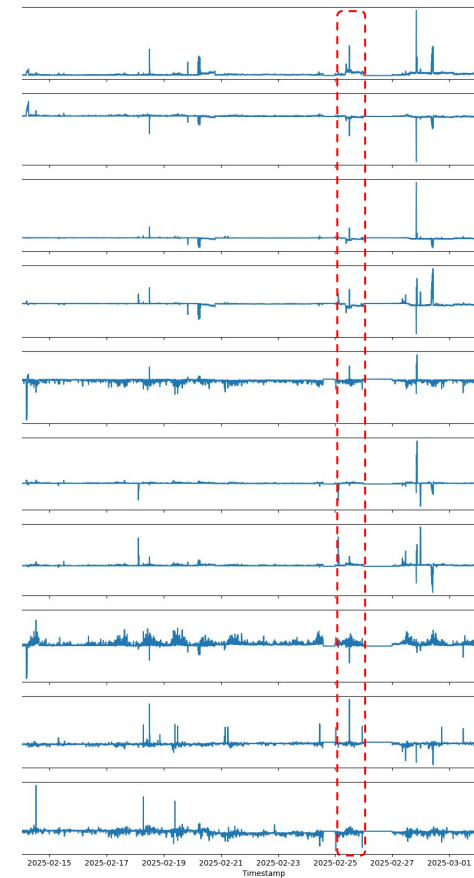
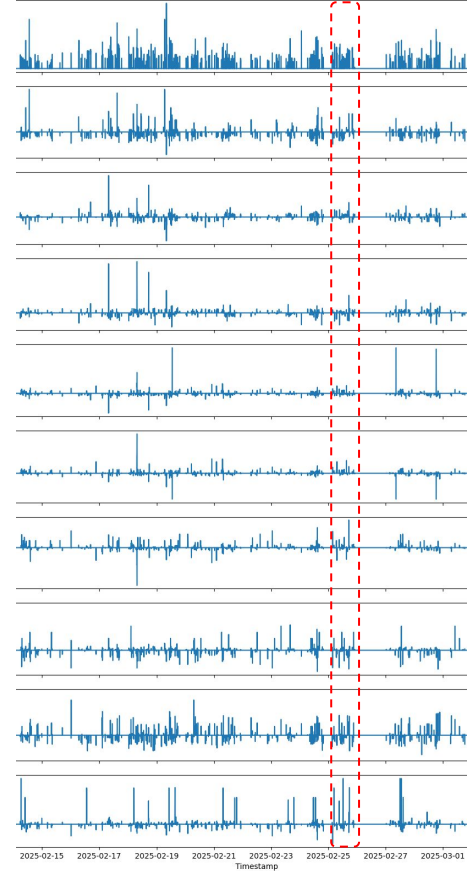
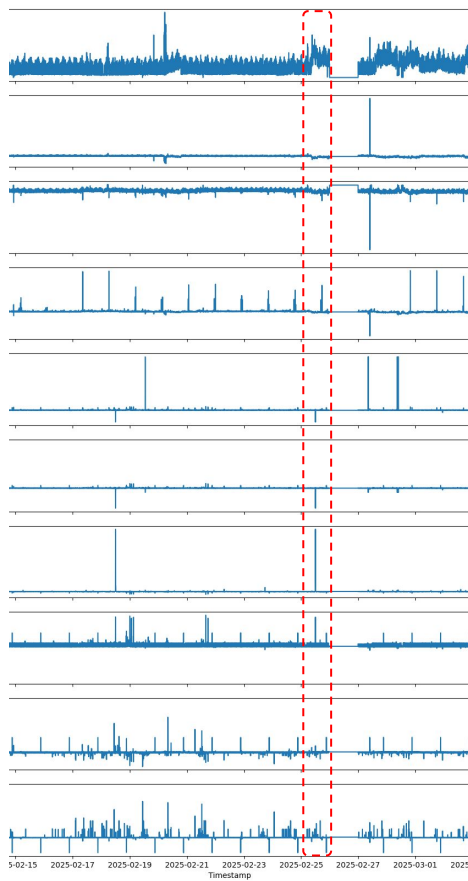


Improving Fault Diagnosis in Complex Systems: Investigating xAI Techniques for Abnormal Sensor Isolation

Dr. Baptiste Lambert
AI4Aerospace Workshop - 19.05.2025 - Toulouse



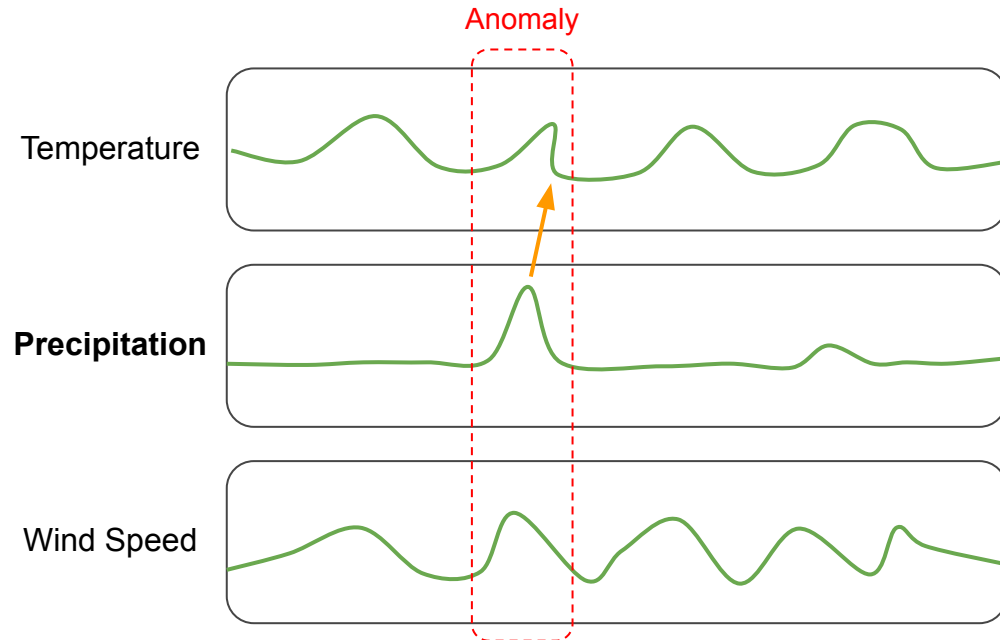
Motivations



Which feature(s) out of the 30 is failing in the red box?

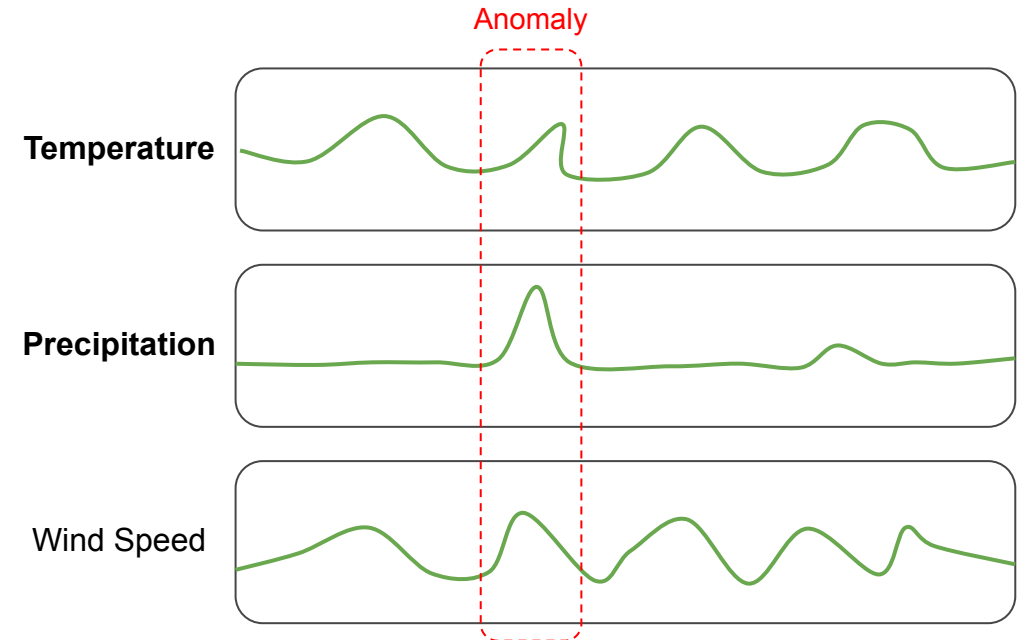
Root Cause Analysis vs Explainable AI (xAI)

