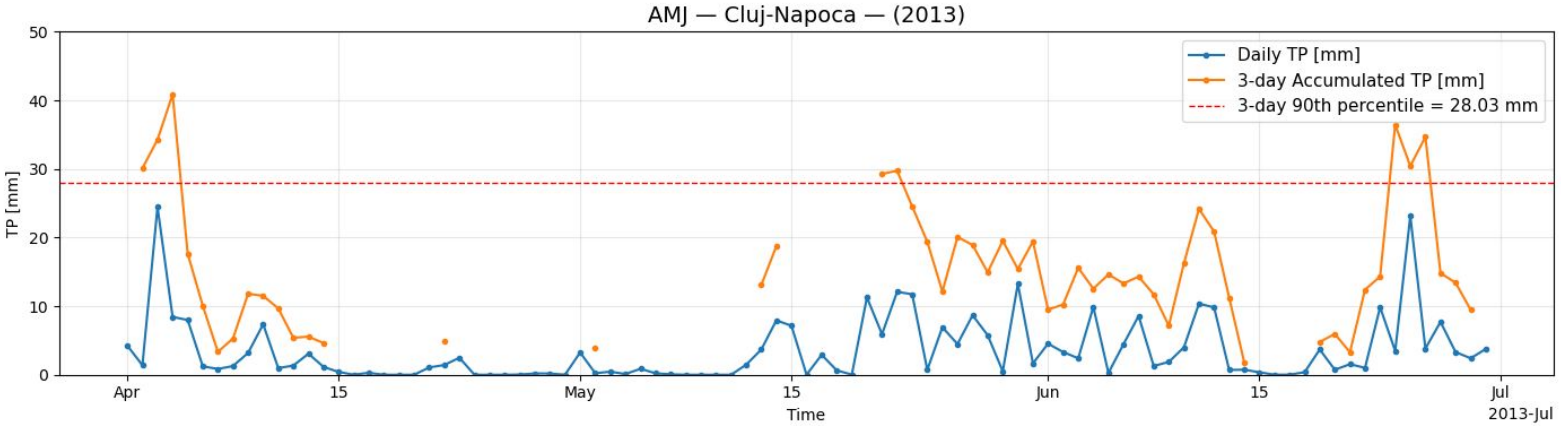
Y

Predictand

target variable

3) Extreme precipitation events

N90p: The number of events per season in which 3-day accumulated precipitation exceeds the **90th-percentile threshold**, where the threshold is computed over wet days during the reference period 1993–2016.



Funded by the European Union
Grant Agreement 101159497

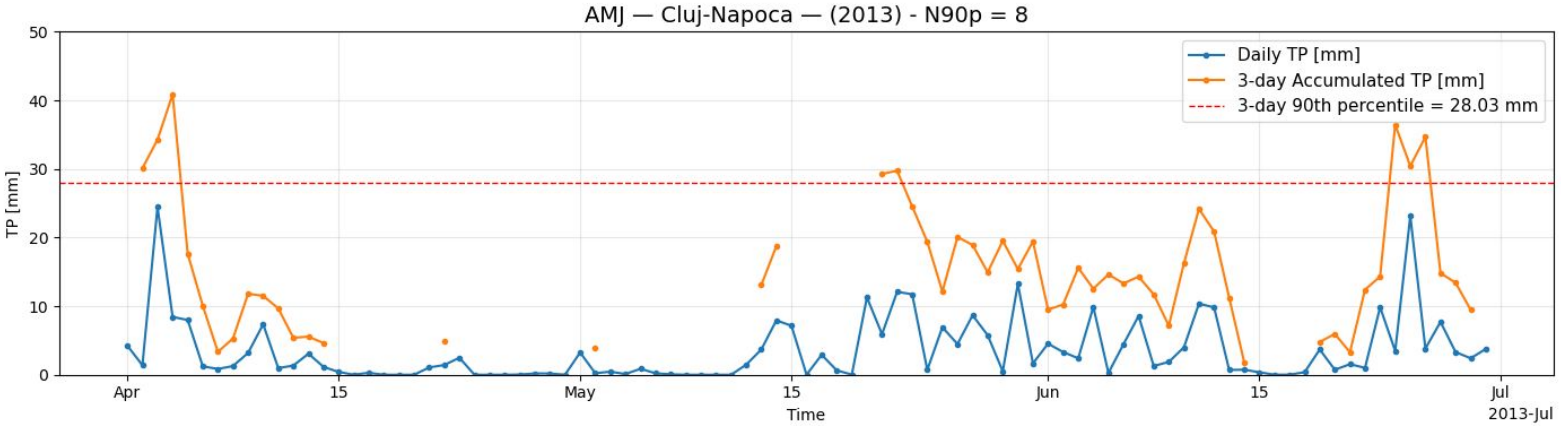
Y

Predictand

target variable

3) Extreme precipitation events

N90p: The **number** of events per season in which 3-day accumulated precipitation exceeds the 90th-percentile threshold, where the threshold is computed over wet days during the reference period 1993–2016.



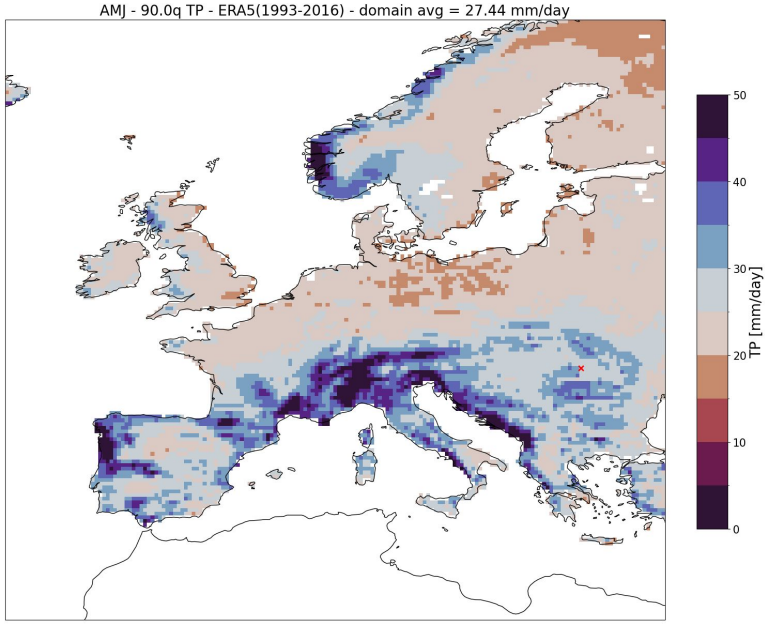
Funded by the European Union
Grant Agreement 101159497

3) Extreme precipitation events

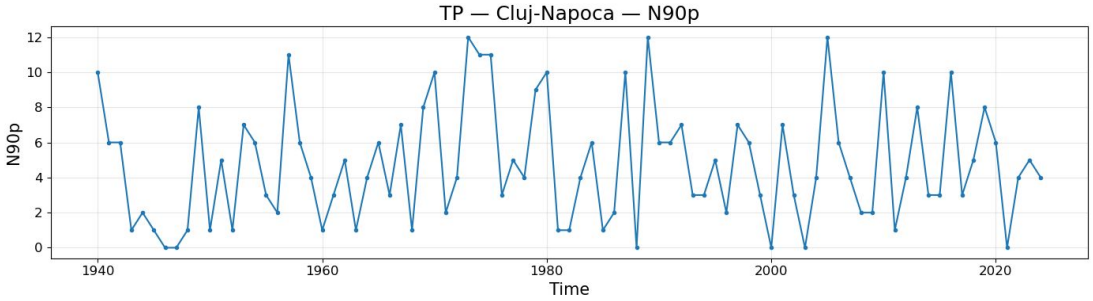
Y

Predictand

target variable



90th percentile for April, May and June (AMJ)



With such a limited dataset, even simple models can suffer from **overfitting**, i.e. learning patterns too closely tailored to the training data, which therefore fail to generalise to unseen data.

UNIVERSITATEA BABEȘ-BOLYAI
cmcc Centro Euro-Mediterraneo sui Cambiamenti Climatici
Universiteit Antwerpen
indeco soft

Funded by the European Union
Grant Agreement 101159497

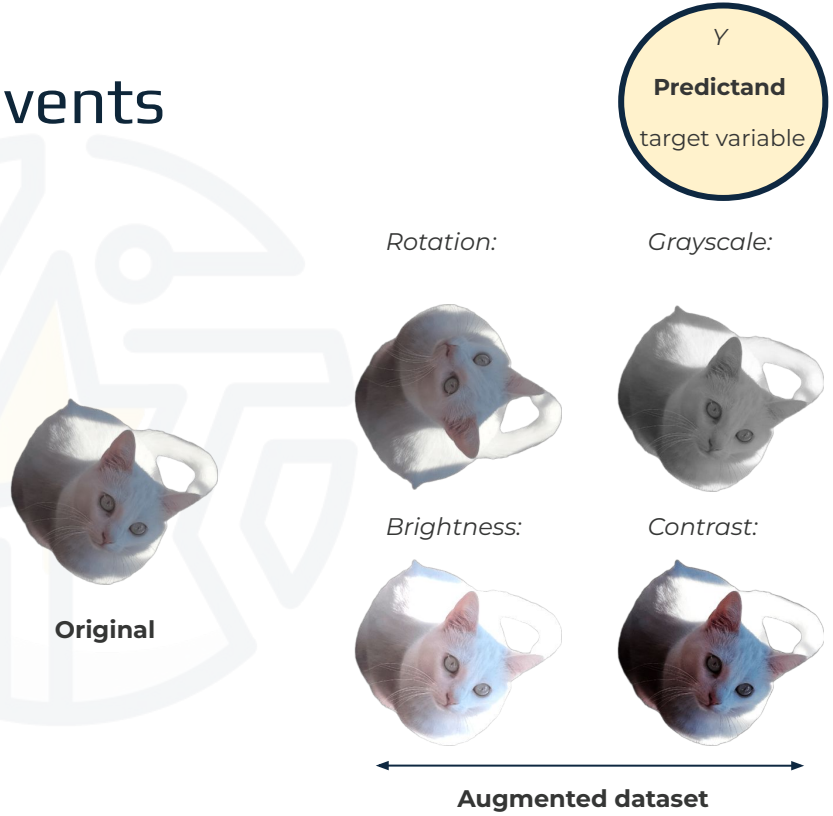
3) Extreme precipitation events

Data augmentation:

The best way to avoid overfitting?
Have more training data.

