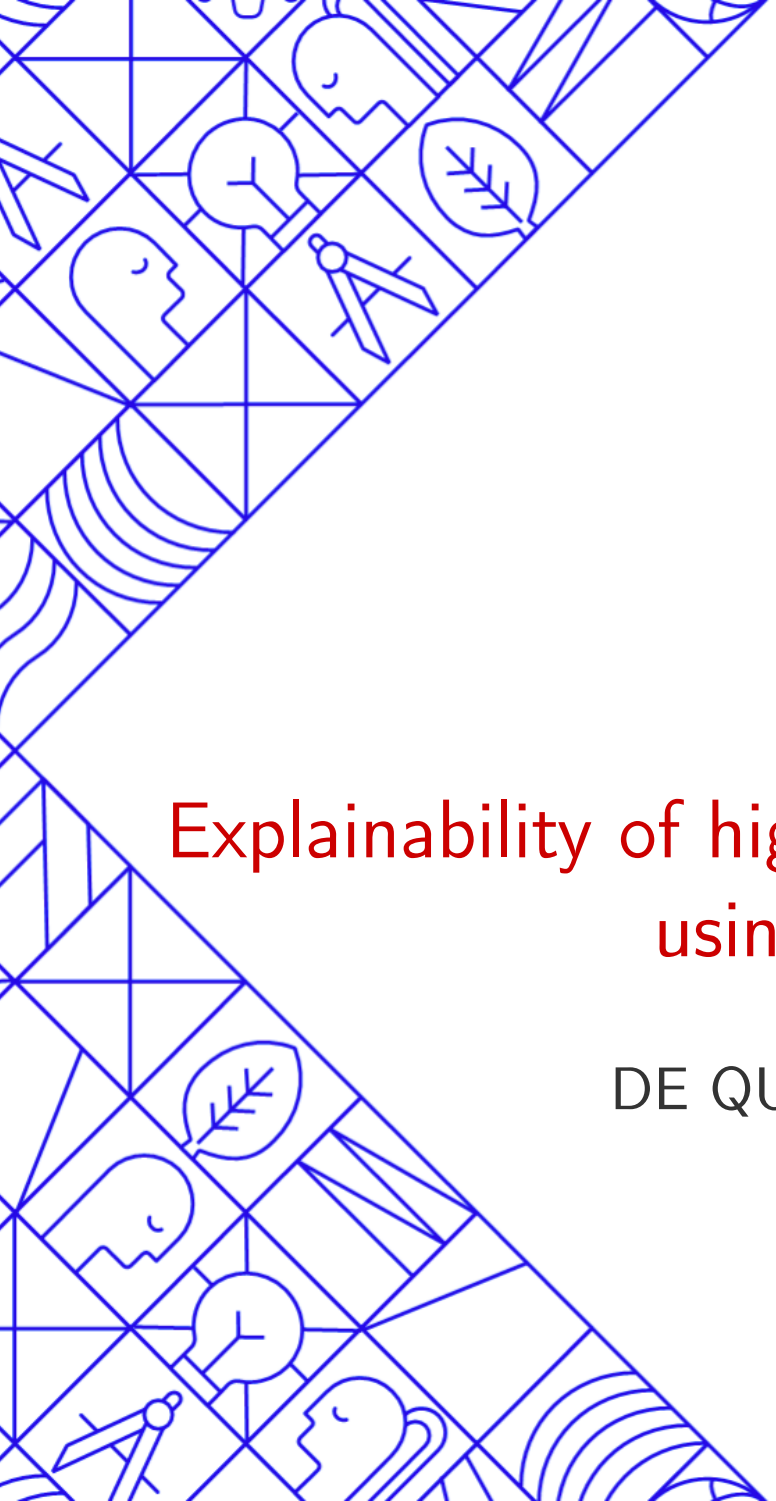


ENSTA

IP PARIS

Explainability of high-dimensional prediction models using neural networks

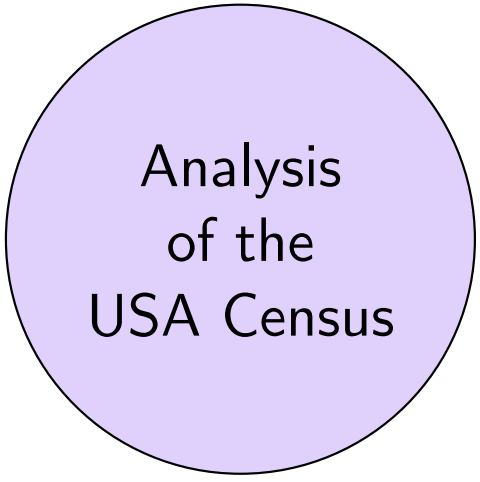
DE QUEIROZ GARCIA, Luana

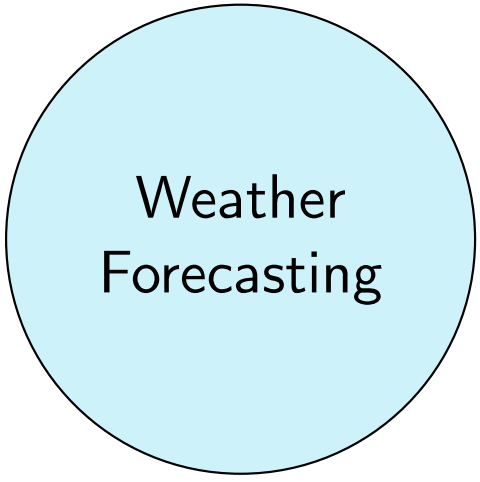
This internship was a result of a partnership between:

1

Introduction

Context





What's the importance of these themes? Why do we use Machine Learning to do these tasks? And how they are linked?

First subject: The USA Census

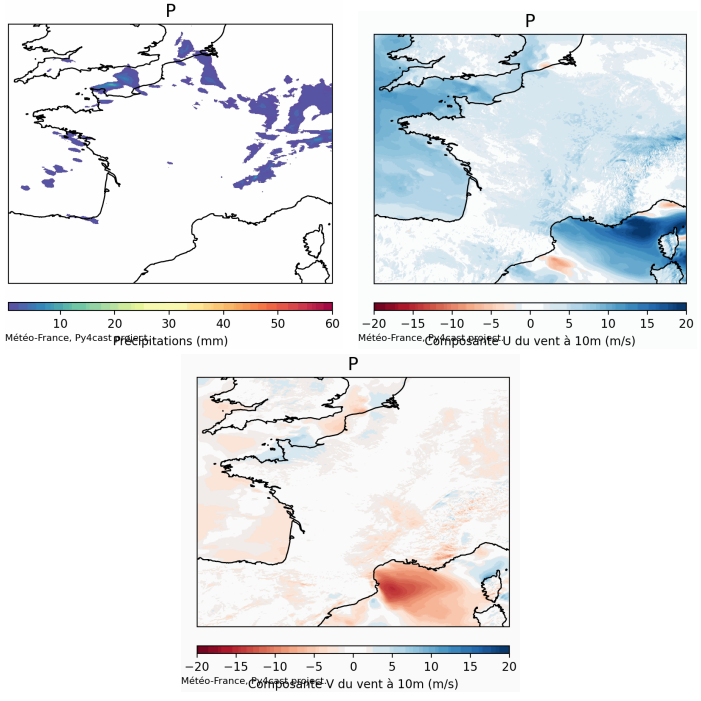
- Data of Analysis: Folktables dataset
- Application: Prediction of a person's income (higher and lower income)
- **Objective: Undestand patterns on the population, that can be used by banks (e.g. loan release)**

Table: Folktables Features

Feature	Description
AGEP	Age
COW	Worker Class
SCHL	Education
MAR	Marital Status
OCCP	Occupation
POBP	Birth Place
RELP	Relationship
WKHP	Work Hours
SEX	Sex
RAC1P	Race

Second subject: Weather forecasting

- Data of Analysis: Titan dataset (AROME and ARPÈGE images)
- Application: Prediction of the weather (variables of temperature, wind, geopotencial, humidity, ...) in the next hour
- **Objective: 100x Faster weather predictions, less computational resources, investigation of extreme climate events**



TP 0m U10 10m

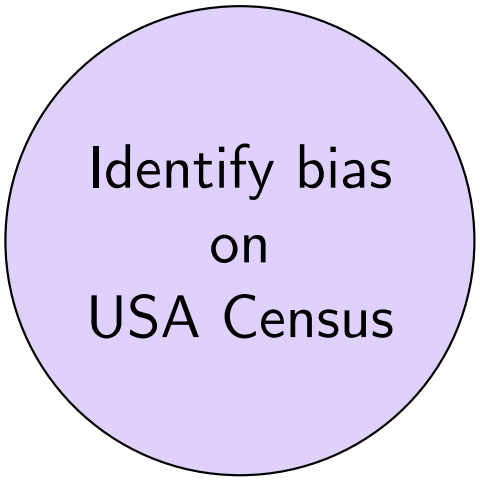
V10 10m

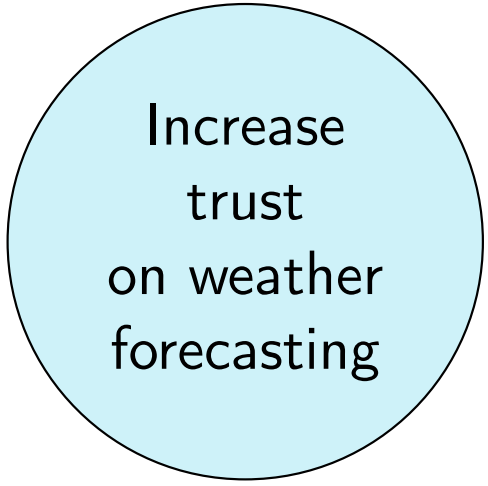
Figure: Titan image channels

What's eXplainable AI (XAI)

The XAI is a domain that gives tools to increase the user's **interpretability and trust** on the results of Artificial Intelligence (AI) models.

Critical domains require investigation...