Root Cause Analysis



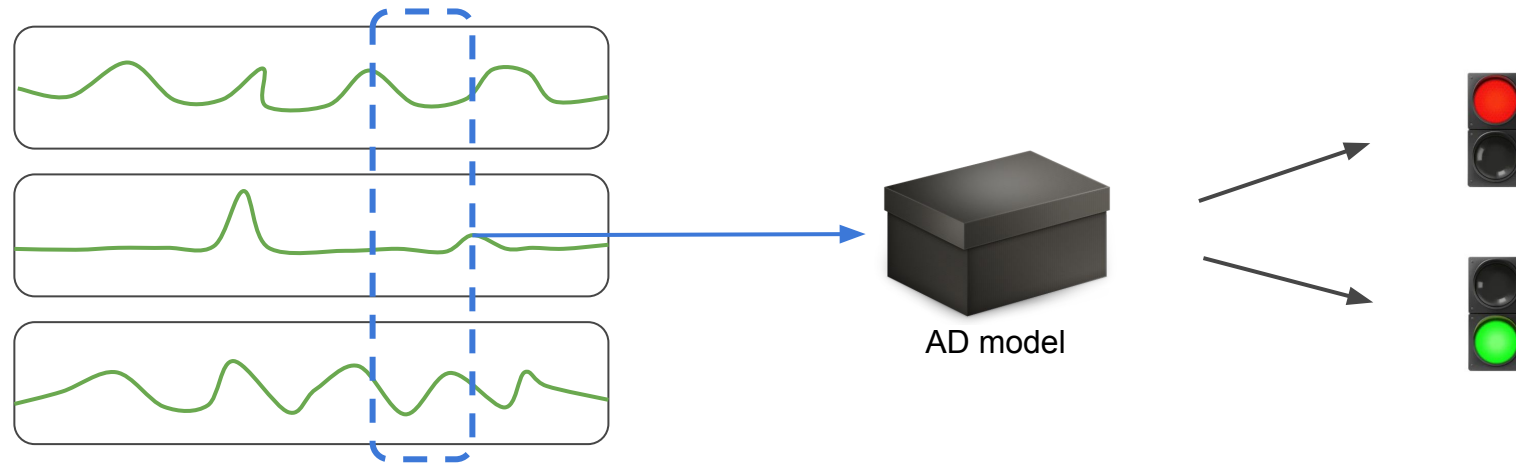
- Find the underlying reasons of the anomaly,
- Root cause(s) of the anomaly can be outside the window detected by the AD model.

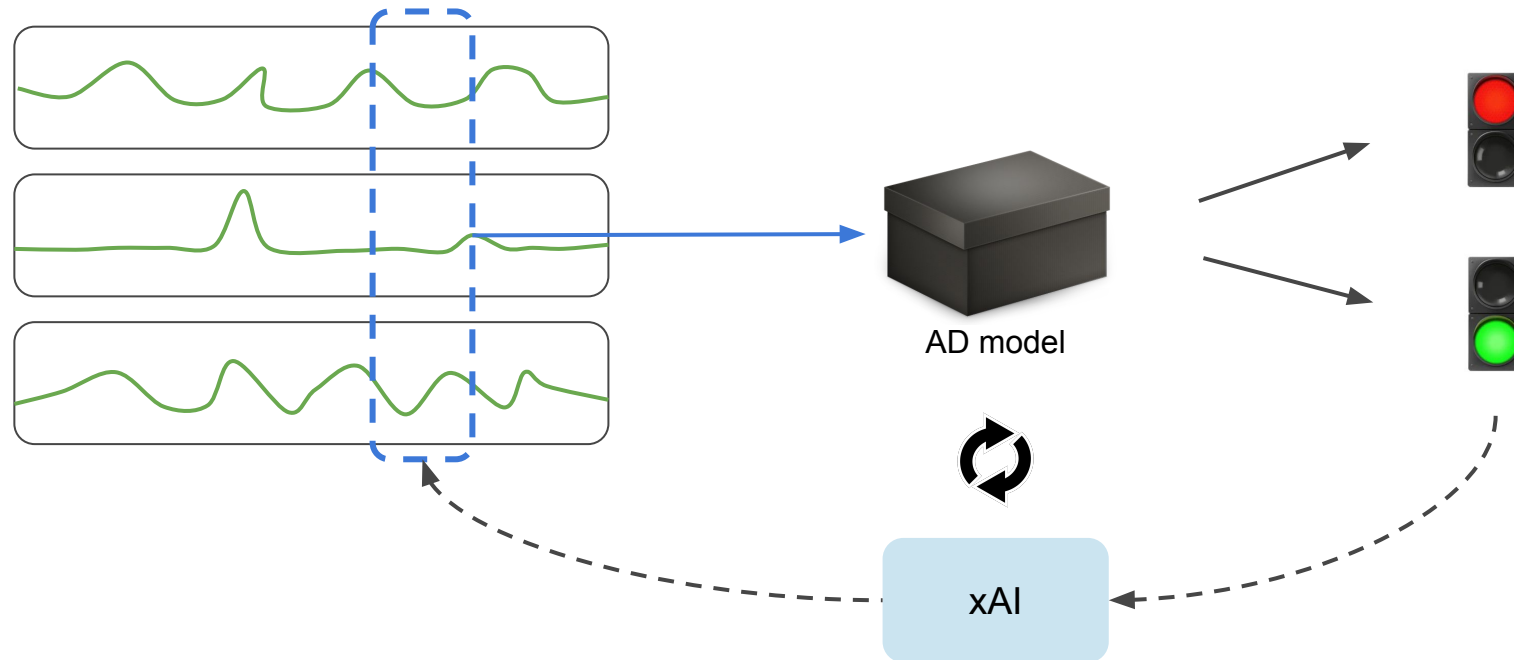
AD-xAI Analysis



- Find the variables that influence the AD model prediction at the anomaly,
- Relationship between variables is not considered.

Workflow



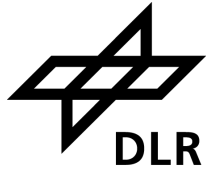


Benchmark challenges:

- Labeled dataset: anomaly and affected variables,
- Explanation quality metrics,
- Unsupervised and model agnostic xAI methods.

Dataset

Dataset: Atacama Desert Weather



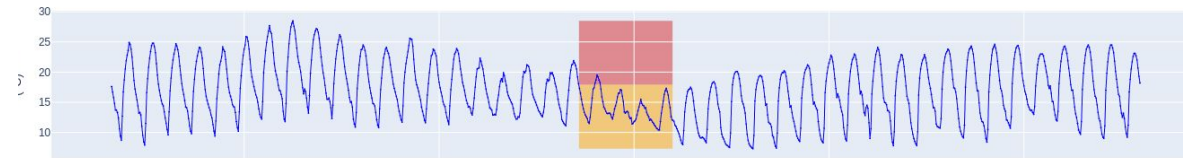
Dataset:

- Weather data from 01.01.2000 to 01.01.2025,
- Hourly averaged variables: temperature, wind speed, pressure, relative humidity, cloud cover, precipitation,
- Anomaly if precipitation $> 2.0\text{mm/h}$ (5-7% yearly avg.) $\rightarrow$ 7 instances,
- Focused on 3 recorded flash floods:
 - 23.03.15 - 26.03.15,
 - 09.08.15,
 - 15.03.22 - 17.03.22.
- Anomaly can be primarily explained by high precipitation.

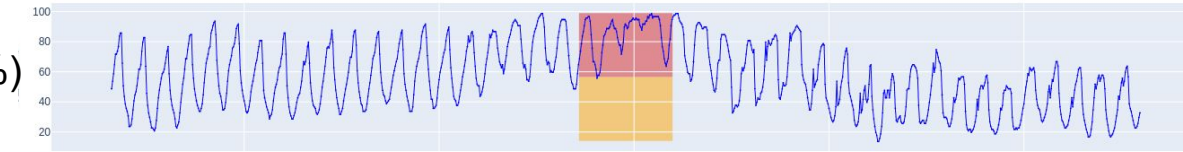


Example of Anomaly: 23.03.15→26.03.15

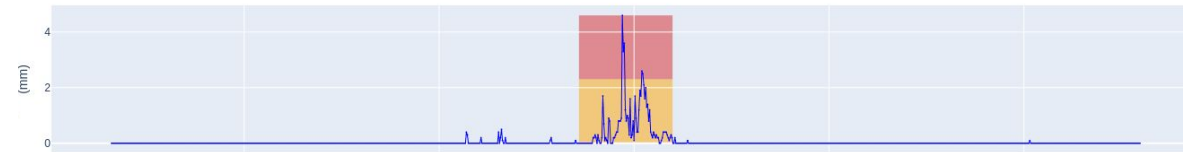
Temperature (°C)



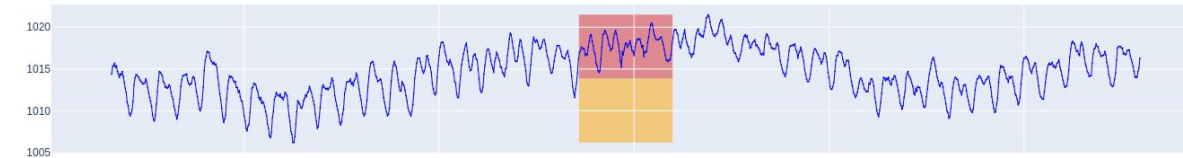
Relative Humidity (%)