But new "real" data is often impossible or expensive to obtain. A possible idea: generate artificial data from the available dataset (**data augmentation**).

More samples to fit makes it harder for the model to memorise noise, reducing overfitting.



Funded by the European Union
Grant Agreement 101159497

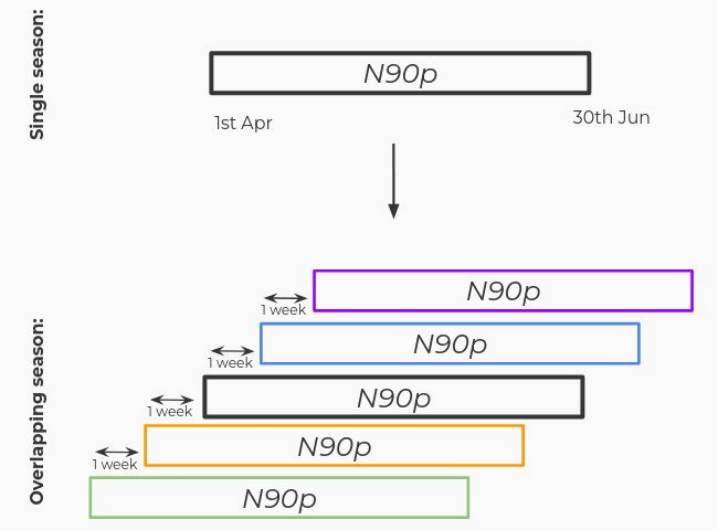
Y
Predictand
target variable

3) Extreme precipitation events

Data augmentation:

How can we create additional training samples in this context?

We can consider **overlapping, shifted seasons** around the one of interest: these can be expected to be driven by similar processes and dynamics, but to carry different target and associated predictor values.



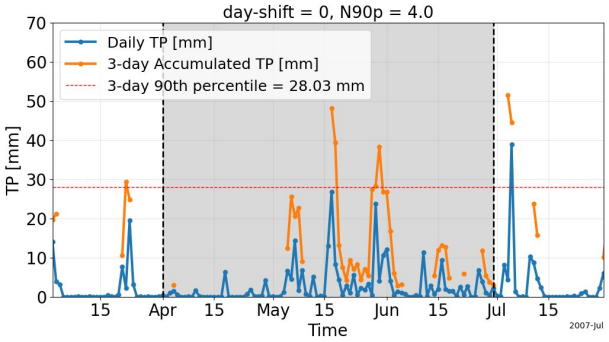
Funded by the European Union
Grant Agreement 101159497

3) Extreme precipitation events

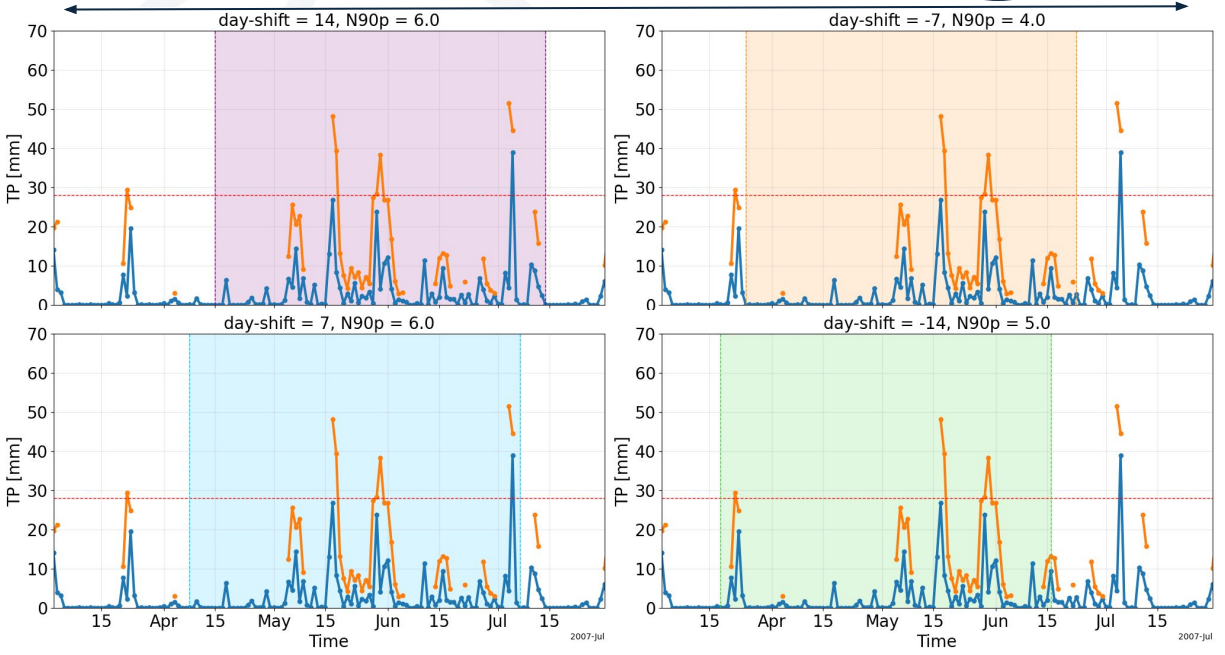
Y
Predictand
target variable

Data augmentation:

Original



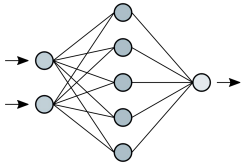
Augmented dataset

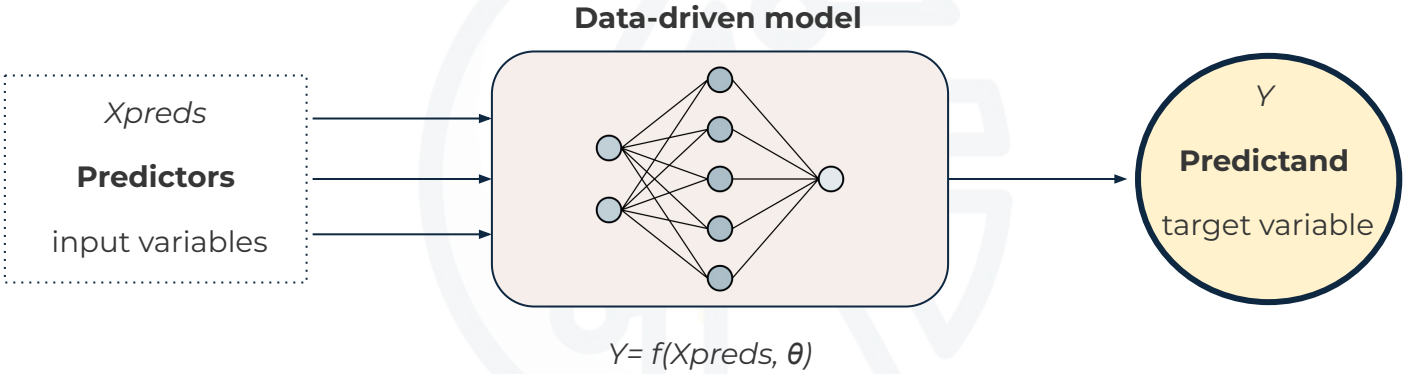


UNIVERSITATEA BABEȘ-BOLYAI
cmcc Centro Euro-Mediterraneo sui Cambiamenti Climatici
Universiteit Antwerpen
indeco soft