Main ideas explored

1. Understanding the results of the AI “black boxes”
   - 1.1 The need to develop adapted eXplainable AI (XAI) techniques
   - 1.2 Verify that the model’s results are made with the "right reasons"
2. Adaptable XAI techniques (low and high dimensional data)
   - 2.1 Used for classification and regression tasks
   - 2.2 Explanations that are representable

XAI techniques

- SHAP, Lime (perturbation-based)
- **Anchors** (example-based)
- Smooth Gradient, Integrated Gradient (gradient-based)

Why exploring Anchors?

Explanations with a promise of being compact and representable, other than being a high precision result, and human understandable!

More precise for low-dimensional data, and adaptable for high-dimensional data.

Diving deeper on Anchors

Considering a binary classifier $f : \mathcal{X} \to \{0, 1\}$. For a given instance x, the objective is to find a predicate A **that explains the prediction** $f(x)$ **by identifying a minimal set of decisive features**.

A predicate A is deemed a valid anchor if it meets a precision threshold, meaning it **guarantees the model's output is consistent under local perturbation**. Precision is formally defined as:

$$\text{prec}(A) = \mathbb{E}_{D(z|A)}[\mathbf{1}_{f(x)=f(z)}] \tag{1}$$

where $D(\cdot|A)$ is the conditional distribution of inputs satisfying A. The anchor condition is satisfied with high confidence if $P(\text{prec}(A) \geq \tau) \geq 1 - \delta$.

Diving deeper on Anchors

To select the most impactful and parsimonious explanation for end-users, the method introduces a coverage measure. The coverage of A **is defined as the probability that the predicate holds under the data distribution**:

$$\text{cov}(A) = \mathbb{E}_{D(z)}[A(z)] \tag{2}$$

For a given data distribution D, parameters τ and δ, the optimal anchor is the one with maximum coverage subject to the precision constraint:

$$\arg\max_{A} \text{cov}(A) \mid P(\text{prec}(A) \geq \tau) \geq 1 - \delta \tag{3}$$

2

XAI for the USA Census

Analyse Fairness on the USA Census

- Can we find bias in the ML models?
- If so, can we use XAI to identify where's the problem?

We used the 'SEX' as the sensible variable.

Trained Models

- **Logistic Regression**
- **XGBoost** (eXtreme Gradient Boosting)
- **Hist Gradient Boosting** (Skrub's Scikit-learn implementation)
- **Simple Neural Network**

Fairness Metrics

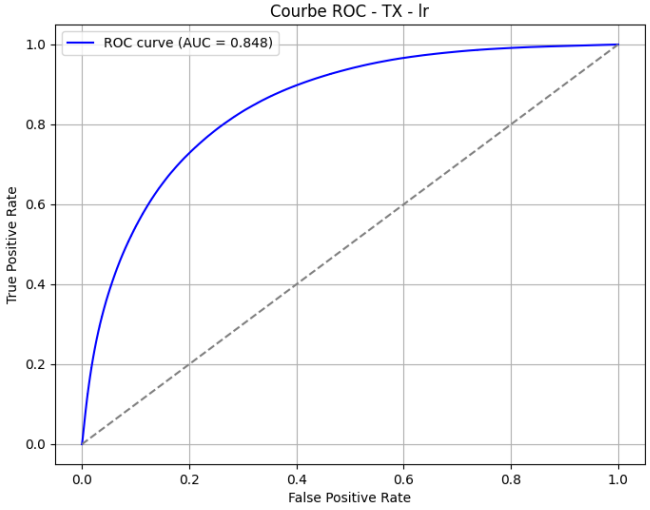
- **Accuracy**: The proportion of correct predictions (both true positives and true negatives) among all predictions.
- **Disparate Impact (DI)**: Measures the ratio between the **proportion of positive outcomes for the protected group (women) versus the privileged group (men)**. Values close to 1 indicate fairness, while values below 1 suggest bias against the protected group.
- **Equality of Odds**: Examines whether both groups have equal true positive rates and equal false positive rates. Values closer to 1 indicate better fairness.
- **Sufficiency**: Assesses whether the probability of the true outcome is the same across groups given the predicted outcome. Values closer to 1 indicate better fairness.

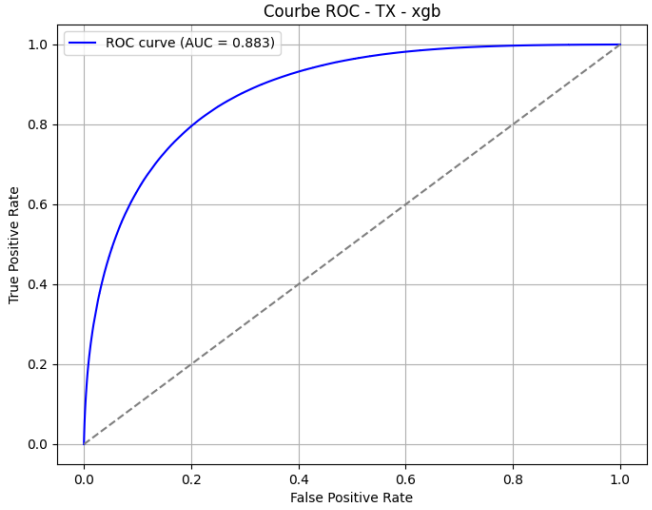
Performance and Fairness metrics

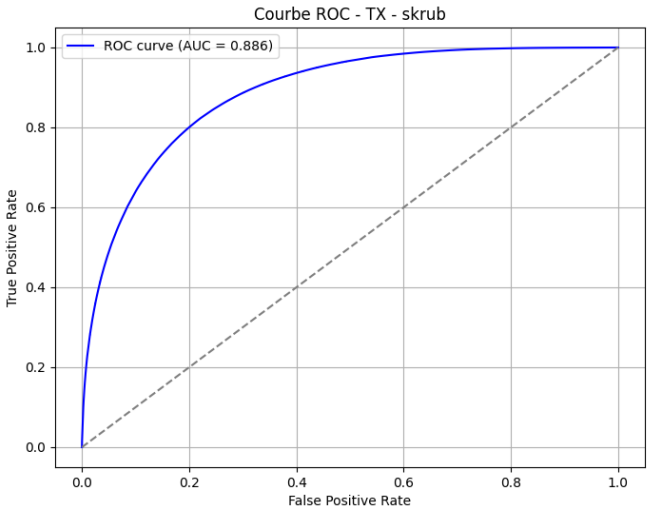
Table: Model Performance Comparison Across States

Model	Training	Testing	Accuracy	Disparate Impact	Equality of Odds	Sufficiency
Logistic Regression	CA	CA	0.56	0.67	0.84	0.95
	TX	TX	0.52	0.46	0.65	0.90
	NY	NY	0.51	0.66	0.82	0.93
XGBoost	CA	CA	0.64	0.72	0.91	0.96
	TX	TX	0.60	0.58	0.83	0.93
	NY	NY	0.61	0.75	0.92	0.96
HistGradientBoosting	CA	CA	0.63	0.71	0.90	0.94
	TX	**TX**	**0.61**	**0.54**	0.78	0.96
	NY	NY	0.60	0.68	0.89	0.97
Neural Network	CA	CA	0.52	0.88	1.03	0.85
	TX	TX	0.50	0.61	0.87	0.94
	NY	NY	0.51	0.76	0.93	0.92