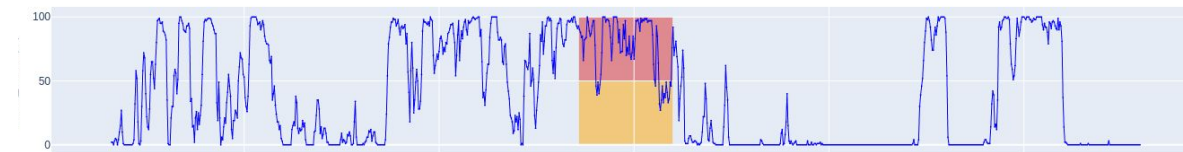
Precipitation (mm)



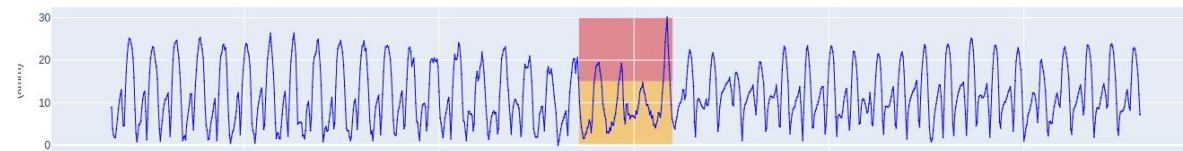
Pressure (hPa)



Cloud Cover (%)



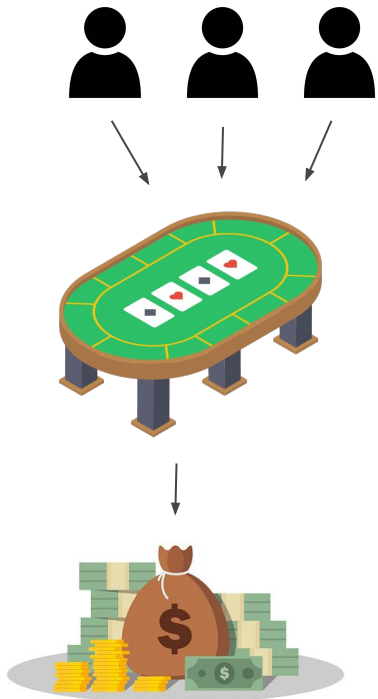
Wind Speed (km/h)



xAI Methods

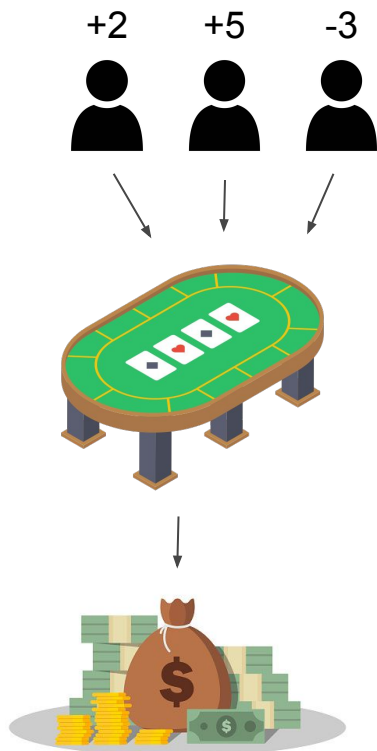
xAI: SHapley Additive exPlanations (SHAP)

- Explanation based on Shapley values (game theory) and feature permutation.



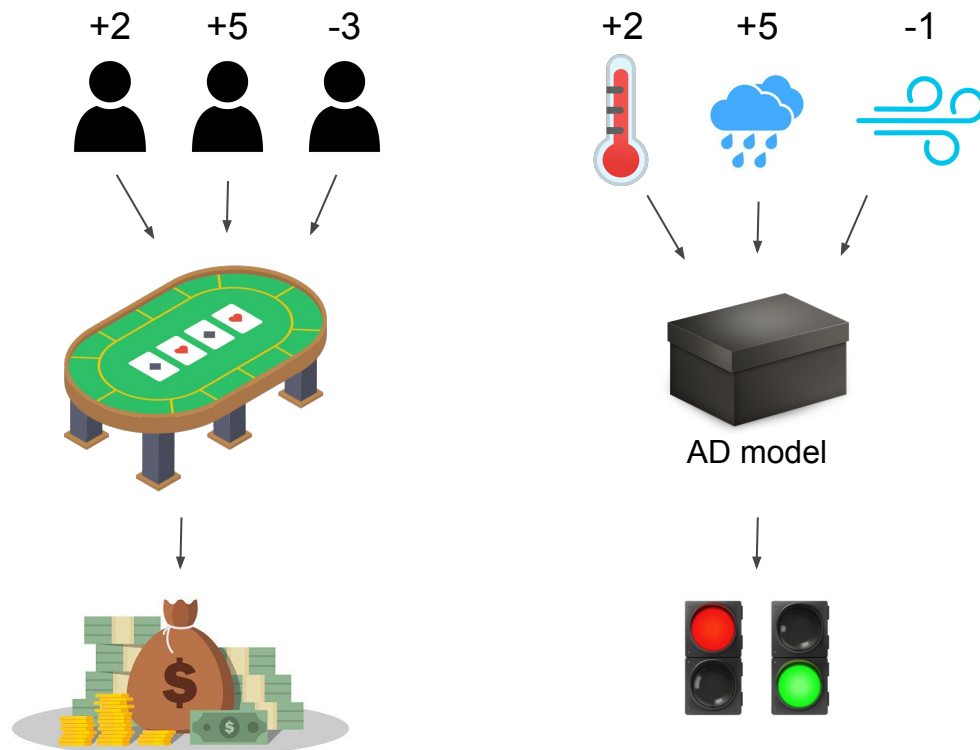
xAI: SHapley Additive exPlanations (SHAP)

- Explanation based on Shapley values (game theory) and feature permutation.



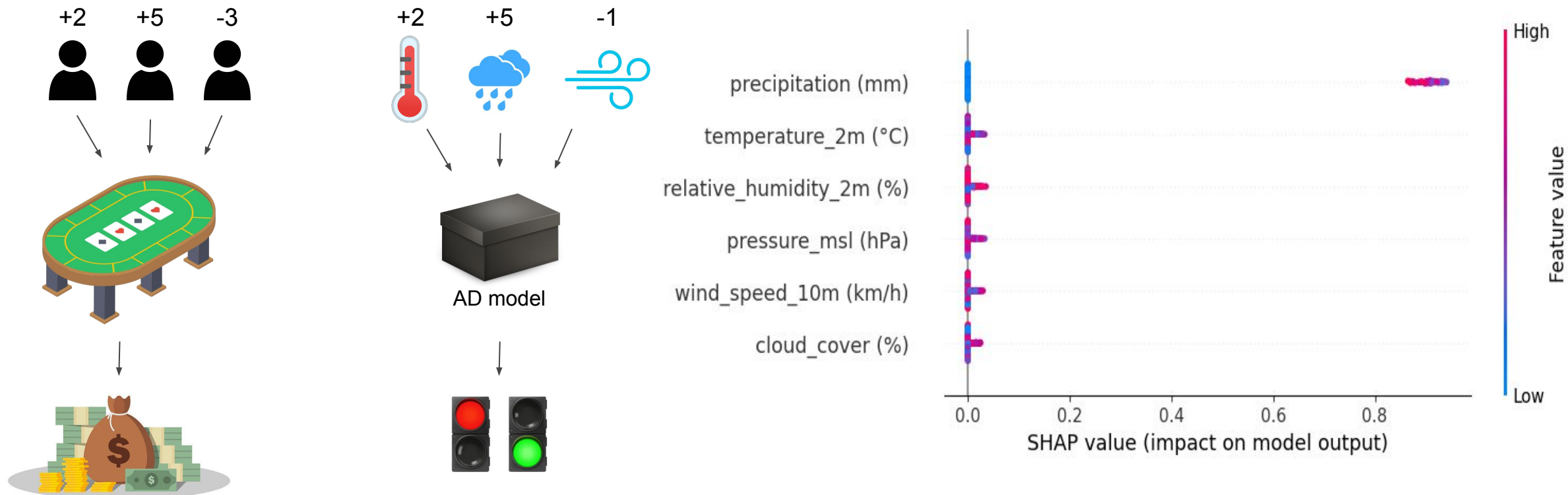
xAI: SHapley Additive exPlanations (SHAP)

- Explanation based on Shapley values (game theory) and feature permutation.



xAI: SHapley Additive exPlanations (SHAP)

- Explanation based on Shapley values (game theory) and feature permutation.
- Two approaches were investigated:
 - Standard SHAP: SHAP apply on each observation,
 - WindowSHAP: time series optimized SHAP.