Funded by the European Union
Grant Agreement 101159497

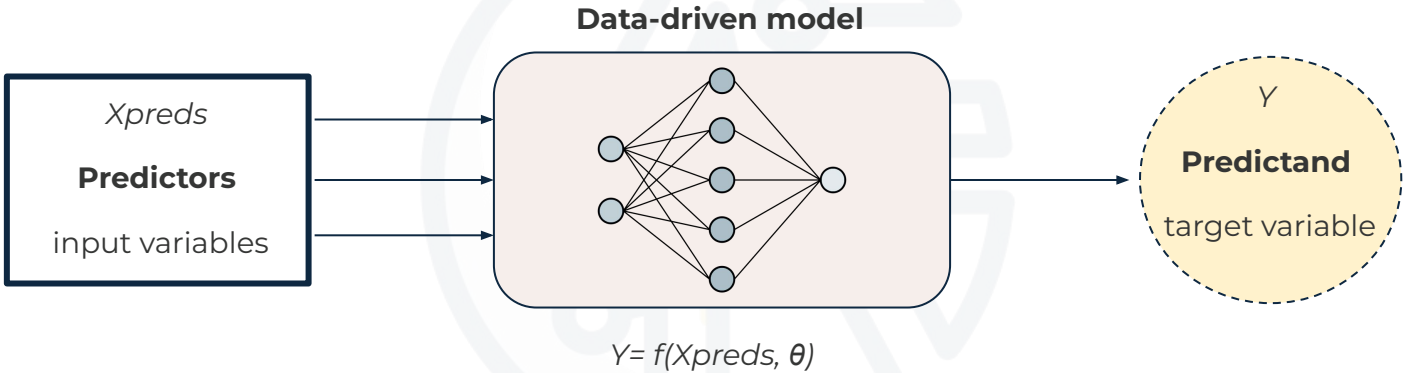
Seasonal prediction: statistical predictions





Funded by the European Union
Grant Agreement 101159497

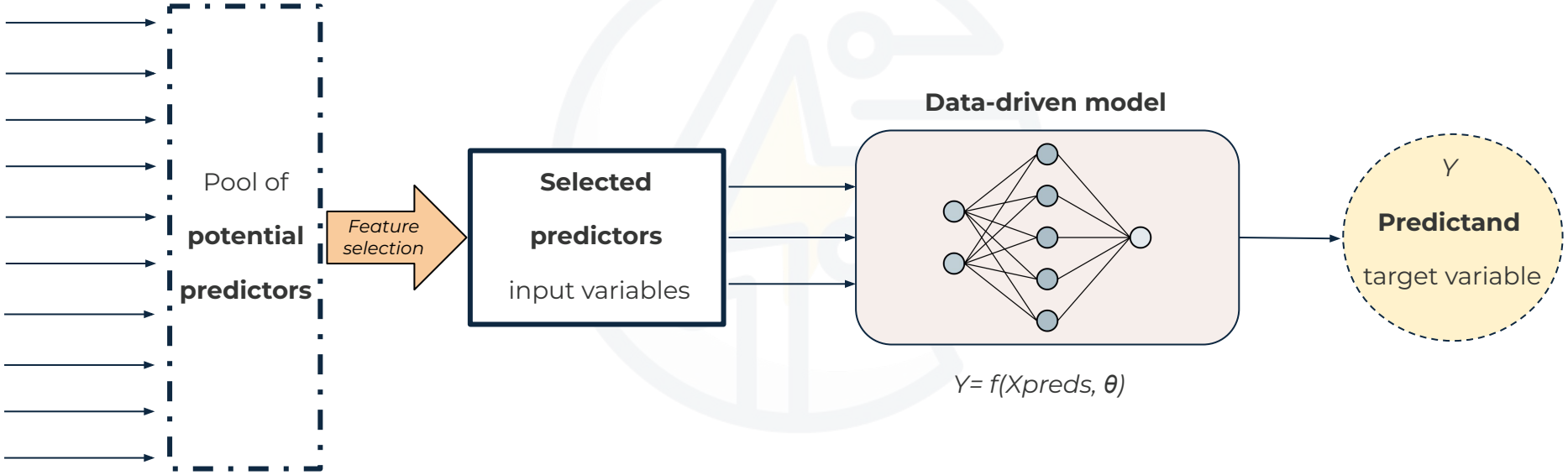
4) Feature selection



Since we cannot rely on advanced data-driven models, which cannot be trained with so little data, we shift our focus to identifying the **most impactful drivers** (predictors) for the target variable at each grid point.

Funded by the European Union
Grant Agreement 101159497

4) Feature selection



Funded by the European Union
Grant Agreement 101159497

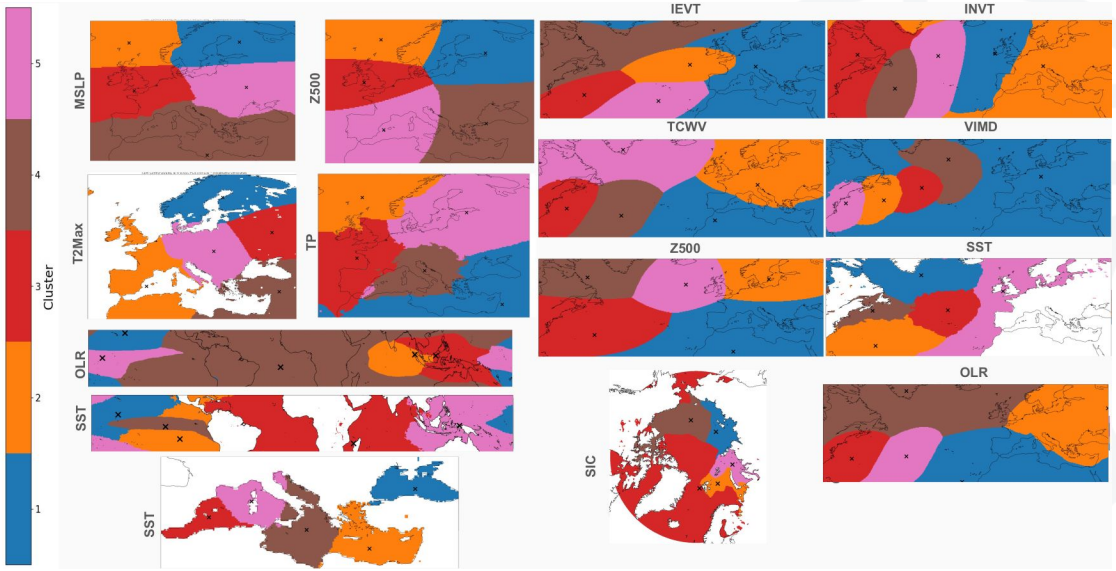
4) Feature selection

- A pool of **potential predictors** was selected to represent the atmospheric and oceanic conditions known to influence the European climate.
- In addition, a few **precipitation-oriented predictors** were included, with particular attention to the mechanisms driving moisture transport, especially from the North Atlantic, that can lead to intense precipitation over Europe.

Variable	*Region*
Vertical integral of eastward water vapor flux (IEVT)	*North Atlantic*
Vertical integral of northward water vapor flux (INVT)	*North Atlantic*
Total column water vapor (TCWV)	*North Atlantic*
Vertical integrated moisture divergence (VIMD)	*North Atlantic*
Sea surface temperature (SST)	*North Atlantic, Tropics, Mediterranean and Black Sea*
500 hPa geopotential height (Z500)	*Europe, North Atlantic*
Mean sea level pressure (MSLP)	*Europe*
Total Precipitation (TP)	*Europe*
Max 2-meter temperature (T2max)	*Europe*
Outgoing longwave radiation (OLR)	*North Atlantic, Tropics*
Sea Ice Cover (SIC)	*Arctic*

Funded by the European Union
Grant Agreement 101159497

4) Feature selection



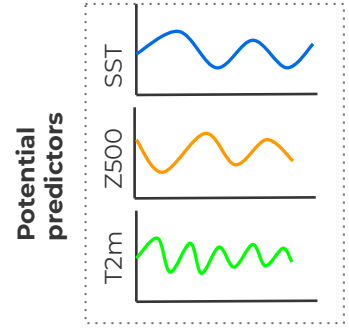
- We **cluster** the pool of potential predictors to reduce their spatial dimensionality.
- This is done with an **enhanced k-means** algorithm using a weighted, multi-dimensional distance (*Squintu, et al., in preparation*).
- For each resulting cluster, a single **weekly time series** is then computed as the area-weighted spatial average of its grid points.

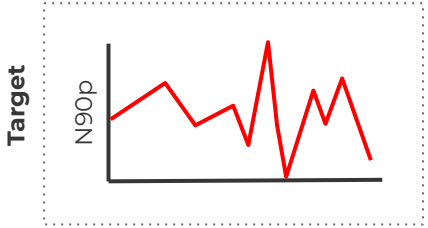
Funded by the European Union
Grant Agreement 101159497