Model Performance







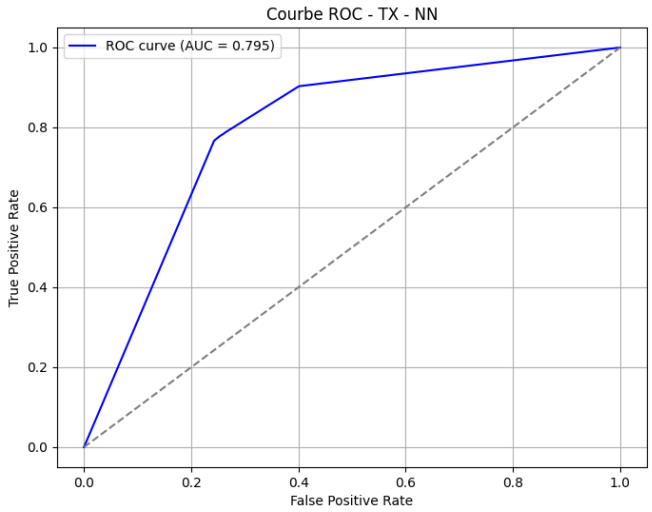


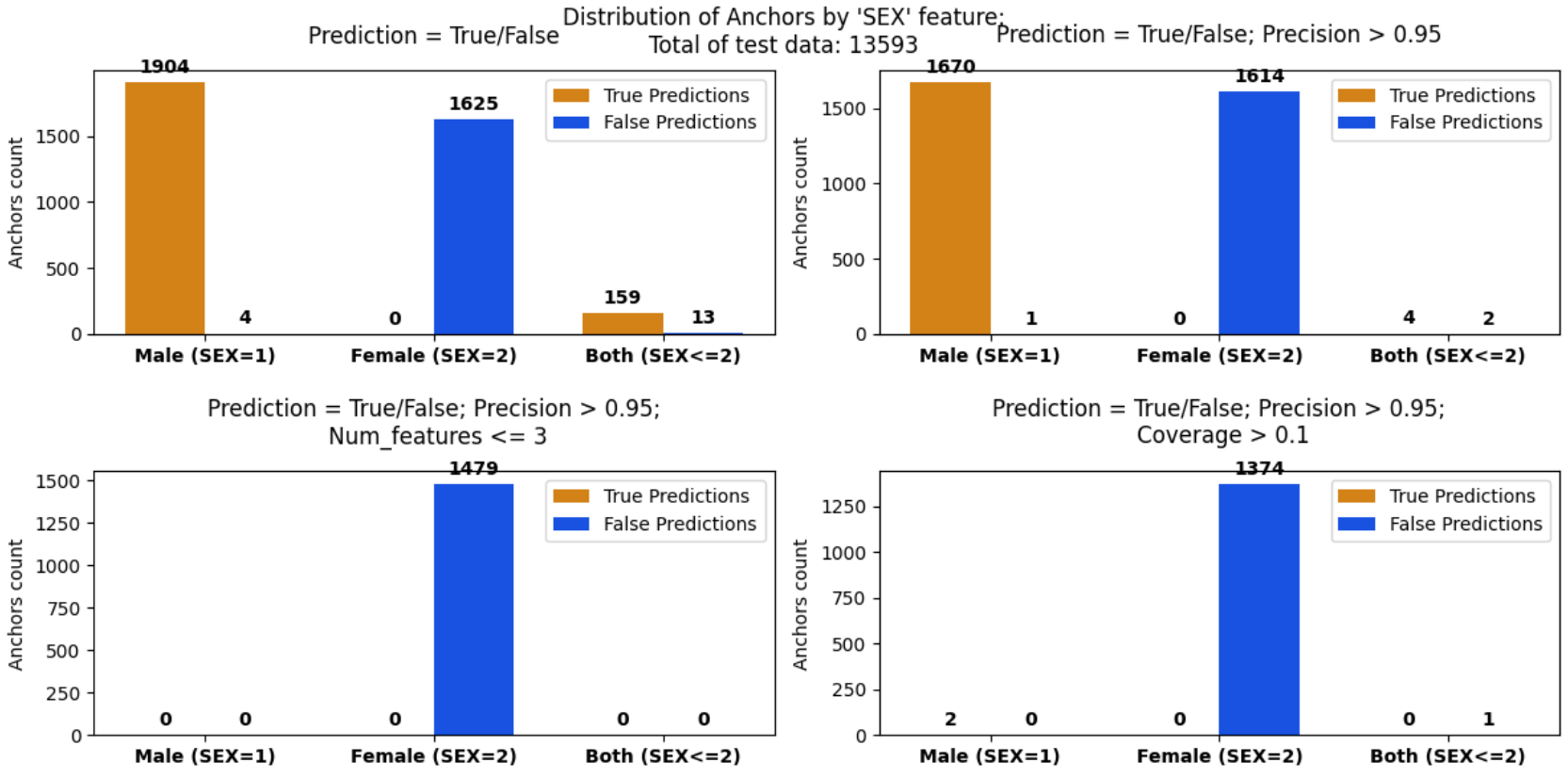
Figure: Comparison of the models trained on the sub dataset of the Texas state

ENSTA INSTITUT POLYTECHNIQUE DE PARIS

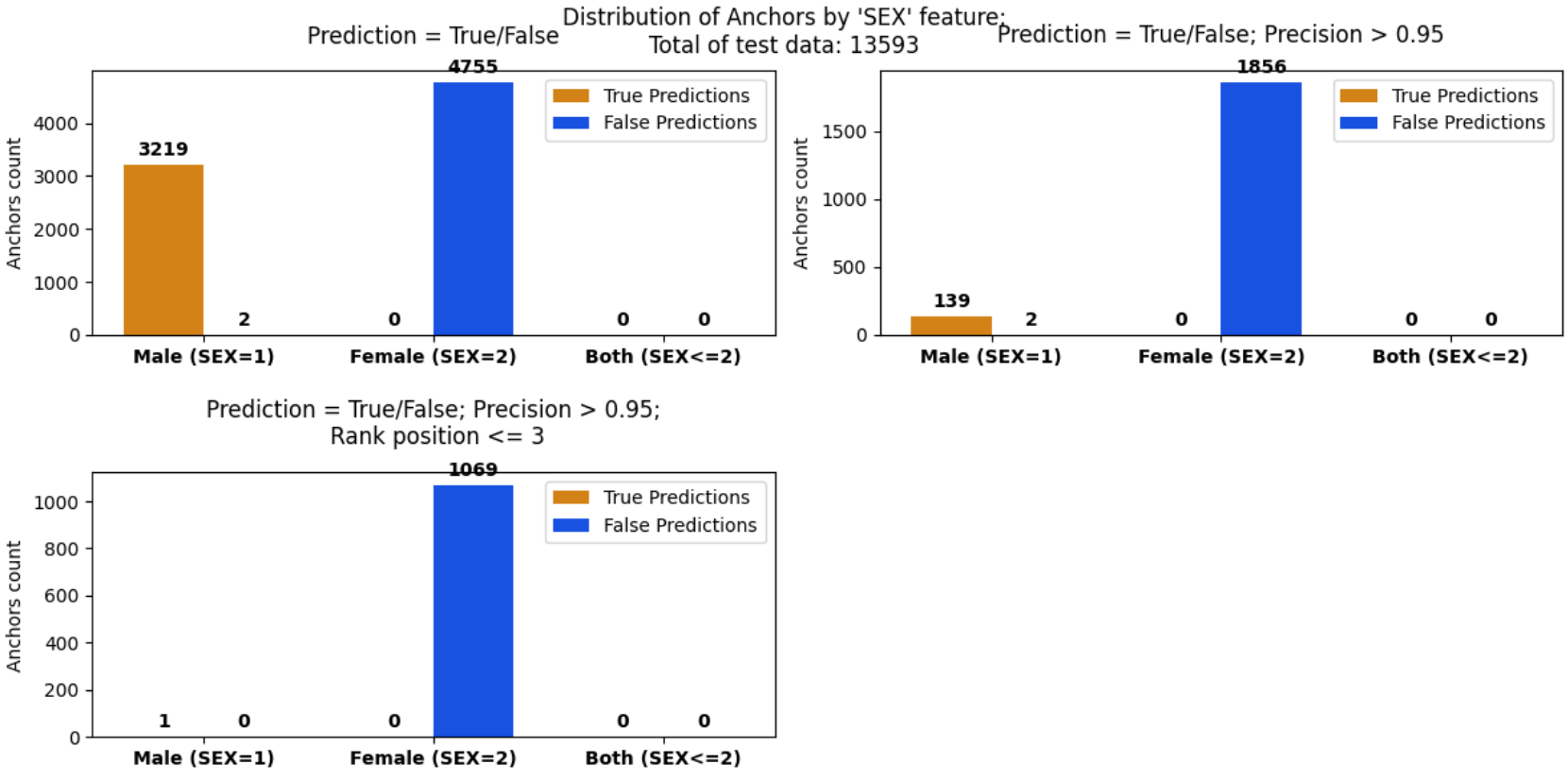
Applying XAI

- We will refer to a **gendered anchor** when the generated Anchors explanation contains 'SEX' in its list of features, which is the sensitive variable we are analysing.
- In a similar way, we will refer to **gendered SHAP values** when the generated SHAP explanation contains 'SEX' at the top of the feature ranking.

Distribution of Gendered Anchors (Anchors vs SHAP)



Distribution of Gendered SHAP values



What does it mean?

It was interesting to see how many explanations contained gendered explanations, meaning that the technique identified that feature as decisive for the prediction.

- With Anchors this implies that if the value of 'SEX' changes, the prediction changes too.
- We can also notice that the gendered explanations are divided mainly between **women with a false model prediction**, low income, and **men with a true prediction**, high income.
- **This reaffirms what the DI metric was showing:** this imbalance in the proportion of True predictions for the protected group versus the privileged group.

What can we do to go further in the investigation?

Analyse the profile of the test dataset to see how the Anchors and SHAP explanations are **organized across different groups of people**. To understand better the demographic problem in our study case.

3

XAI and Meteo

Analyse truthfulness of Results

- Can we explain the ML models decisions on a human understandable way?
- If so, can we use XAI to identify relations between regions, how they influence each other?

We used the Titan data from November 2023 and the UNetR++ convolutional neural network