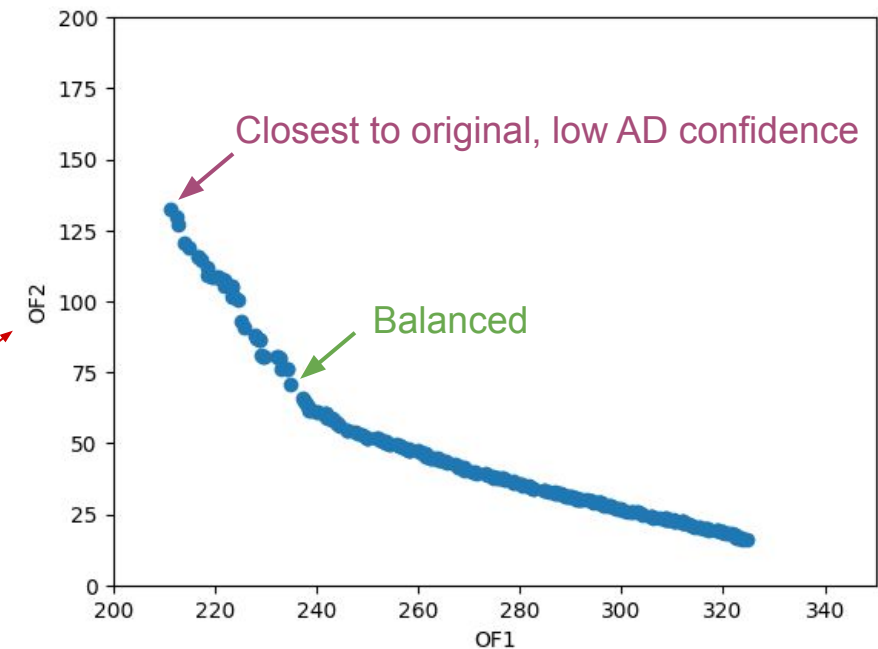
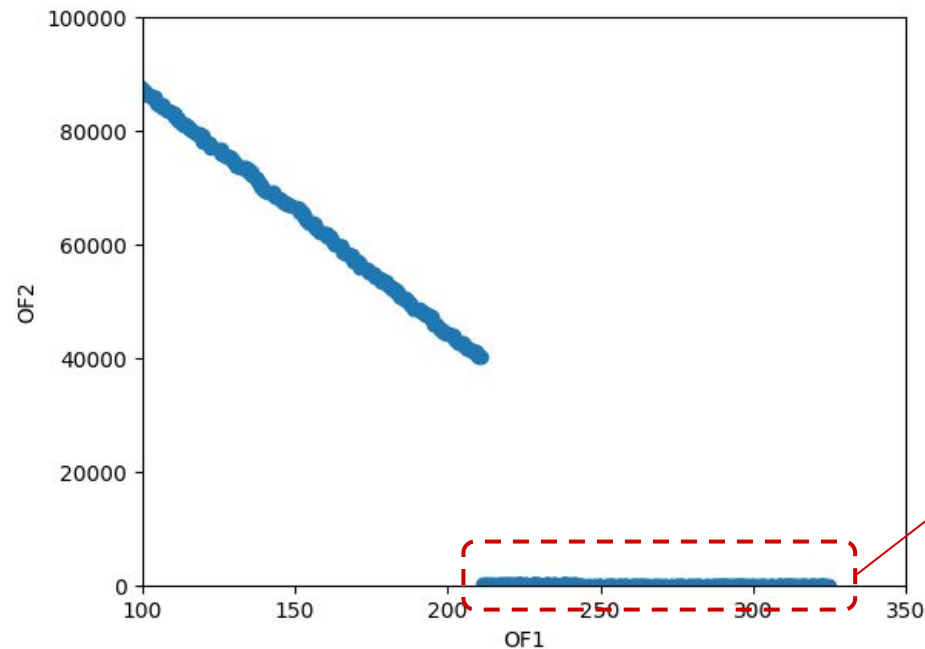
xAI: Counterfactual Explanation (CFE)

Minimization of:

$$OF_1(x, x_{faulty}) = \underbrace{\|x - x_{faulty}\|_2^2}_{\text{Minimal changes to original}}$$

$$OF_2(x, x_{faulty}) = \underbrace{AD_{score}(x)}_{\text{Anomaly score}} + \underbrace{\lambda AD_{label}(x)}_{\text{Abnormal penalization}}$$

- Solve using NSGA-II



xAI: “Counterfactual Explanation” (CFE population)

$$OF_1(x, x_{faulty}) = \|x - x_{faulty}\|_2^2$$

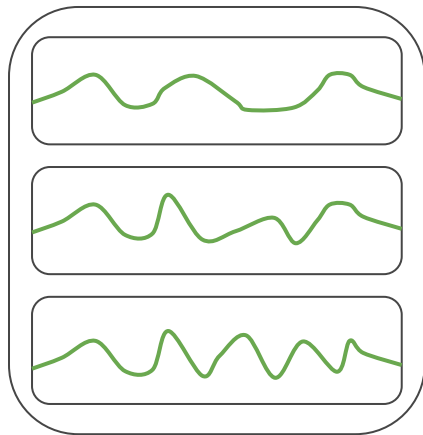
Minimal changes to original

$$OF_2(x, x_{faulty}) = \underbrace{AD_{score}(x)}_{\text{Anomaly score}} + \underbrace{\lambda AD_{label}(x)}_{\text{Abnormal penalization}}$$

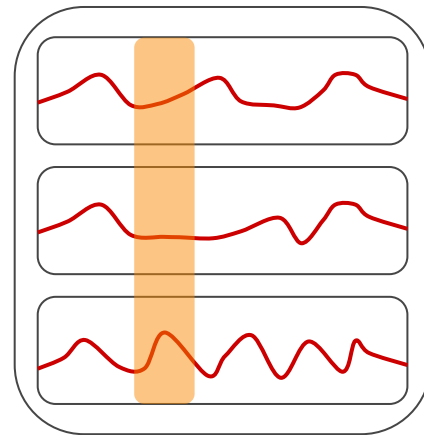
Anomaly score

Abnormal penalization

- Good enough counterfactual can be find in the data:

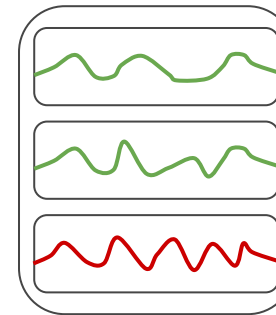


Normal sample

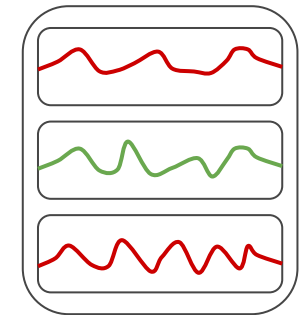


Abnormal sample

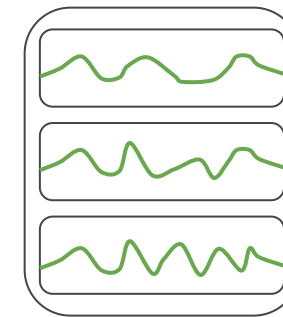
$OF = (0, 0.6)$



Candidate 1:
 $OF = (0.5, 0.7)$

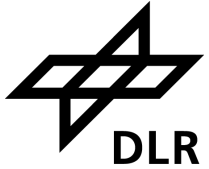


Candidate 3:
 $OF = (0.1, 0.2)$



Candidate 2:
 $OF = (0.3, 0.1)$

Anomaly Detection Models

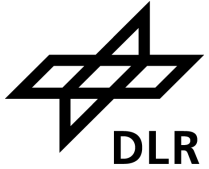


	F1	Recall	Precision	PR-AUC	PR-VUS
GMM	0.833	0.714	1	0.715	0.351
KNN	0.833	0.714	1	0.717	0.354
KDE	0.833	0.714	1	0.722	0.361
LUNAR	0.833	0.714	1	0.716	0.35
OCSVM	0.833	0.714	1	0.717	0.361
CBLOF	0.727	0.571	1	0.678	0.356
ANOGAN	0.25	0.143	1	0.172	0.044
HBOS	0.2	0.143	0.333	0.054	0.007
IFOREST	0.015	0.714	0.007	0.031	0.007
STRAY	0.014	0.369	0.007	0.004	0.186

- AD model hyperparameters tuned using Bayesian optimization.
- All models at least partially capture the 3 anomaly of interest.
- All models are from PyOD library.