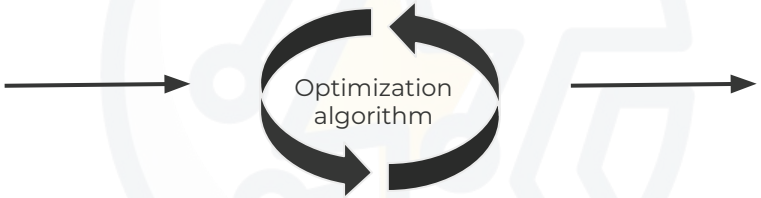
4) Feature selection

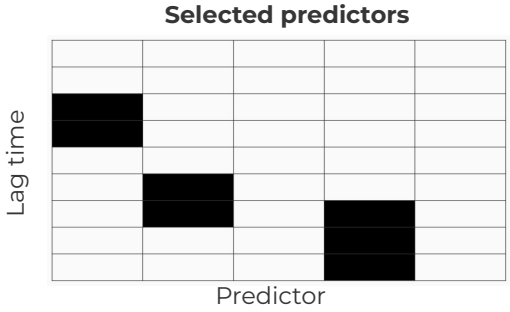
McAdam et al., 2025

Training set









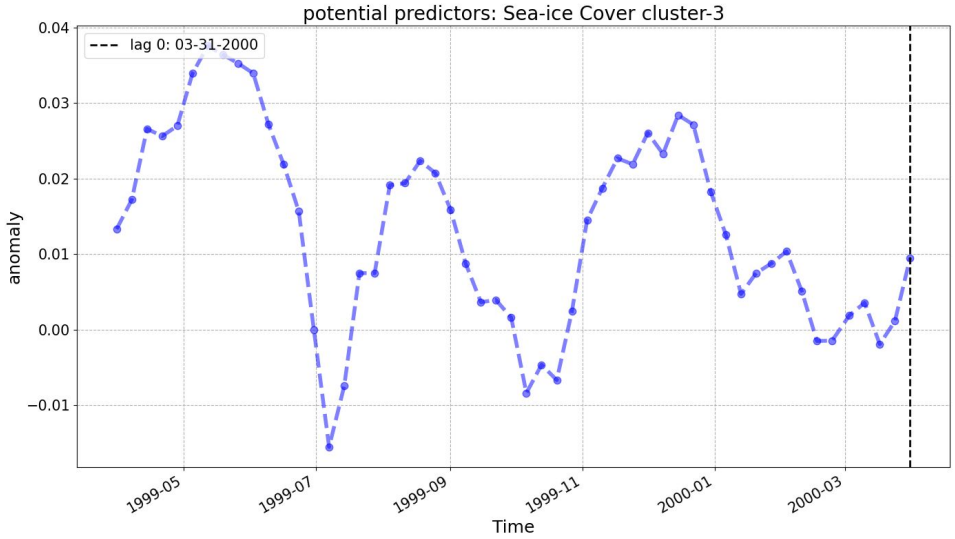
Optimisation-based **feature selection** of most influential drivers on seasonal timescale (Probabilistic Coral Reef Optimization Algorithm with Substrate Layers, PCRO-SL).

Funded by the European Union
Grant Agreement 101159497

4) Feature selection

For each potential predictor, the selection corresponds to a configuration of **three parameters**: the variable cluster, the sequence length, and the time-lag.

- **Variable cluster**: whether a given cluster variable is selected or not.
- **Sequence length**: the length of the period over which the cluster variable is considered.
- **Time-lag**: how far before the target season this period ends.

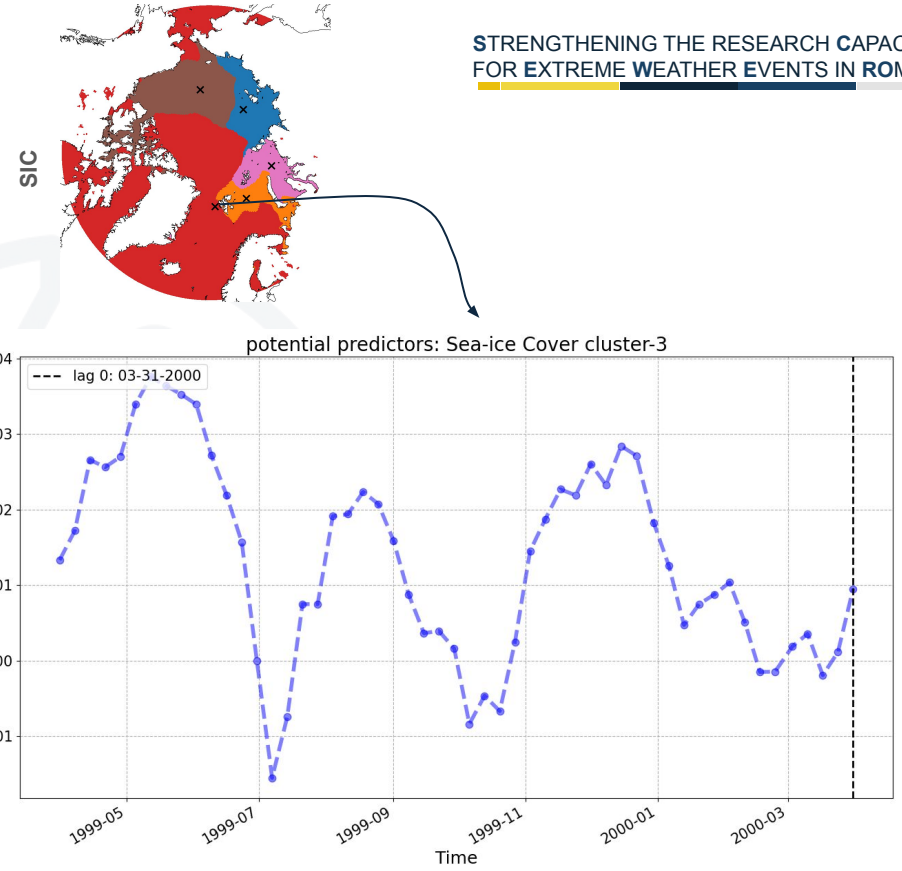


Funded by the European Union
Grant Agreement 101159497

4) Feature selection

For each potential predictor, the selection corresponds to a configuration of **three parameters**: the variable cluster, the sequence length, and the time-lag.

- **Variable cluster**: *True*
- **Sequence length**: *4 weeks*
- **Time-lag**: *5 weeks*

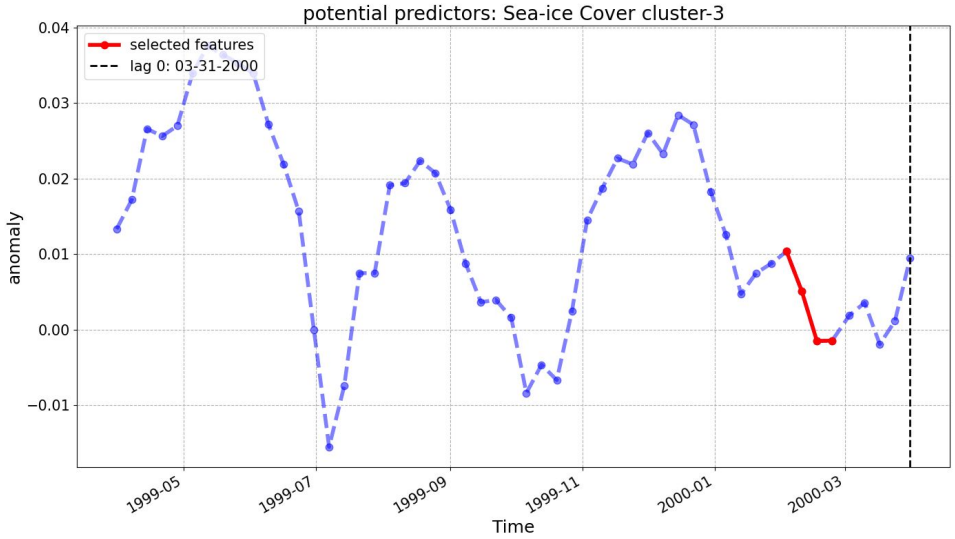


Funded by the European Union
Grant Agreement 101159497

4) Feature selection

For each potential predictor, the selection corresponds to a configuration of **three parameters**: the variable cluster, the sequence length, and the time-lag.

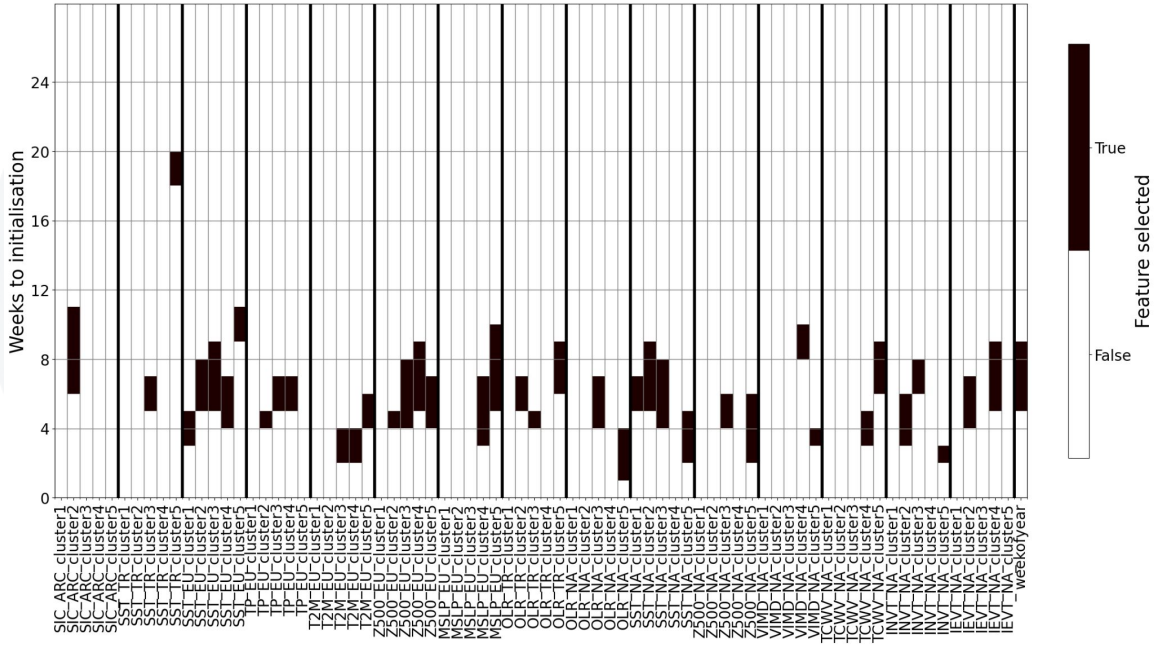
- **Variable cluster**: *True*
- **Sequence length**: *4 weeks*
- **Time-lag**: *5 weeks*