The Titan data

21 channels of AROME images + 16 channels of ARPÈGE images

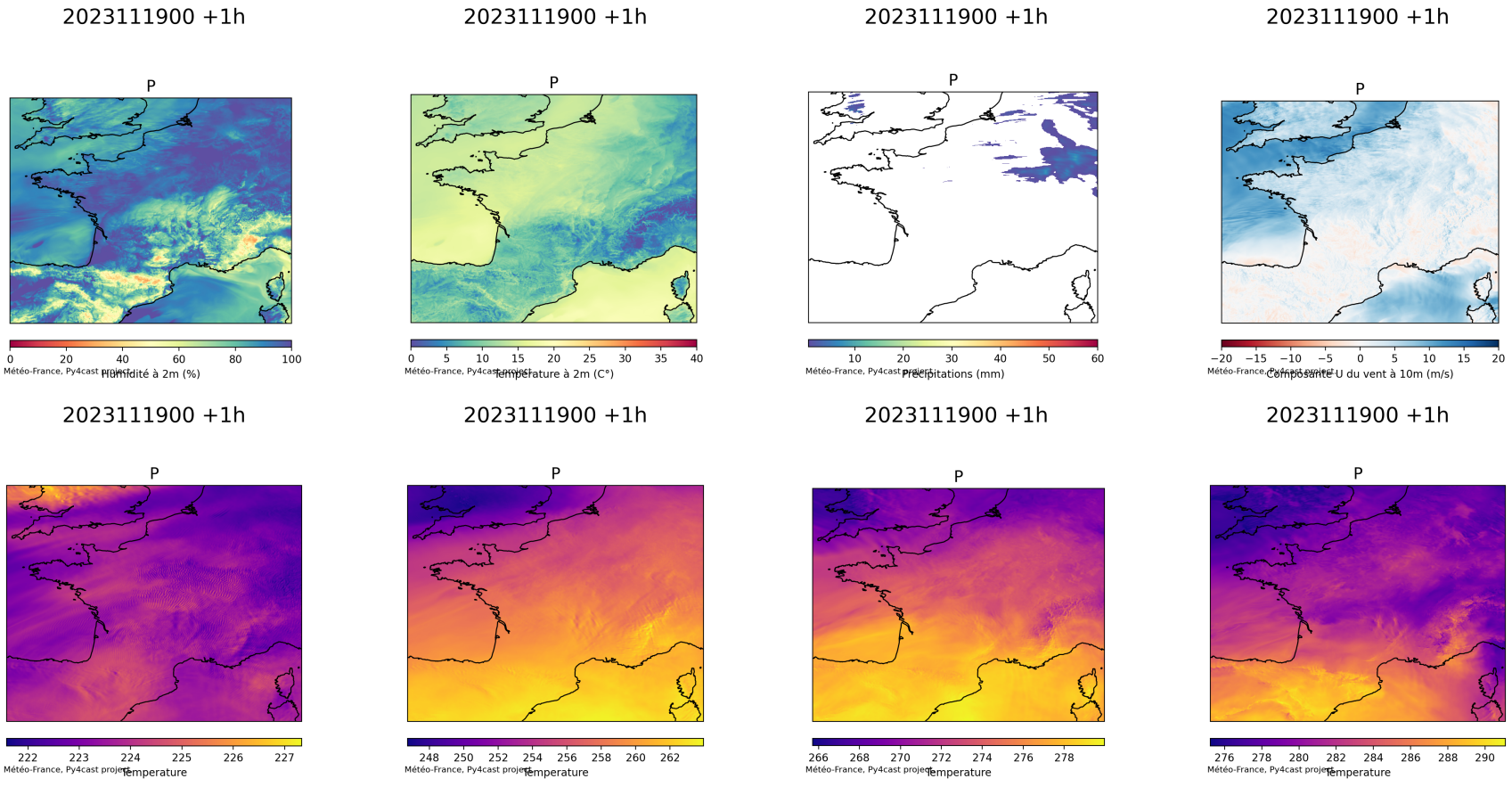


Figure: Titan image channels

Combining AROME and ARPÈGE

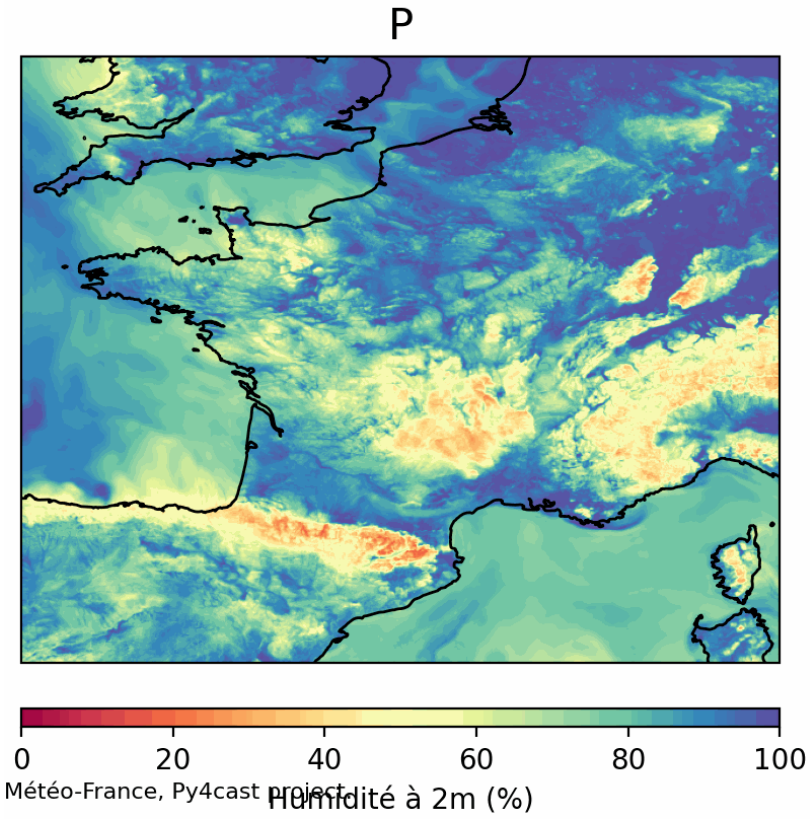
Let t be the present time and $t+1$ be a future time.

- Training without boundary conditions:
  - Input: aro_t (AROME at time t)
  - Output: $aro_{t+1} - aro_t$ (The change in AROME from t to $t+1$)
- Training with boundary conditions:
  - Input: $[aro_t, arp_t]$ (AROME and ARPÈGE at time t)
  - Output: $aro_{t+1} - aro_t$ (The change in AROME from t to $t+1$)

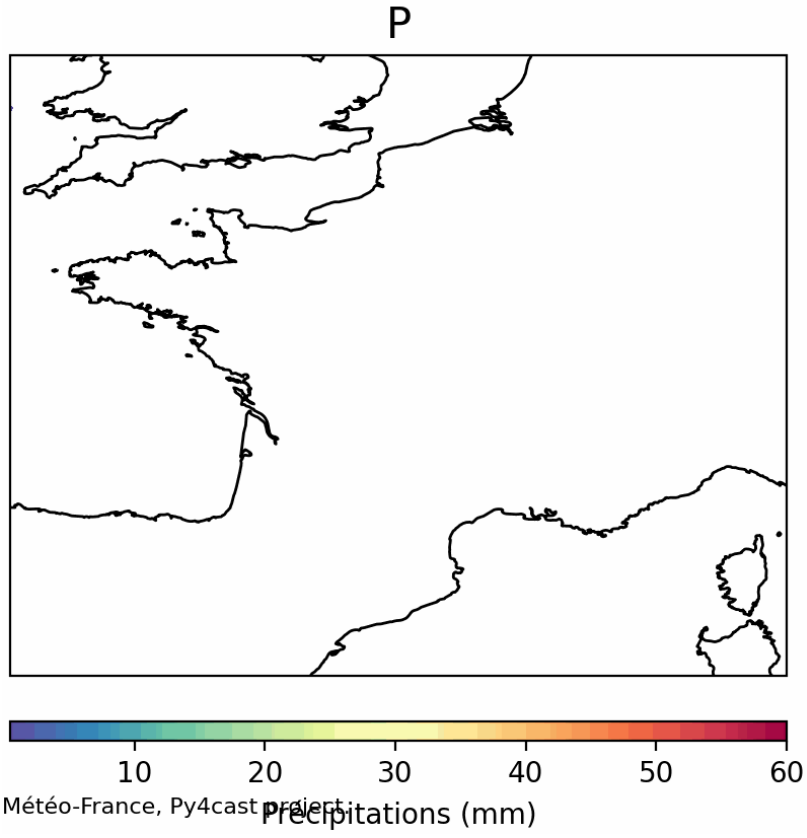
In summary, the data can be combined in two primary ways: a simpler approach using only AROME data, and a more precise, operational-style approach that leverages **ARPÈGE data to provide essential boundary conditions**.

Py4cast Predictions

2023021500 +1h



2023021500 +1h



Applying Anchors on Weather Forecasting

We elaborated two extensions of the Anchors formalization, which will be particularly useful for the weather forecasting problem.

- Deterministic Precision Constraint:

$$\arg\max_{A} \text{cov}(A) \mid \text{prec}(A) \geq \tau \tag{4}$$

- Extension to Regression and Probabilistic Output:

$$\text{prec}(A) = \mathbb{E}_{D(z|A)}[\mathbf{1}_{|f(x)-f(z)|>\tau}] \tag{5}$$

where τ is a threshold above which the conditions of A are supposed to have a significant impact.

Generating data perturbations

We suppose that $g_{i,j,c,R}(f(x_t)) == 1$, given a region (i, j) in the map, and we want to explain why using anchors. We will sample K different counterfactual observations z_k, for $k \in 1, \ldots, K$.

- **Center of Perturbation:** $(\widehat{i_k}, \widehat{j_k})$
- **Channel of Perturbation:** $\widehat{c_k}$
- **Perturbation Value:** $\widehat{v_k}$

Perturbation Mask

We introduce a second hyperparameter, σ_e, which models the spatial extent of the perturbation. The procedure for generating a counterfactual instance z_k is then as follows:

- **Unaffected Channels:** For all channels $c \neq \widehat{c_k}$, the data remains unchanged: $z_k(:,:,c) = x_t(:,:,c)$.
- **Mask Creation:** A mask $m \in \mathbb{R}^{I \times J}$ is generated. Its values are initially drawn from a 2D Gaussian distribution centered at $(\widehat{i_k}, \widehat{j_k})$ with covariance matrix $\begin{bmatrix} \sigma_e & 0 \\ 0 & \sigma_e \end{bmatrix}$. These values are then linearly rescaled to the interval $[0, 1]$.
- **Application:** The perturbed channel $\widehat{c_k}$ in z_k is created by a mask-based blending between the original values and the new value $\widehat{v_k}$:

$$z_k(\bar{i}, \bar{j}, \widehat{c_k}) = m(\bar{i}, \bar{j}) * \widehat{v_k} + (1 - m(\bar{i}, \bar{j})) * x_t(\bar{i}, \bar{j}, \widehat{c_k}) \qquad (6)$$

for all $\bar{i} \in 1, \ldots, I$ and $\bar{j} \in 1, \ldots, J$.