Metrics

xAI Performance Metrics



Metric:

$$1 - \|IP(x) - IP_{true}\|_2$$

	IP from xAI	IP true
Temperature	0.01	0
Humidity	0.09	0
Precipitation	0.8	1
Pressure	0.05	0
Cloud Cover	0.05	0
Wind Speed	0	0

score = 0.93

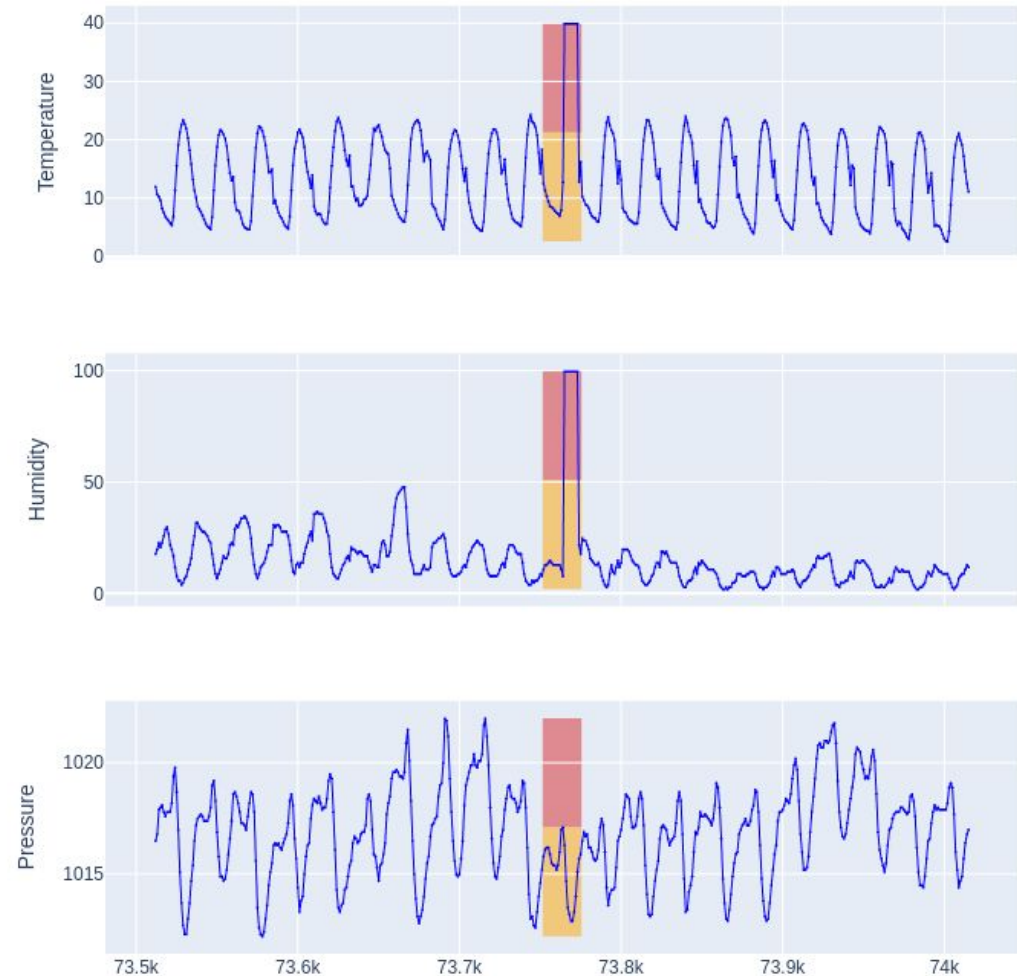
Synthetic Anomalies

Motivations:

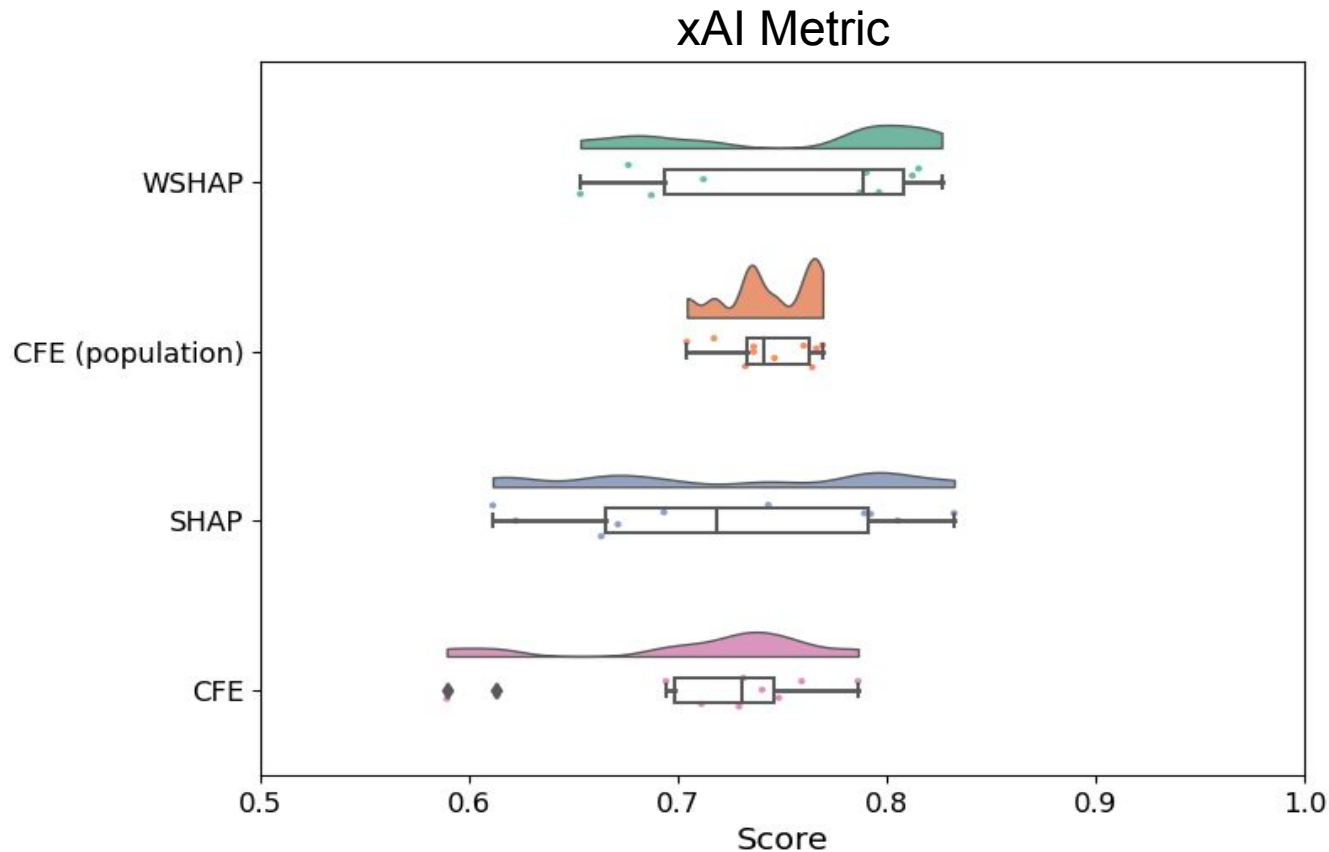
- Failing variables are known,
- Simulate isolated sensor failure: no causal effects.

Dataset:

- Consider only temperature, humidity and pressure from original Atacama dataset,
- 5 anomalies impacting one or two variables simultaneously,
- Original time series was seen as normal by the AD models before injecting anomalies.