Funded by the European Union
Grant Agreement 101159497

4) Feature selection

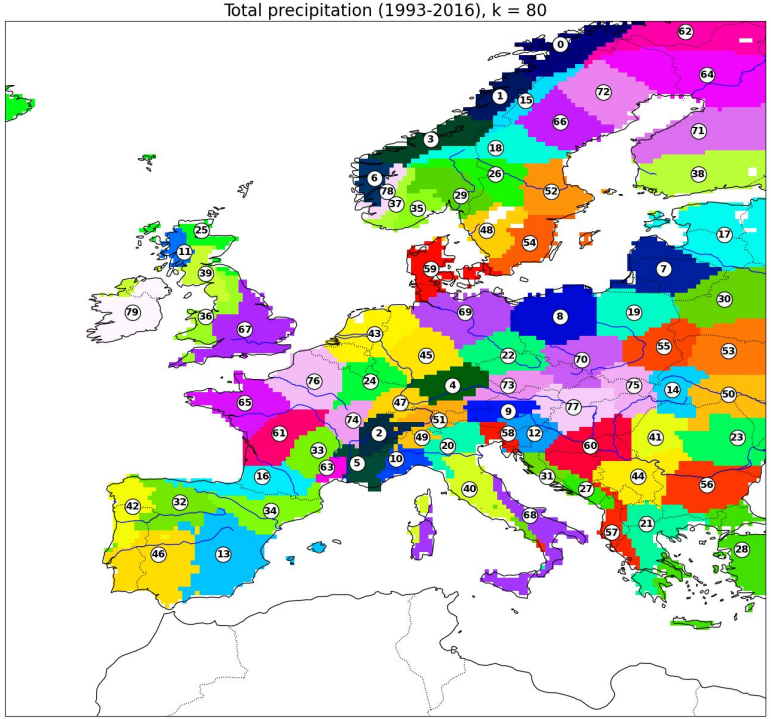
- Example of a possible output of the feature-selection algorithm for a **single grid point**.
- The feature selection is performed separately for the **four seasons** analysed (JFM, AMJ, JAS, OND), since different optimal predictors are expected for different seasons.



Funded by the European Union
Grant Agreement 101159497

4) Feature selection

An additional **post-processing** step is applied to the predictors selected for each grid point.
It is motivated by the need to reduce the intrinsic **randomness** of the feature-selection method, and by the **assumption** that similar predictors should play a comparable role for grid points with similar precipitation behaviour.



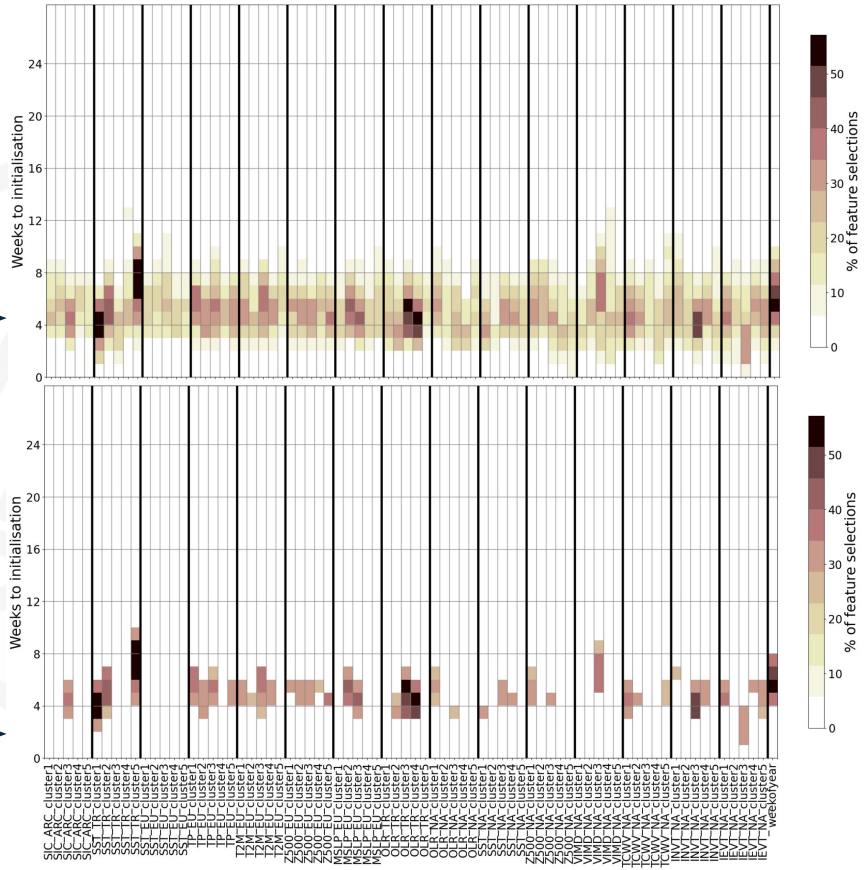
The same **clustering** methodology presented before is applied to the total precipitation (TP) variable, using k = 80 clusters.

Funded by the European Union
Grant Agreement 101159497

4) Feature selection

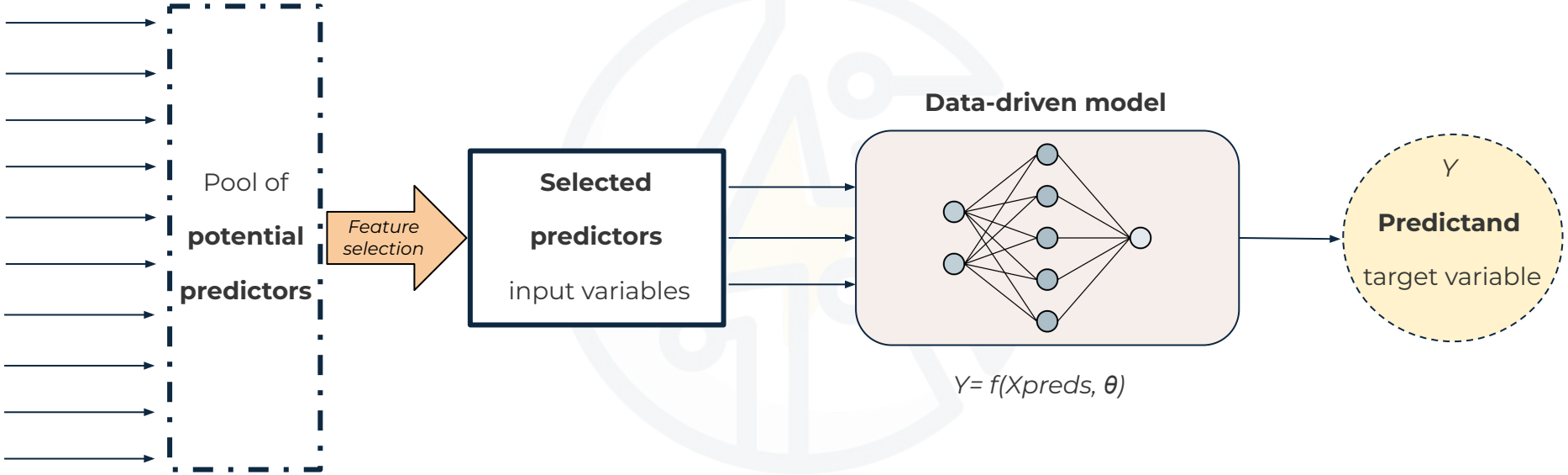
Once the clusters are defined, the predictors selected through the feature-selection procedure are **aggregated** across all grid points belonging to the same cluster.

Only the most frequently selected predictors are retained as the optimal predictor set for that cluster, defining a **common set of predictors** for all grid points within it.



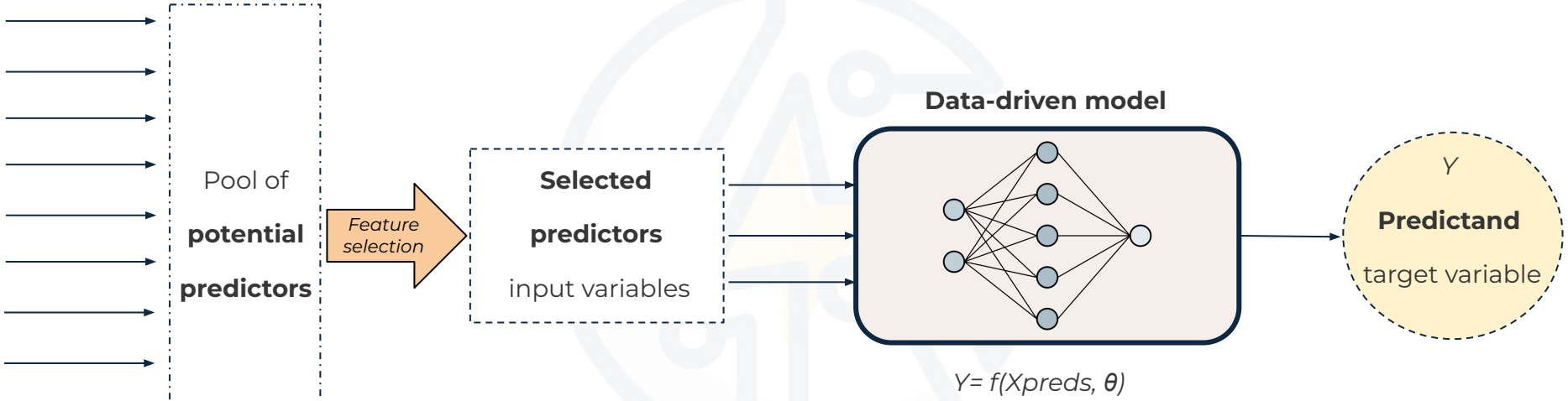
Funded by the European Union
Grant Agreement 101159497

4) Feature selection



Funded by the European Union
Grant Agreement 101159497

5) Data-driven prediction



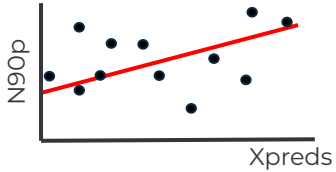
Because the training dataset is small, we must rely on data-driven methods suited to **small-sample regimes**, in order to minimize the risk of overfitting.