Example of a Perturbed Image

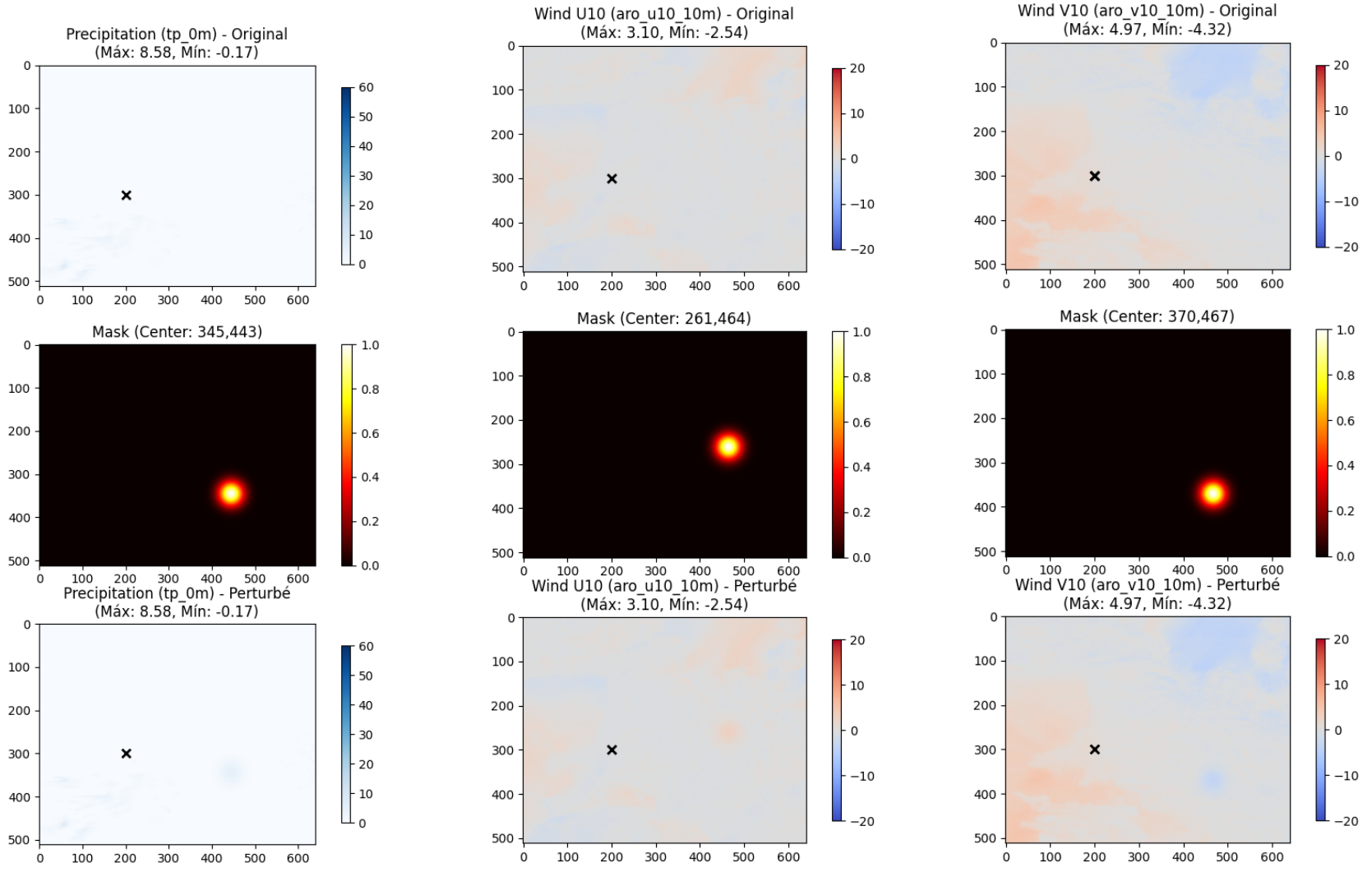


Figure: Example of counterfactuals of Titan image channels from 18/11/2023

Running the Regression Anchors

We generated about 300 conterfactuals per channel, in a radius of about **300 km from the target point**.

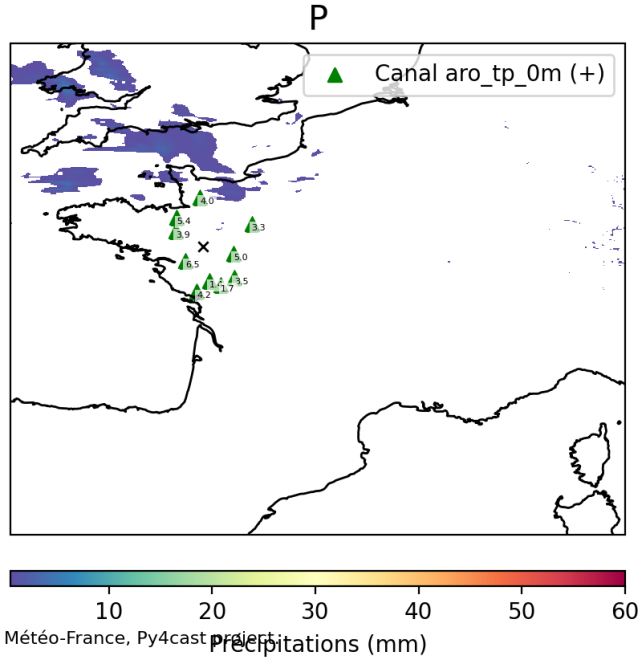
Using a **mask of 40 km** for the perturbations.

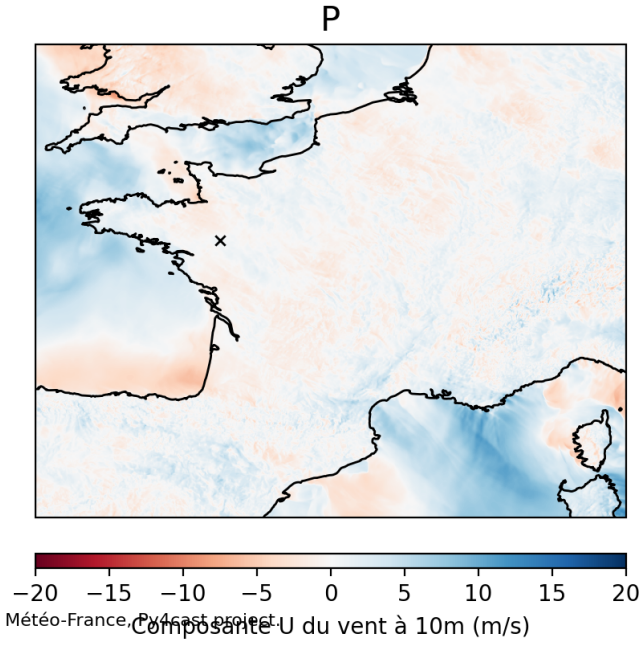
Seeing if the **rain prediction changed** in more than 5%.

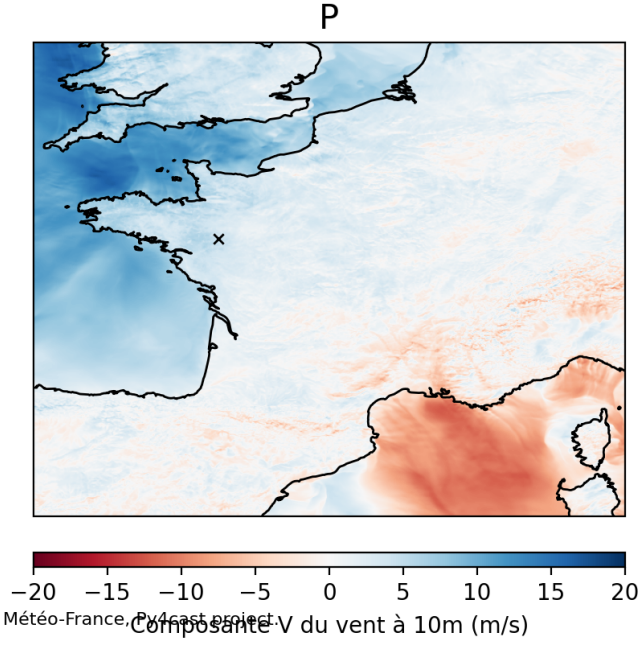
Rain Model

- *aro_tp_0m* (Total Precipitation)
- *aro_u10_10m* (Zonal wind component at 10m)
- *aro_v10_10m* (Meridional wind component at 10m)