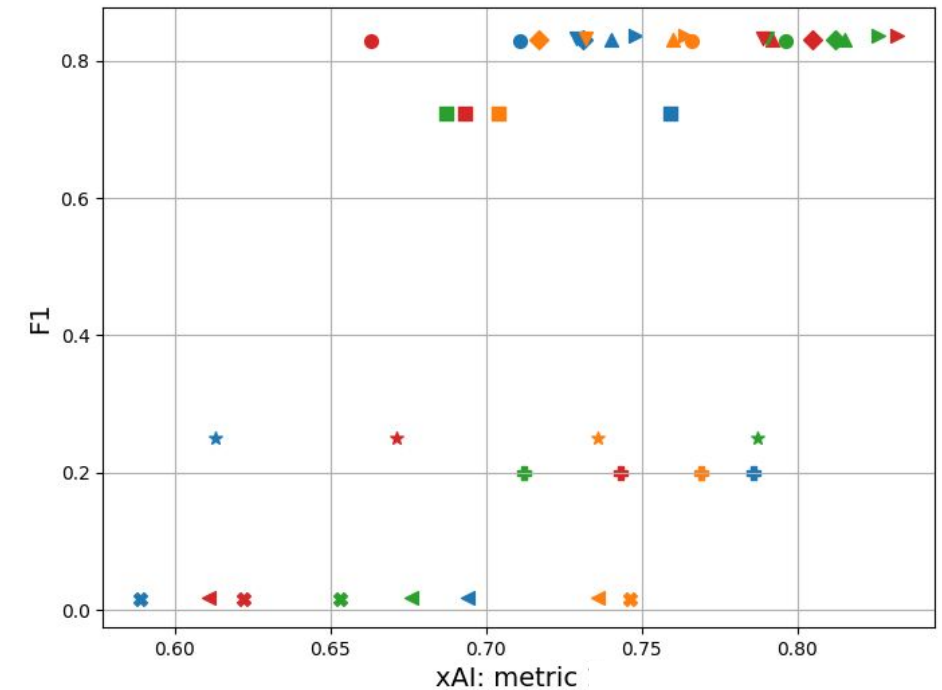
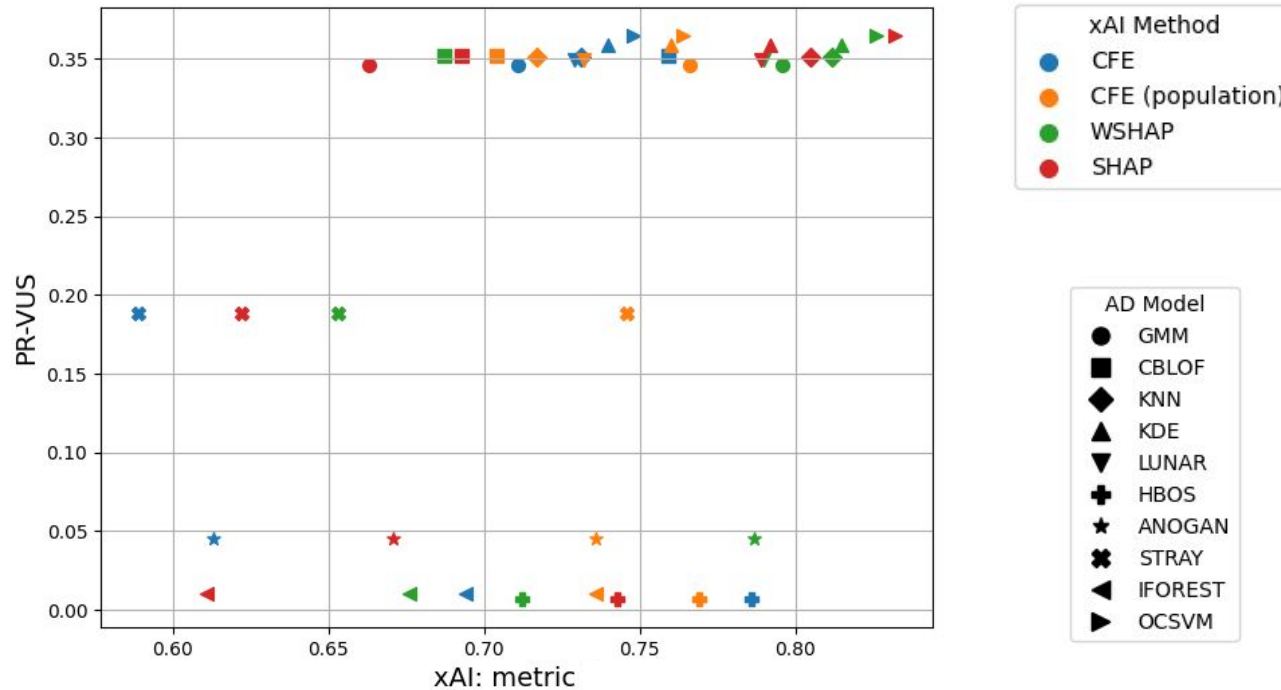


Metrics Testing on Synthetic Data



- For all 10 anomaly detection models,
- Large spreads show sensitivity of the explanation to the AD model type and/or performance.

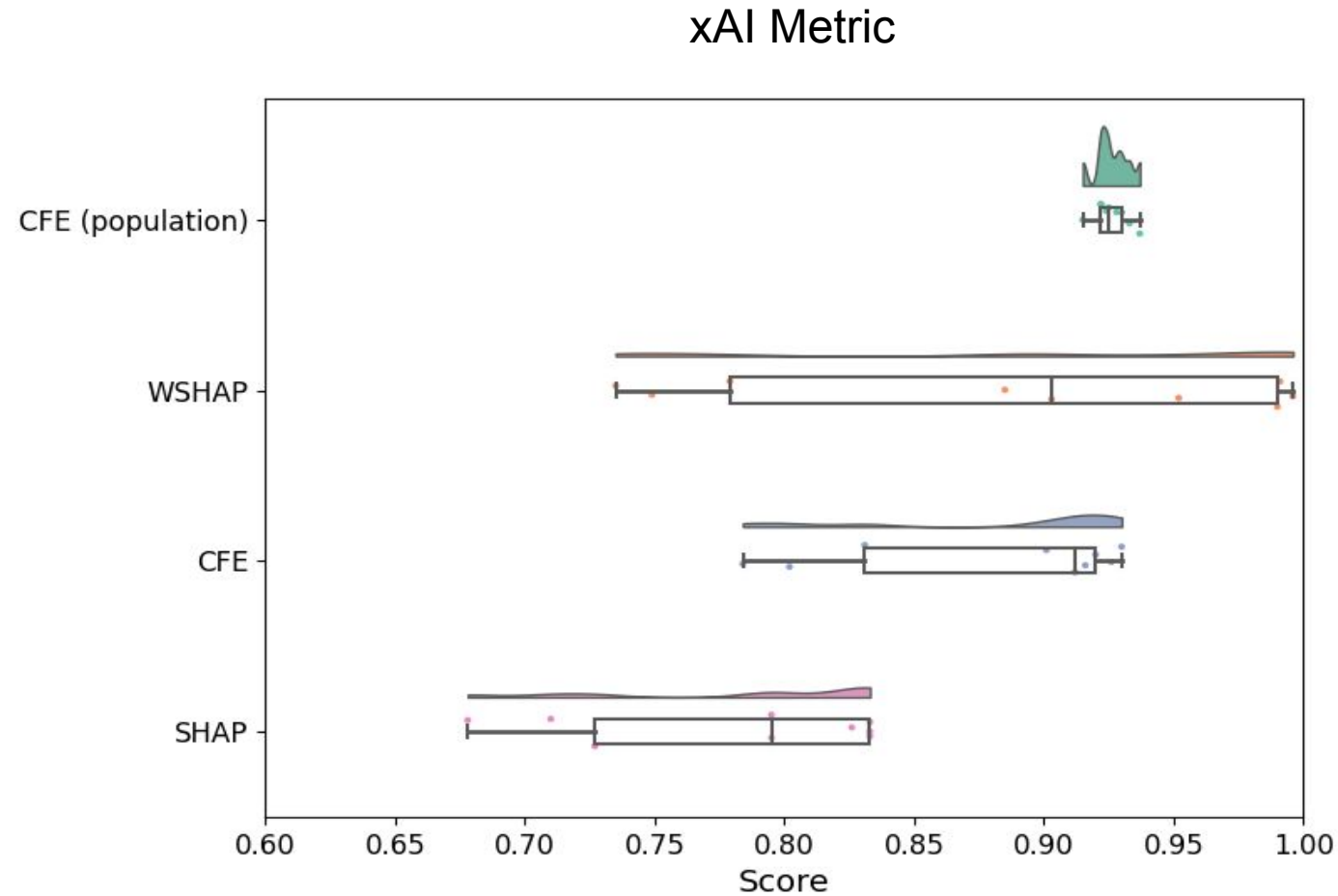
Results: xAI Score vs AD Score



- No correlation between the AD model and the xAI method performances,
- AD-xAI performance relationship could be only with a fixed AD model → need more datasets.

Results: ATACAMA

Results: xAI Performance



For all 10 anomaly detection models

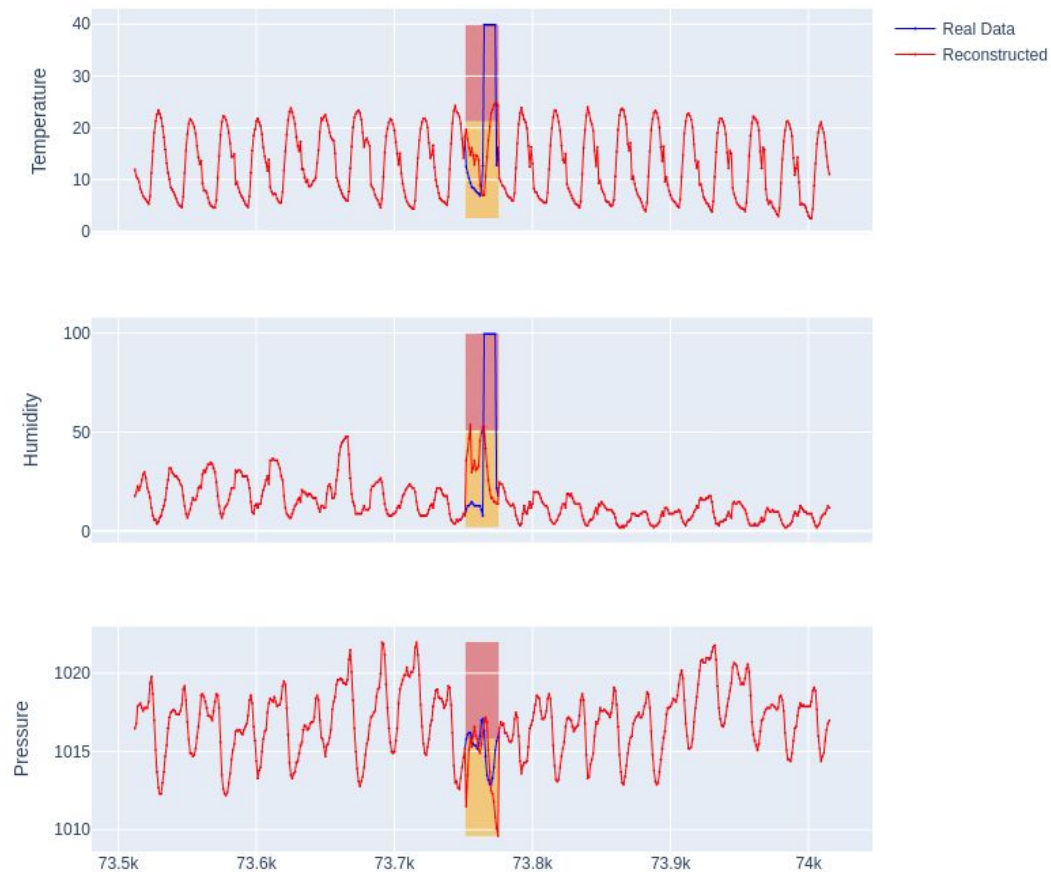
- Find more datasets for benchmark.
- Explore alternative xAI methods.
- Improve counterfactual:
 - Alternative optimization methods,
 - Improved GA optimization: parallelization & memory management.