Funded by the European Union
Grant Agreement 101159497

5) Data-driven prediction

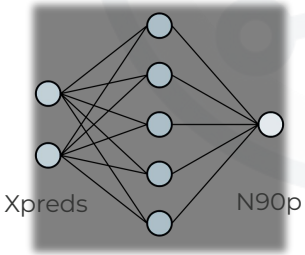
Ridge and **Lasso Regression**:

regularized multivariate linear models that add a penalty term to the mean squared error (L2 norm for Ridge, L1 norm for Lasso) to prevent overfitting and handle potential multicollinearity



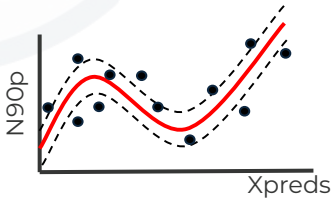
Multilayer Perceptron (MLP):

a shallow neural network restricted to a single hidden layer of 10 nodes to keep the parameter count low.



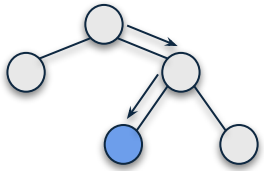
Support Vector Regression (SVR):

a machine-learning algorithm able to capture both linear and non-linear relationships through different kernels, here linear and polynomial.



XGBoost (XGB):

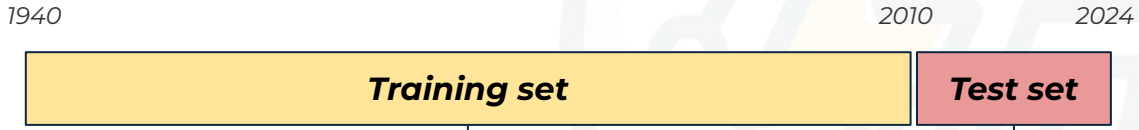
a gradient-boosted ensemble of decision trees, built sequentially so that each new tree corrects the errors of the previous ones.



Funded by the European Union
Grant Agreement 101159497

5) Data-driven prediction

- Both the potential predictors and the target values used in this work are derived from the **ERA5** reanalysis dataset. ERA5 provides global coverage on a 0.25° spatial grid, with no spatial or temporal gaps.
- The period from 1940 to 2024 is considered.



- Feature selection
- Training of the data-driven models

- Evaluation of performance in terms of temporal correlation coefficient (ρ) and root mean square error (RMSE)
- Selection of the best-performing (max ρ) data-driven model for each grid point

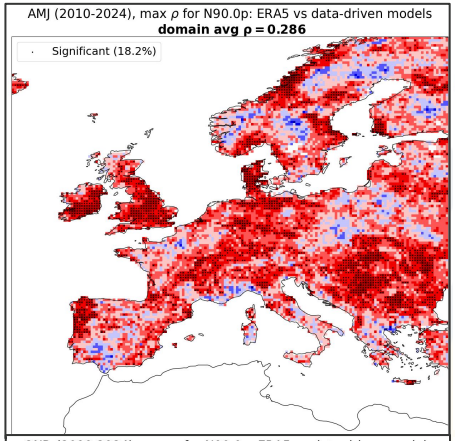
Funded by the European Union
Grant Agreement 101159497

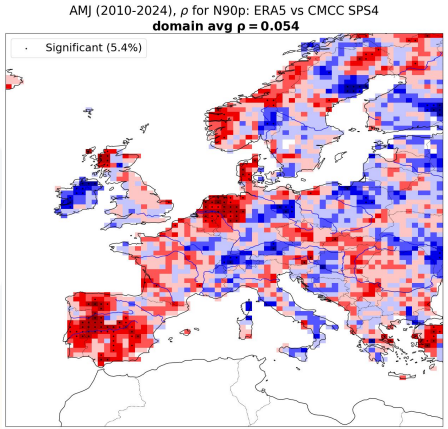
6) Results

Statistical prediction - Data-driven:

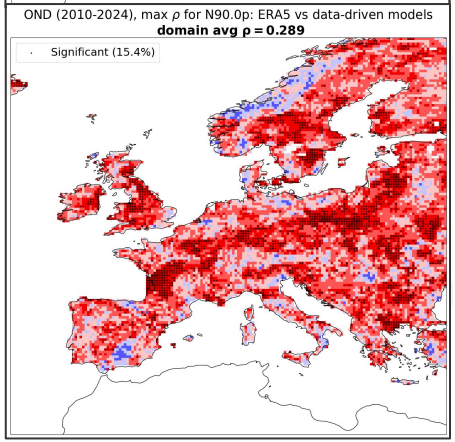
Dynamical prediction - CMCC SPS4:

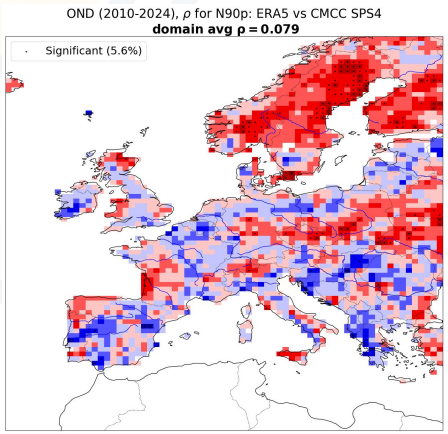
AMJ:

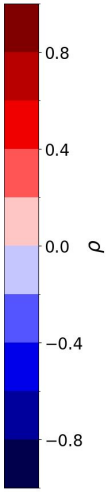




OND:







Funded by the European Union
Grant Agreement 101159497

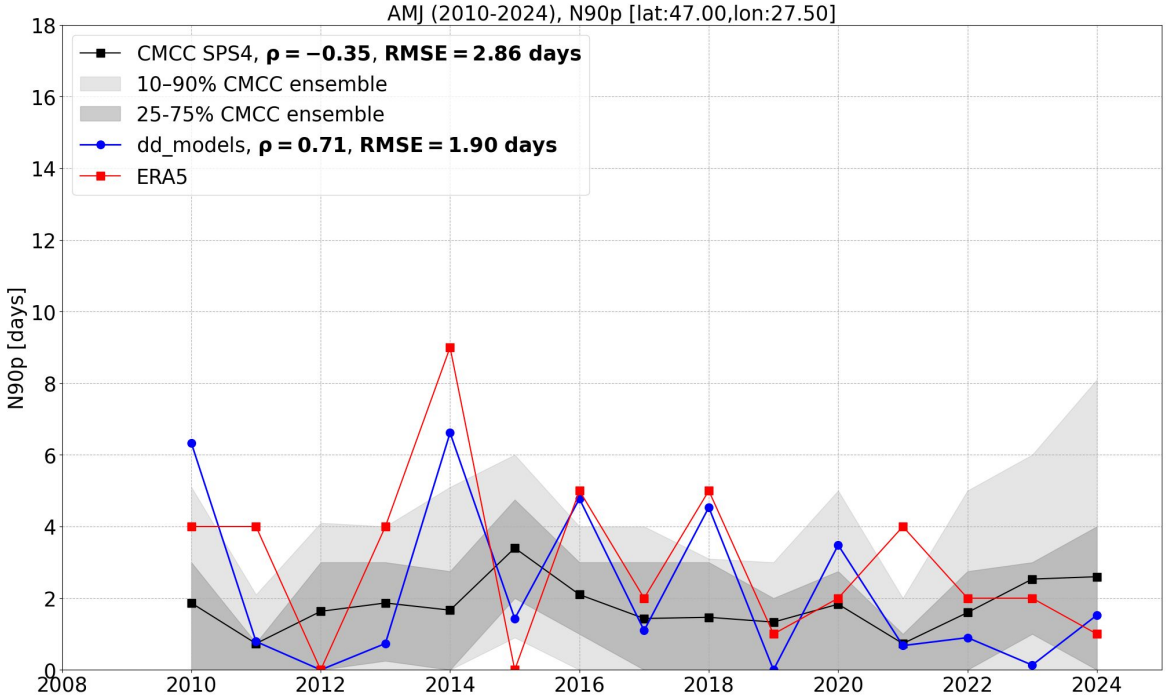
6) Results

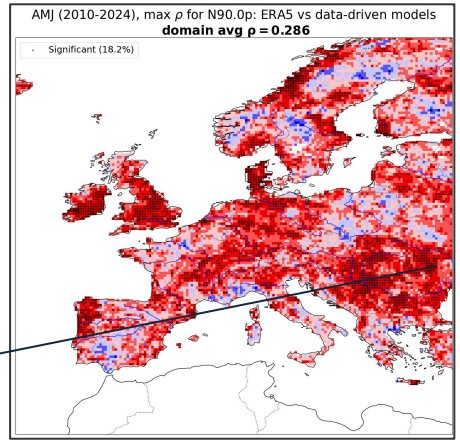
	JFM		AMJ		JAS		OND	
	ERA5	MSWEP	ERA5	MSWEP	ERA5	MSWEP	ERA5	MSWEP
Grid-points (%) with CMCC-SPS4 significant skill	**15.3**	**15.4**	5.4	4.6	5.2	3.6	5.6	5.1
Grid-points (%) with data-driven significant skill	11.1	10.6	**18.2**	**16.5**	**12.4**	**14.4**	**15.4**	**14.4**
$(RMSE_{dd} - RMSE_{SPS4})$ within areas of data-driven significant skill	-0.57	-0.85	-0.71	-0.58	-0.60	-0.42	-0.64	-0.79