Region of Chateaubriant on 18th November

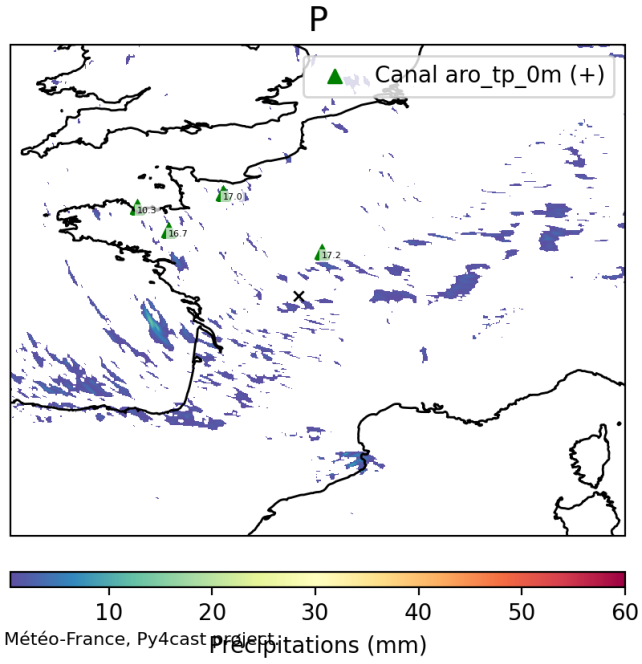


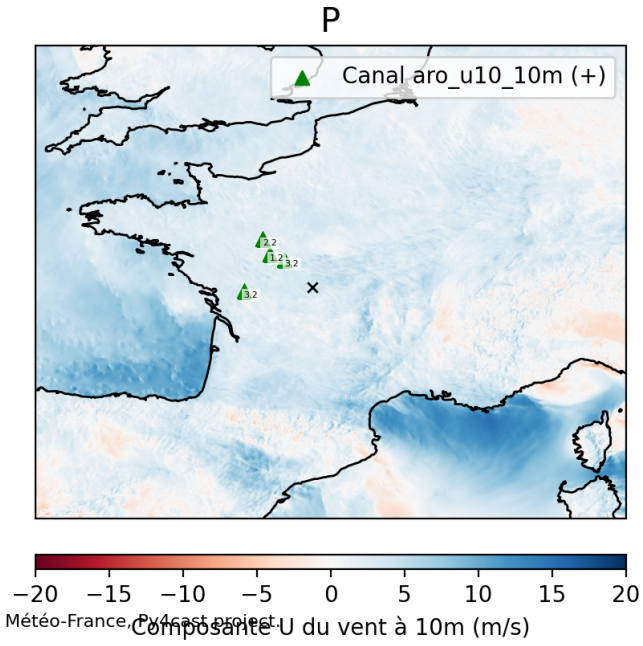


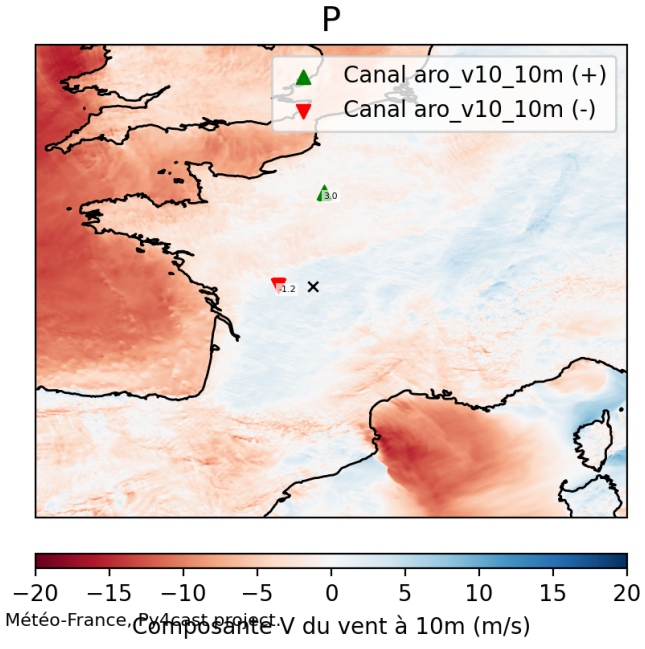


ENSTA INSTITUT POLYTECHNIQUE DE PARIS

Region of Limoges on 21st November







ENSTA INSTITUT POLYTECHNIQUE DE PARIS

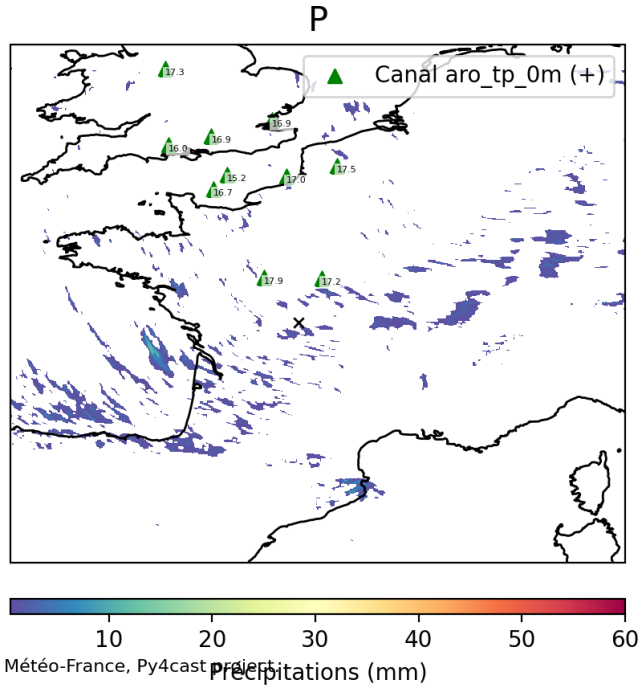
4

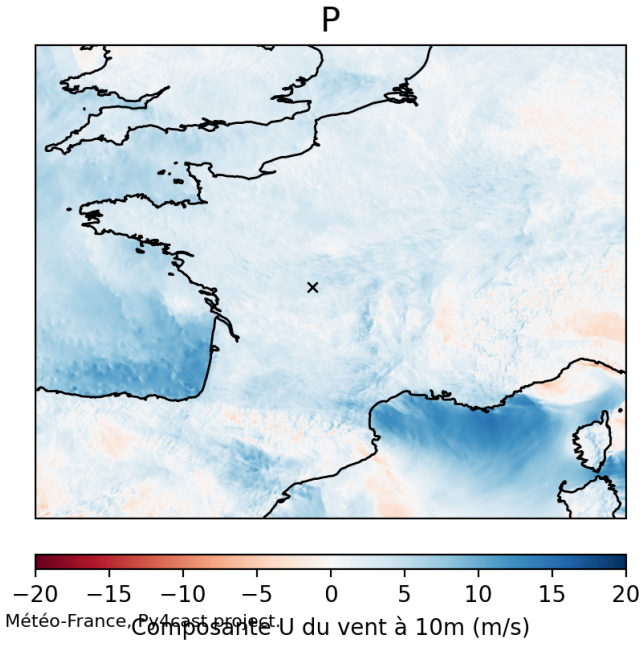
Results

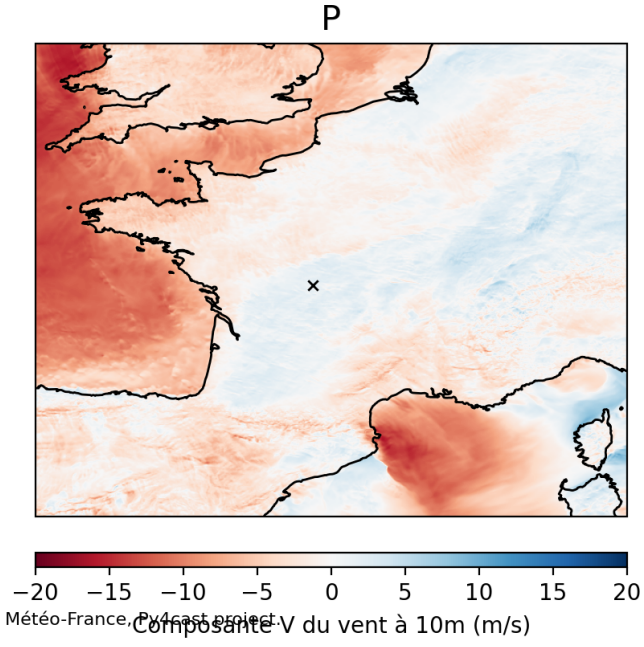
Different Models to Compare

1. **Rain Model (No Boundaries):** Trained exclusively on the three AROME rain channels, as presented in the previous results.
2. **Rain Model (With Boundaries):** Trained on the three AROME rain channels augmented with all ARPÈGE global fields as boundary conditions.
3. **Full Model (No Boundaries):** Trained on the complete set of 21 AROME channels, without boundary conditions.
4. **Full Model (With Boundaries):** Trained on the full suite of AROME channels combined with all ARPÈGE global fields to provide boundary conditions.