Thank you for your attention

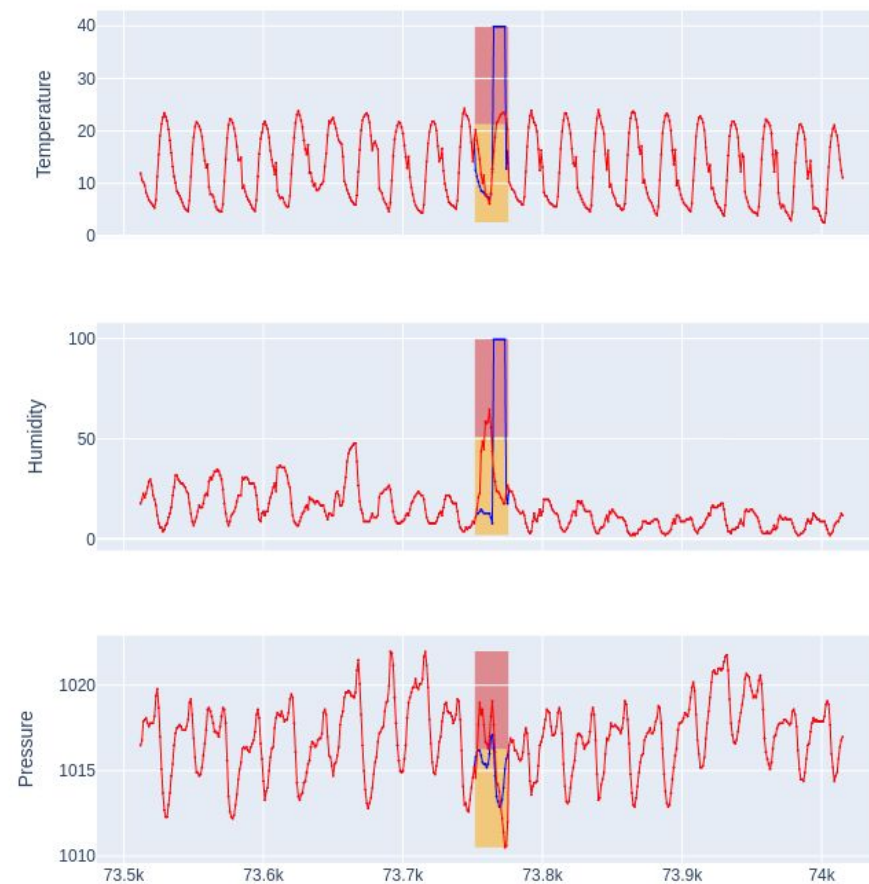
ATACAMA: Synthetic

CFE vs CFE (population)

CFE (population)