Performance has additionally been evaluated against the N90p index extracted from the **MSWEP** dataset, a higher-resolution precipitation product that merges gauge, satellite, and model-based estimates.

Funded by the European Union
Grant Agreement 101159497

6) Results

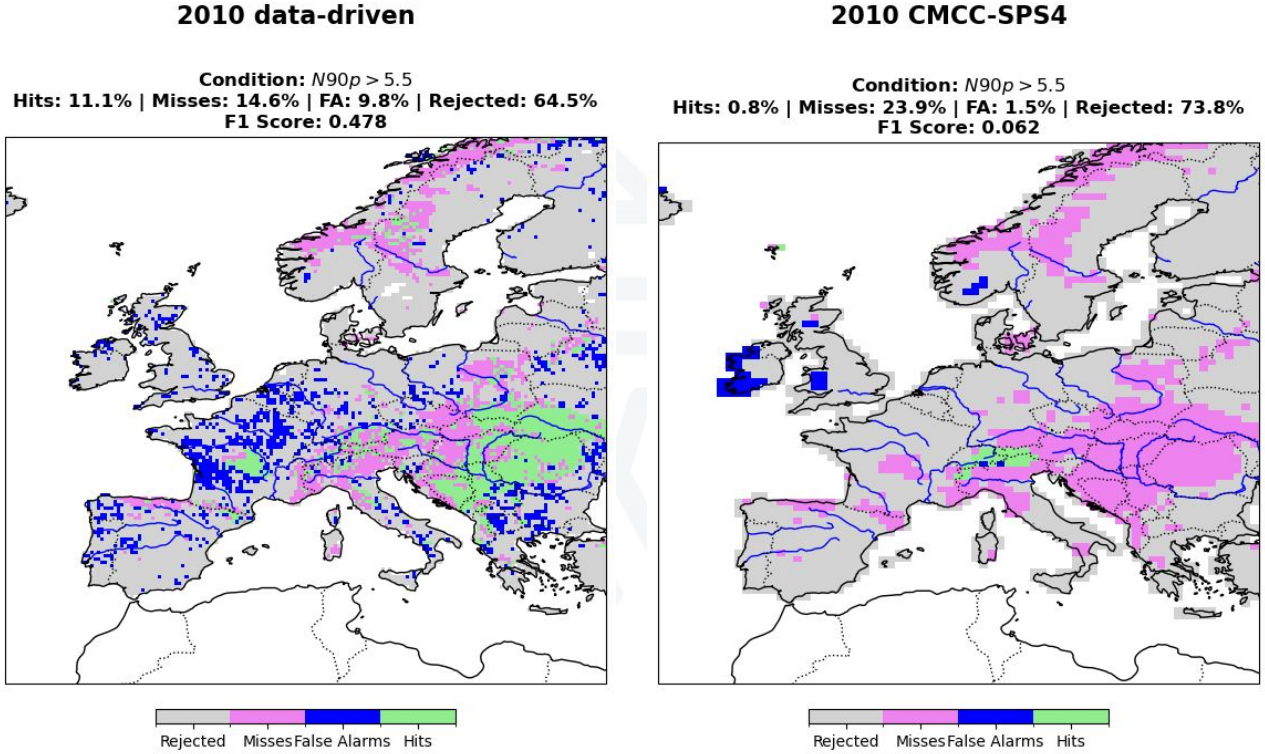




Funded by the European Union
Grant Agreement 101159497

6) Results

Contingency map for the condition N90p ≥ 6 events per season, AMJ

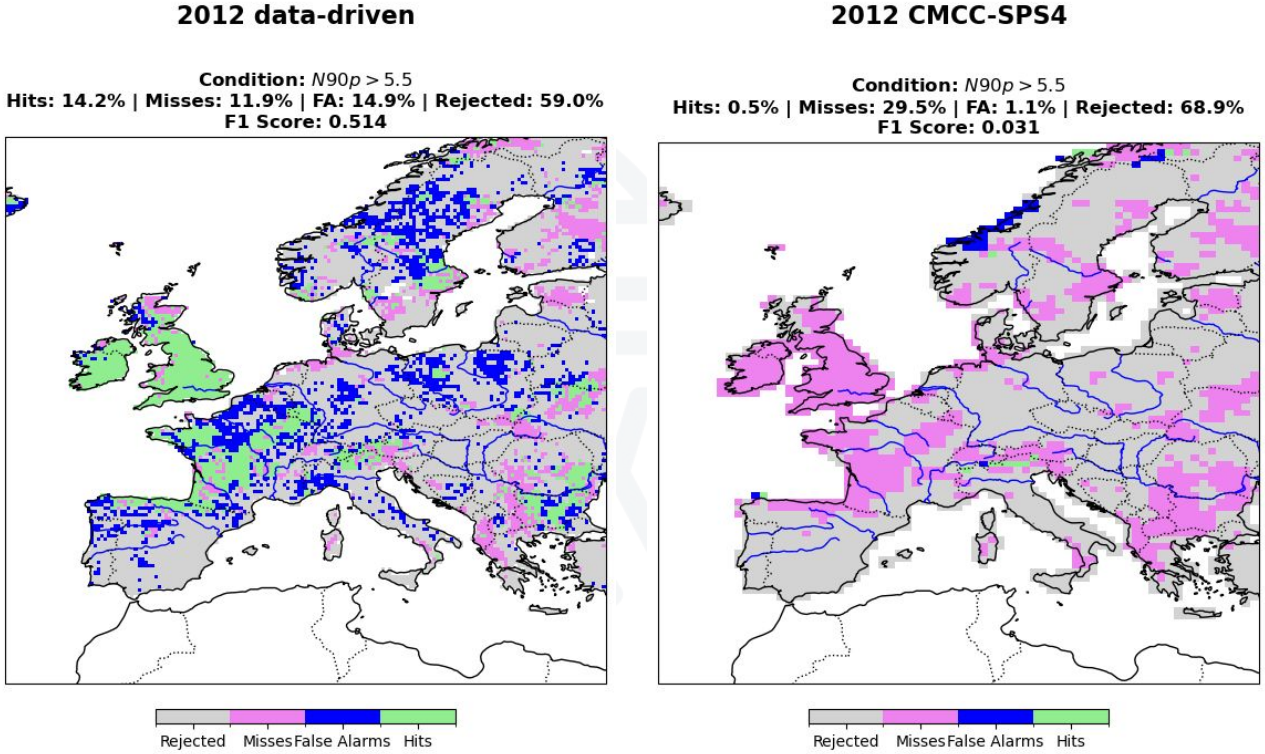


UNIVERSITATEA BABEȘ-BOLYAI
cmcc Centro Euro-Mediterraneo sui Cambiamenti Climatici
Universiteit Antwerpen
indeco soft