Rain Model With Boundaries

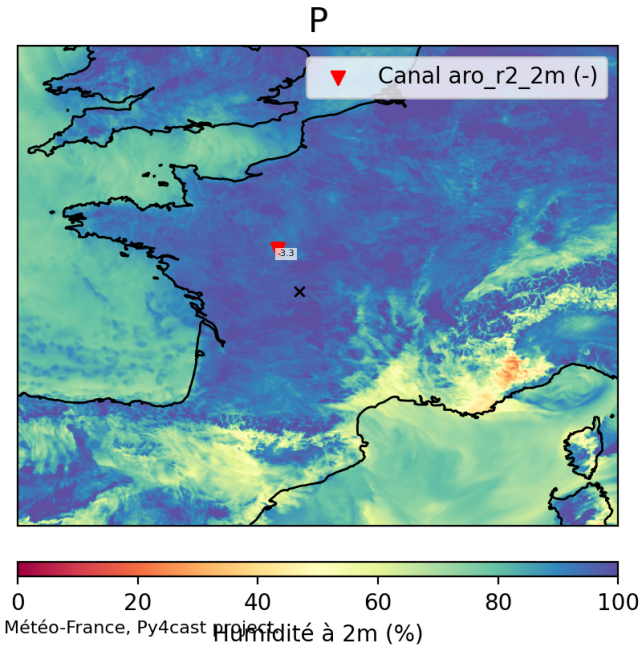


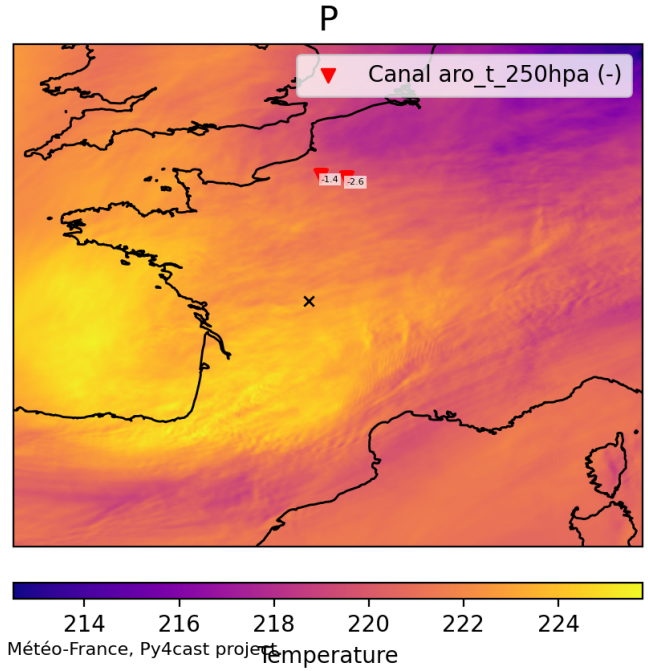


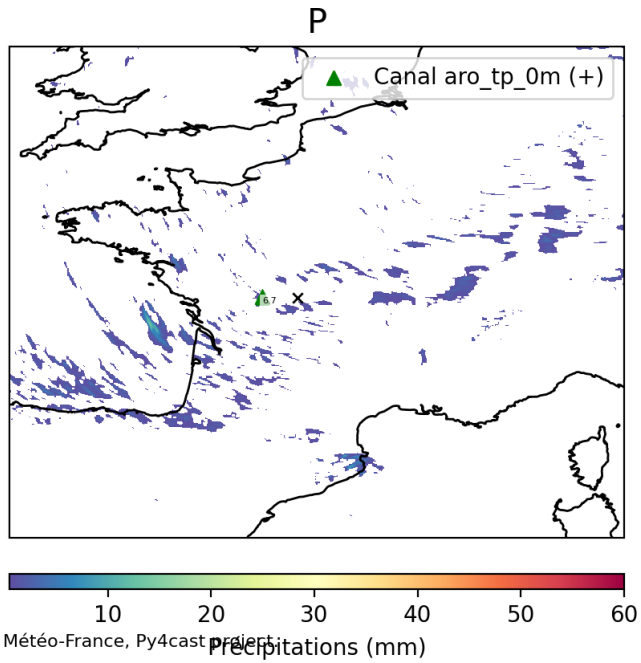


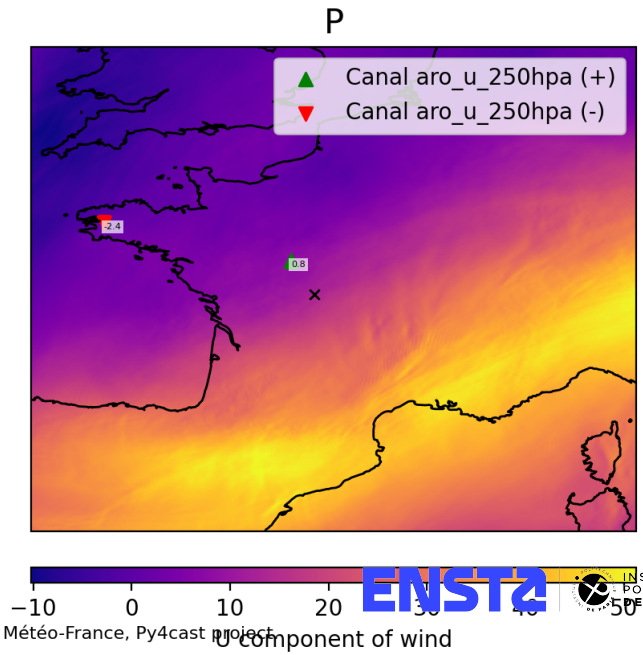
ENSTA INSTITUT POLYTECHNIQUE DE PARIS

Full Model Without Boundaries



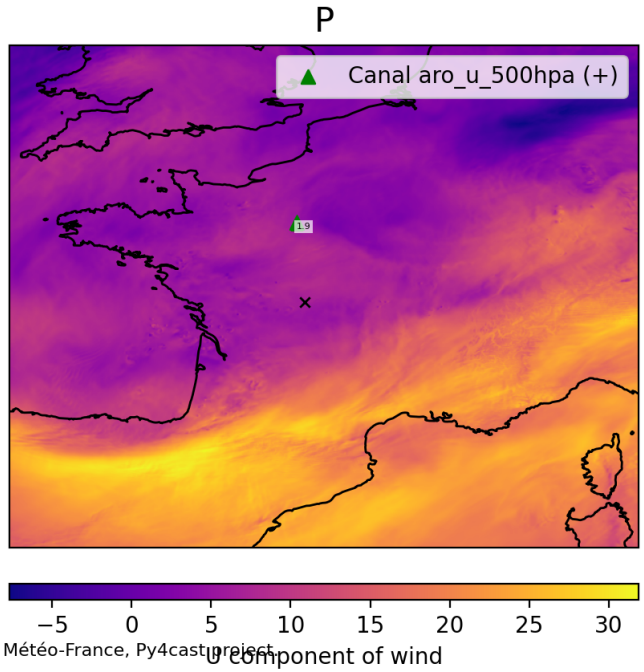


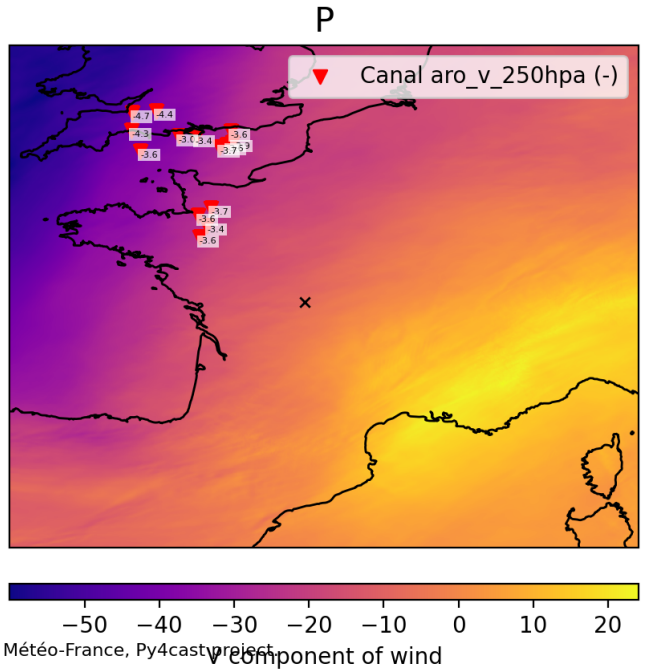


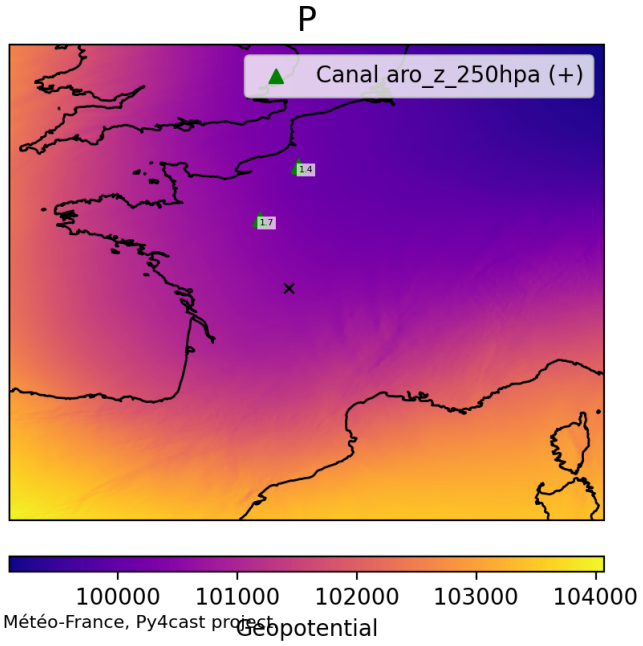


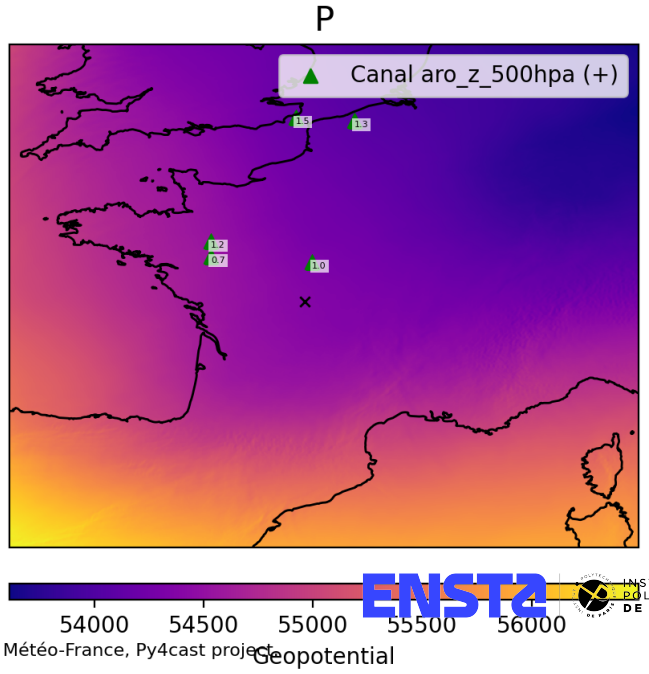
ENSTA INSTITUT POLYTECHNIQUE DE PARIS

Full Model Without Boundaries



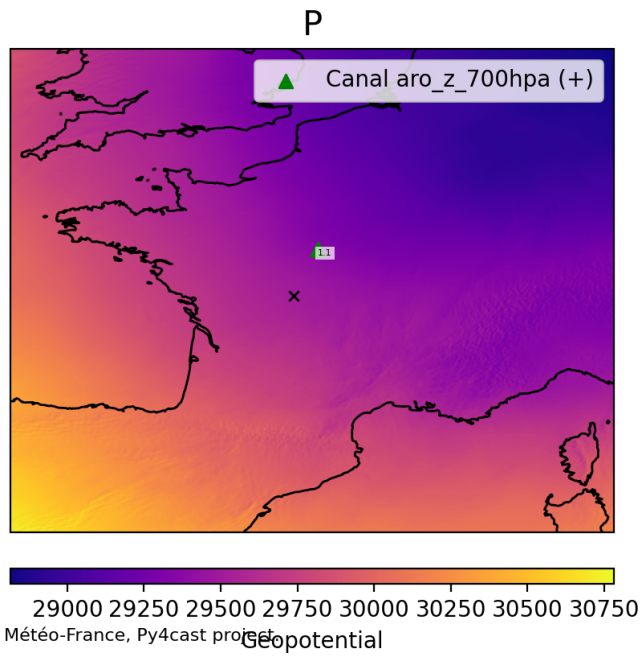


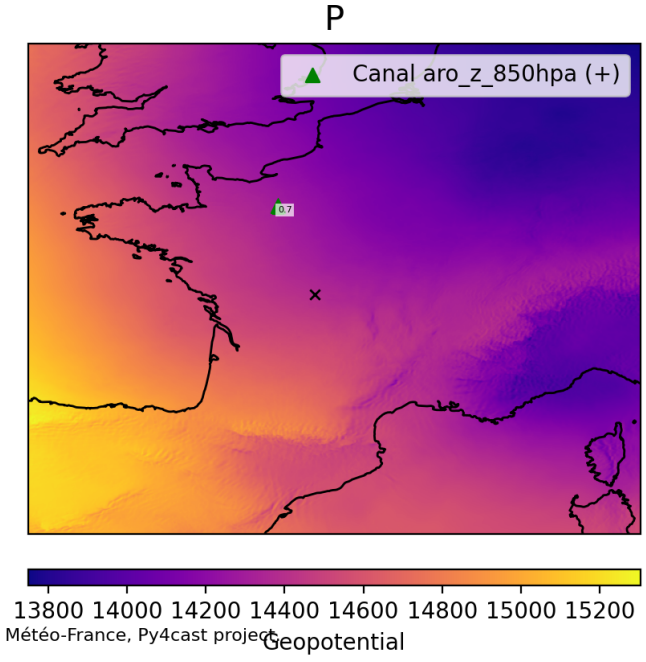




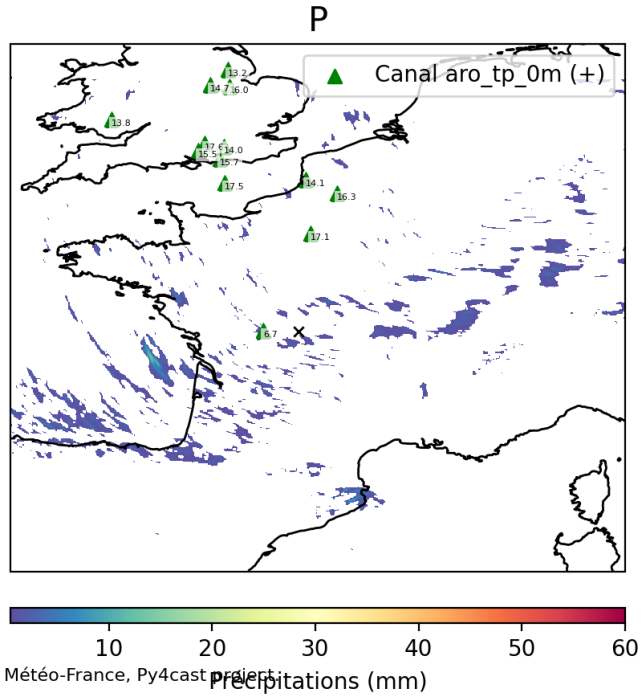
ENSTA INSTITUT POLYTECHNIQUE DE PARIS

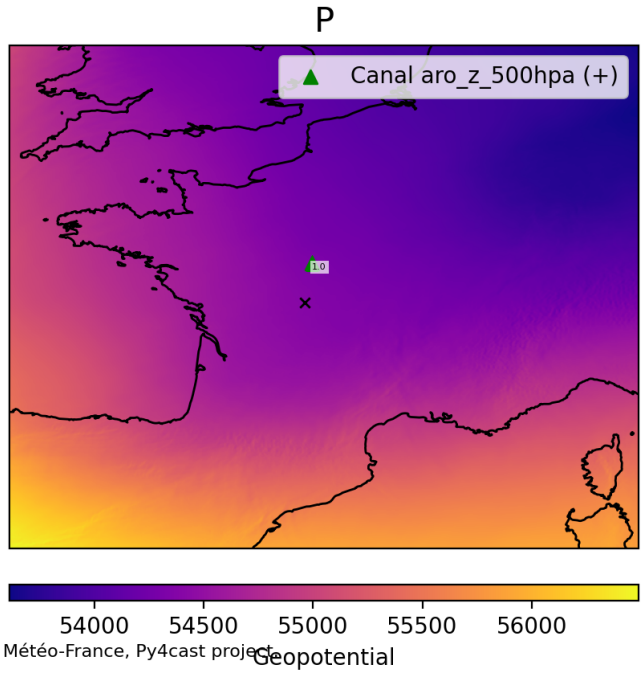
Full Model Without Boundaries

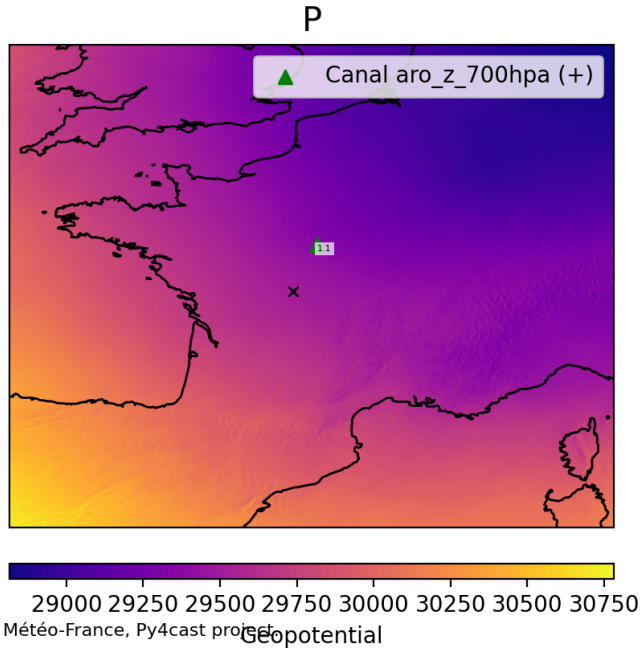




Full Model With Boundaries







ENSTA INSTITUT POLYTECHNIQUE DE PARIS

5

Discussion

Some remarks on the Regression Anchors

The results for the Full Model without Boundaries can be interpreted by considering two factors: the number of anchors per channel (indicating influence) and the magnitude of their effect on the prediction.