Funded by the European Union
Grant Agreement 101159497

6) Results

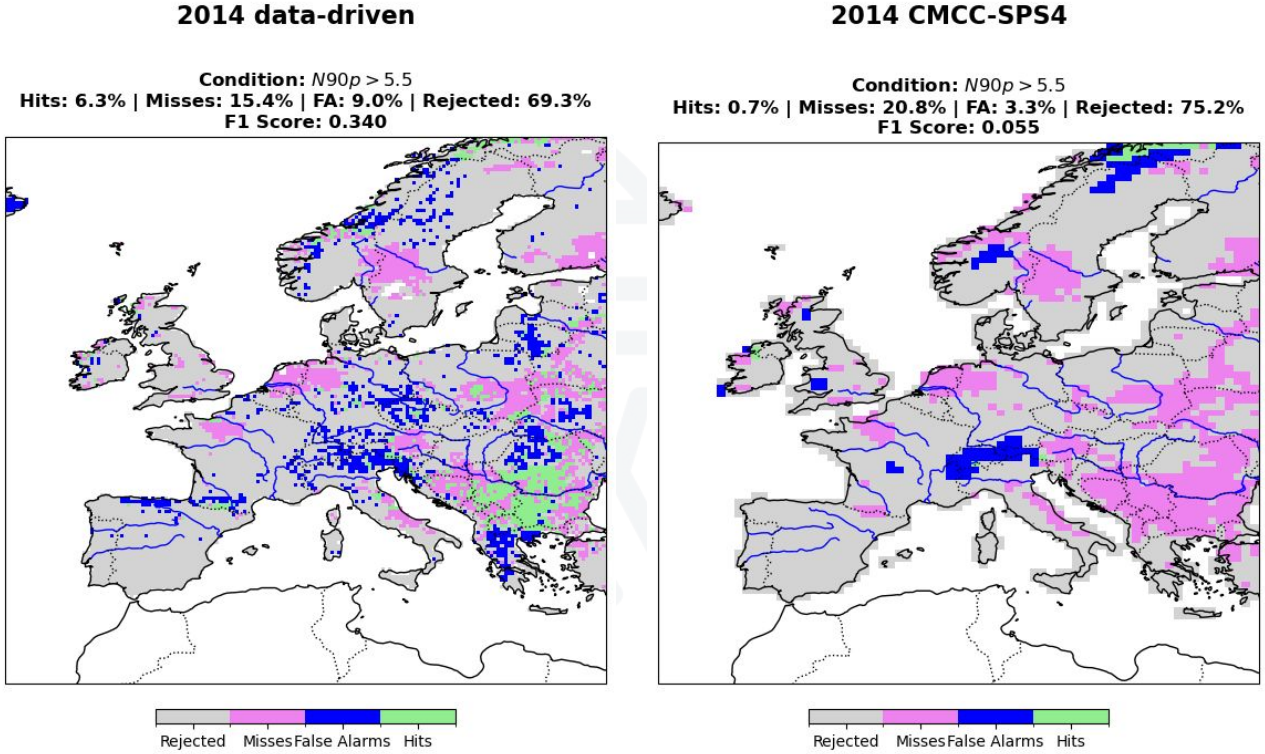
Contingency map for the condition N90p ≥ 6 events per season, AMJ



Funded by the European Union
Grant Agreement 101159497

6) Results

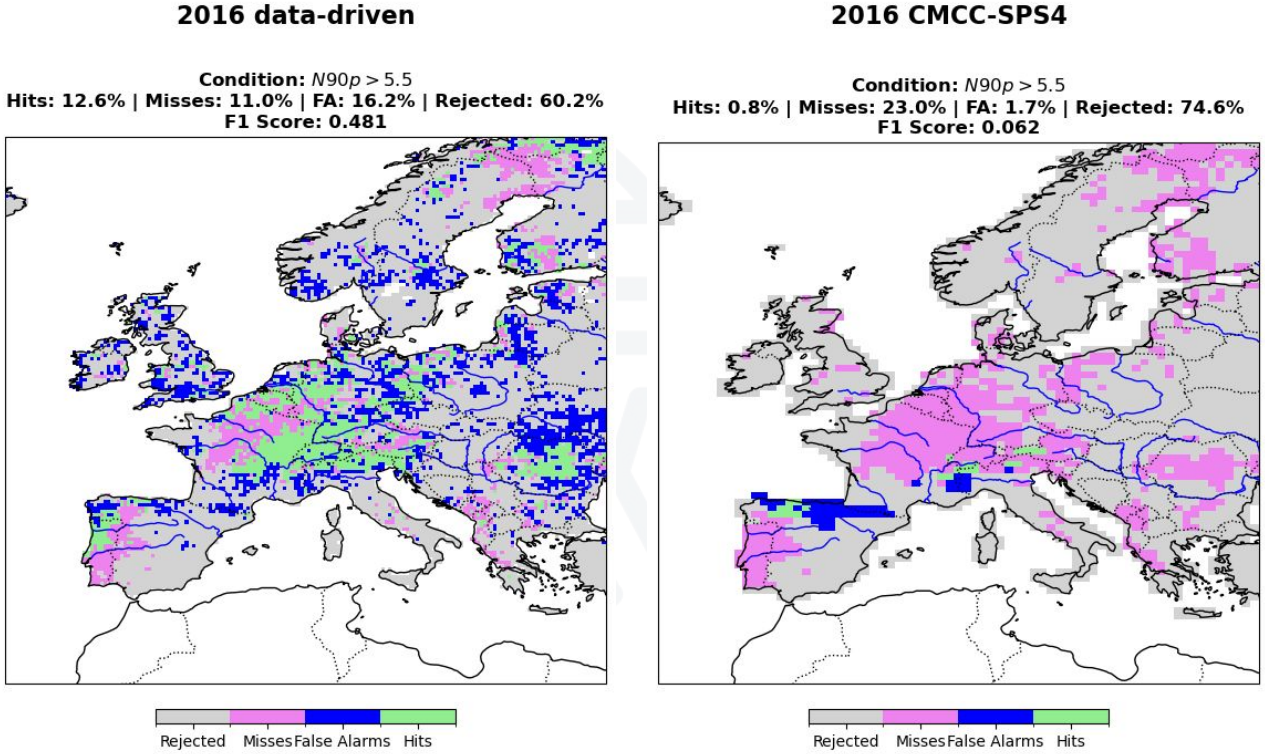
Contingency map for the condition N90p ≥ 6 events per season, AMJ



Funded by the European Union
Grant Agreement 101159497

6) Results

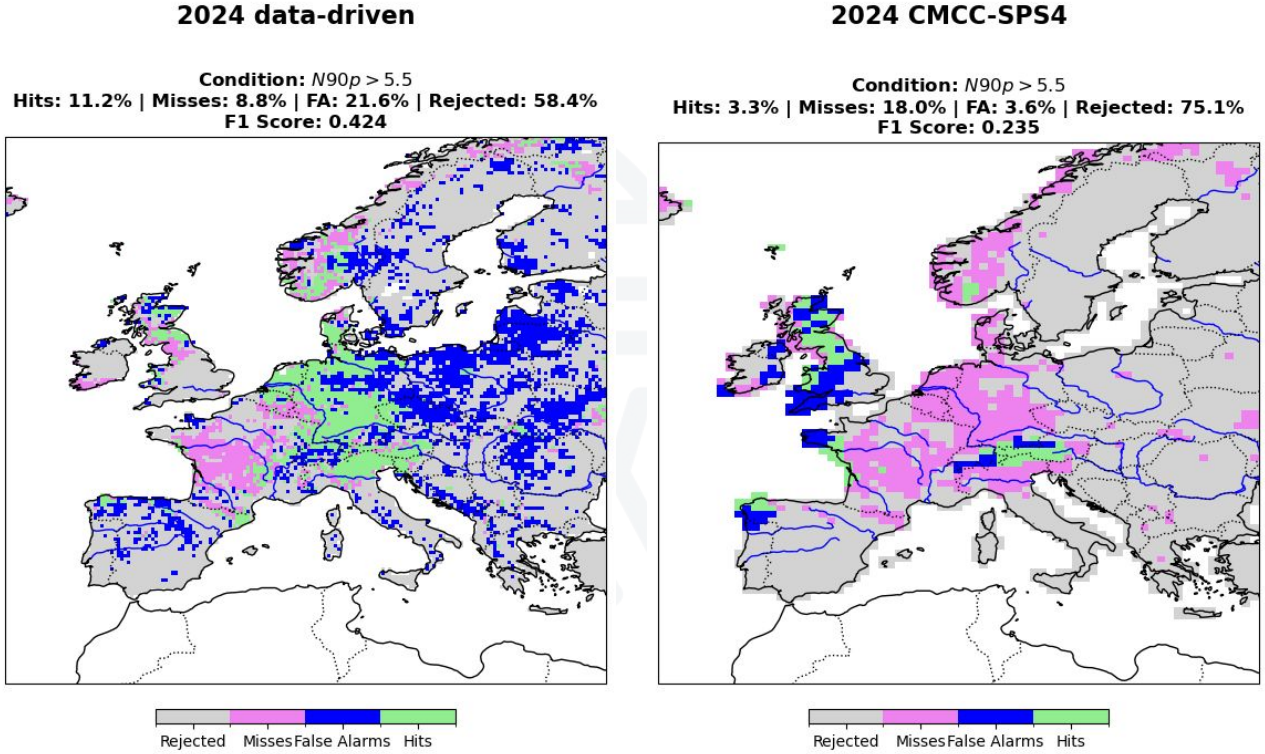
Contingency map for the condition N90p ≥ 6 events per season, AMJ



Funded by the European Union
Grant Agreement 101159497

6) Results

Contingency map for the condition N90p ≥ 6 events per season, AMJ



Funded by the European Union
Grant Agreement 101159497

6) Results

Conclusions:

- This work shows that even simple data-driven models, if trained on the right predictors, can achieve **skill** comparable to, or even superior to, expensive dynamical models.
- The proposed **data-augmentation** mechanism proves effective both for the feature-selection framework and for training data-driven models on reanalysis data in a seasonal prediction context.
- It opens the possibility of identifying new drivers of intense precipitation over Europe and improving the understanding of the underlying **physical processes**.

Funded by the European Union
Grant Agreement 101159497

Thank you!

Enrico Scoccimarro, Leone Cavicchia, Ronan McAdam, Antonello Squintu, Luca Famooss Paolini

UNIVERSITATEA BABEȘ-BOLYAI
cmcc Centro Euro-Mediterraneo sui Cambiamenti Climatici
Universiteit Antwerpen
indeco soft

Funded by the European Union
Grant Agreement 101159497