- Full Model with Boundaries reveals that the model trained with ARPÈGE data produced **more conservative and robust explanations**.
- It generated anchors in **fewer channels and with a lower density per channel**.
- This increased selectivity may be due to the robustness introduced by the **global context provided by the ARPÈGE boundary conditions**.
- Despite this difference, there is an intersection between the anchors found in both models, indicating agreement on the importance of certain channels and locations.

6

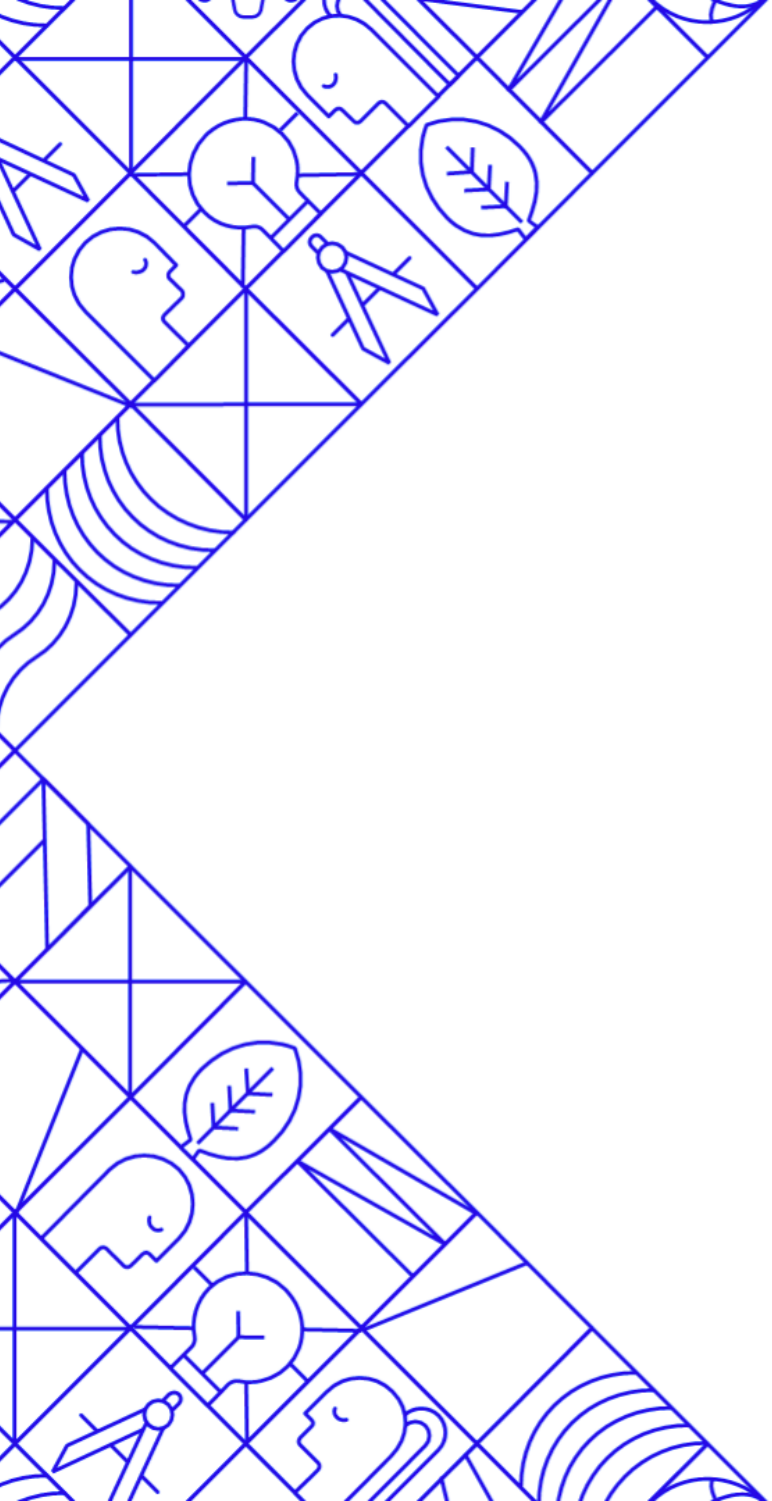
Conclusion

Insights on the Meteo

Furthermore, the method demonstrated potential to:

- Understand the constraints used by the model for the prediction
- Proactive vulnerability mapping
- Highlighting surrounding areas that **highly influence a target location**, Anchors can reveal geographic points of sensitivity

That lead us to identify zones where extreme weather events elsewhere could have a critical downstream impact, thereby improving risk assessment and resource allocation.



Merci!