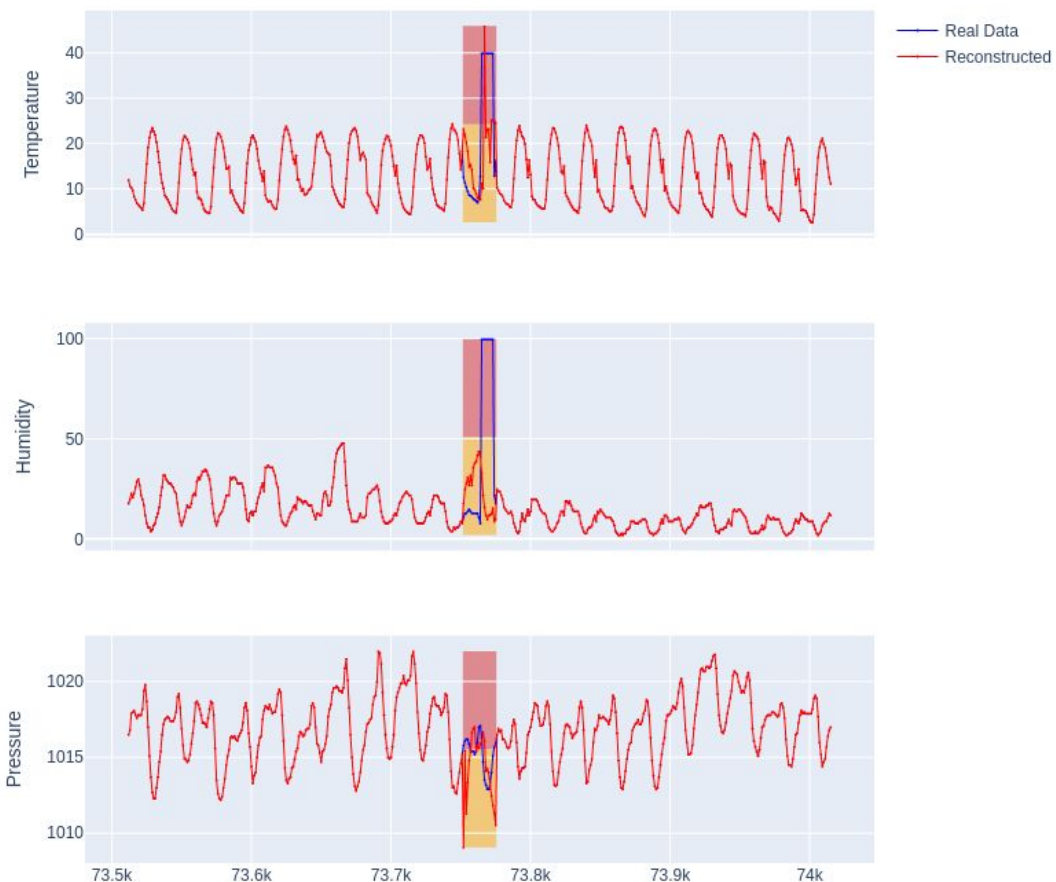


CFE

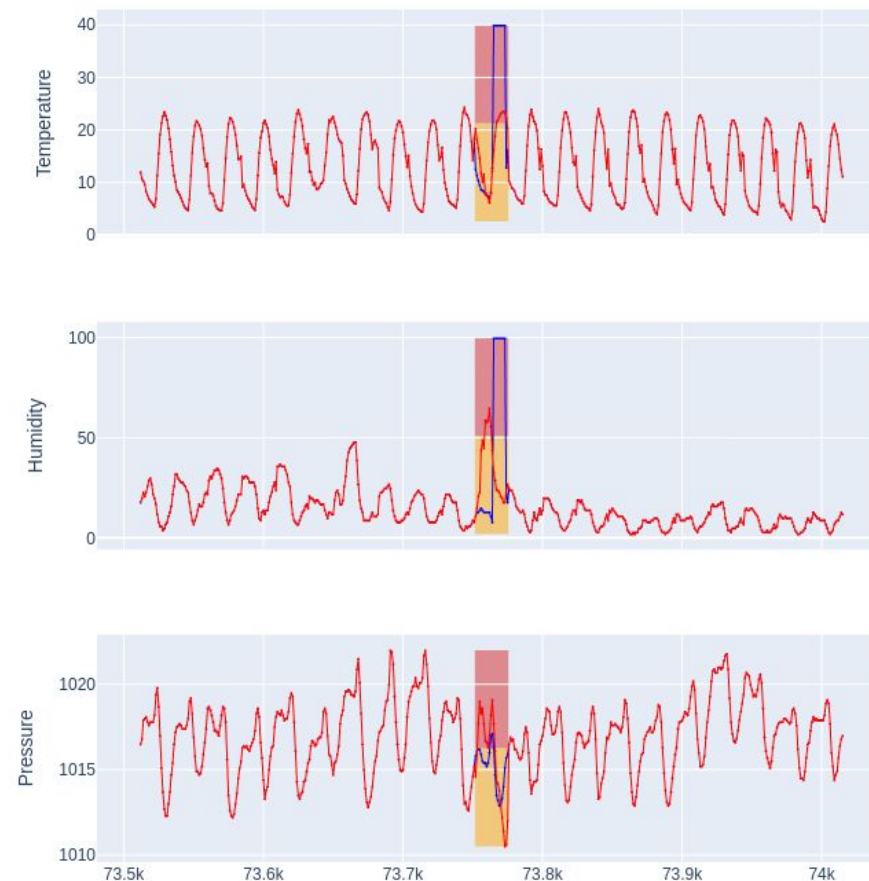


Counterfactual Quality vs AD Model

HBOS



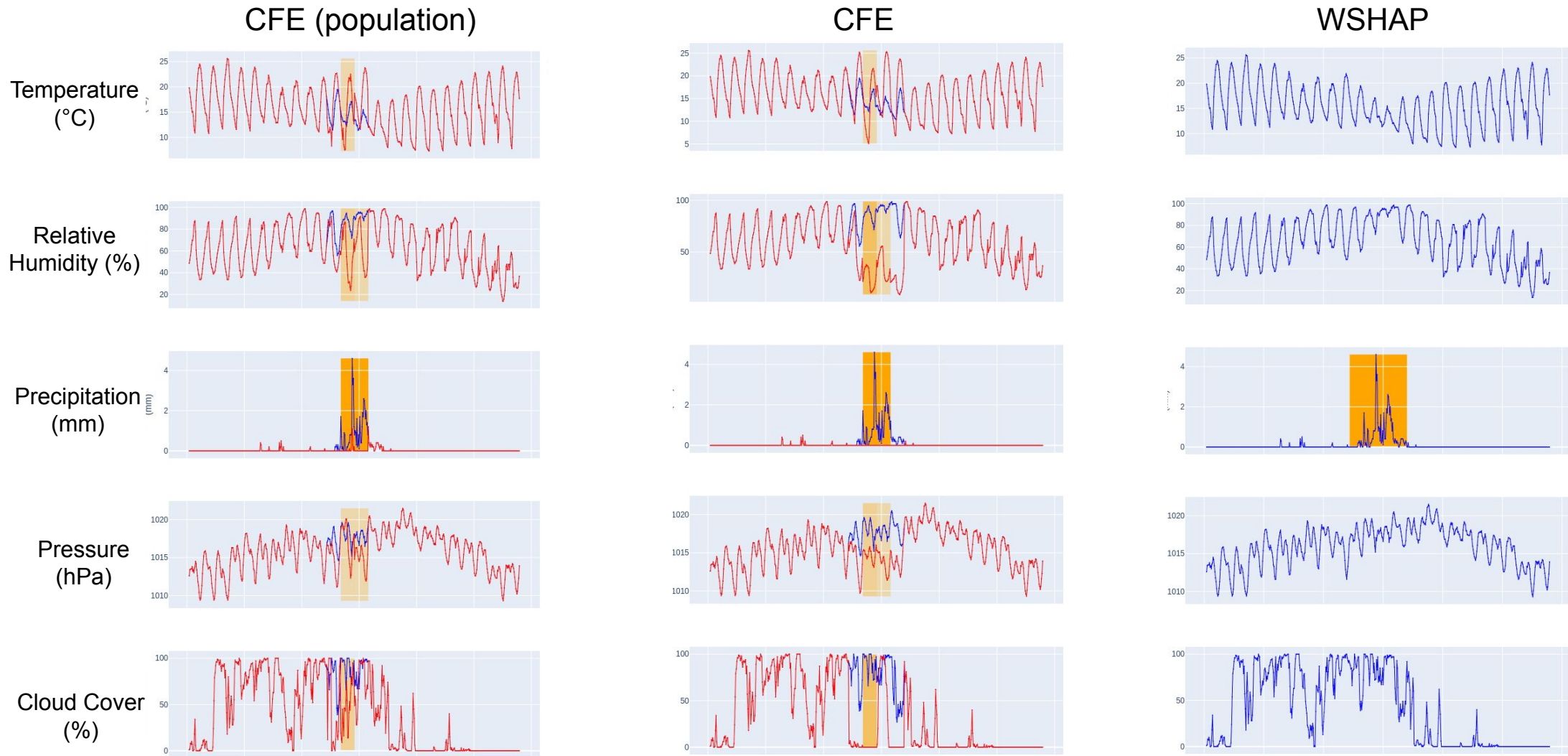
GMM



The background of the slide is a high-resolution photograph of the ATACAMA satellite in orbit. The satellite is a rectangular platform with two long, multi-segmented solar panel arrays extending horizontally from its central body. It is positioned against the backdrop of Earth, showing a mix of green landmasses and white cloud cover. The curvature of the Earth and the thin blue line of the atmosphere are visible at the bottom of the frame.

ATACAMA: Real

Details Analysis (GMM, Anomaly 1)



Blue: original data
Red: counterfactual

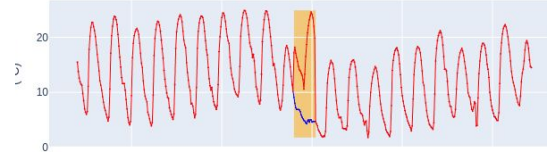
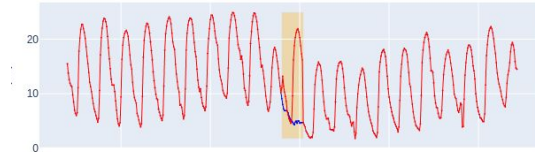
Details Analysis (GMM, Anomaly 2)

CFE (population)

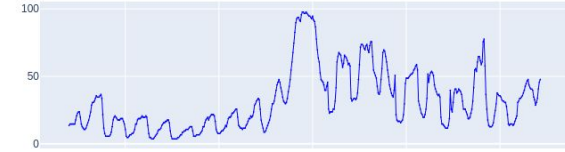
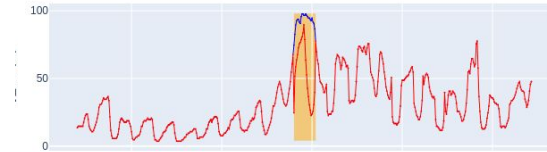
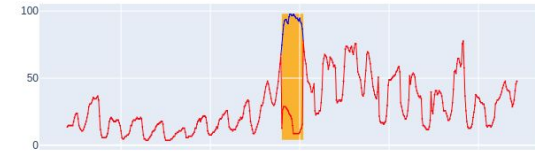
CFE

WSHAP

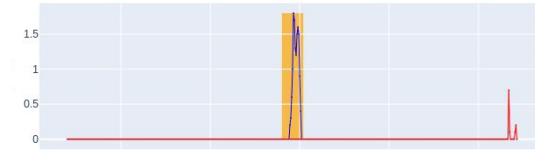
Temperature
(°C)



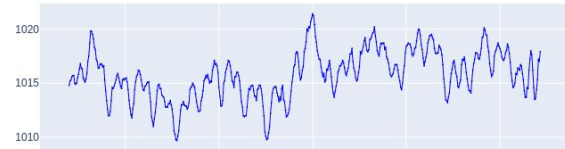
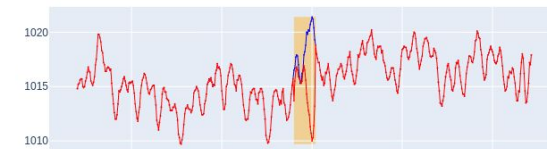
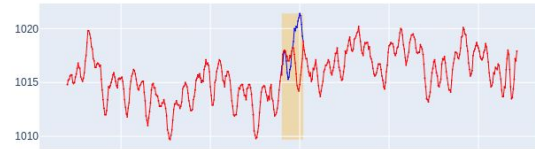
Relative
Humidity (%)



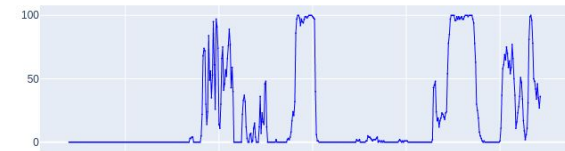
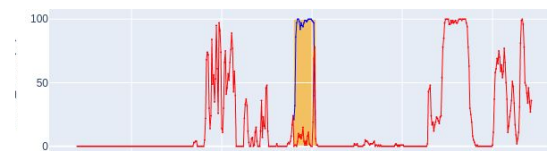
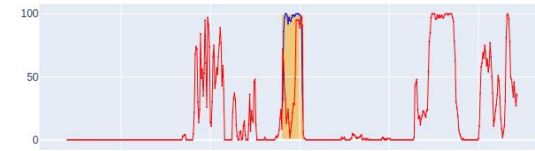
Precipitation
(mm)



Pressure
(hPa)

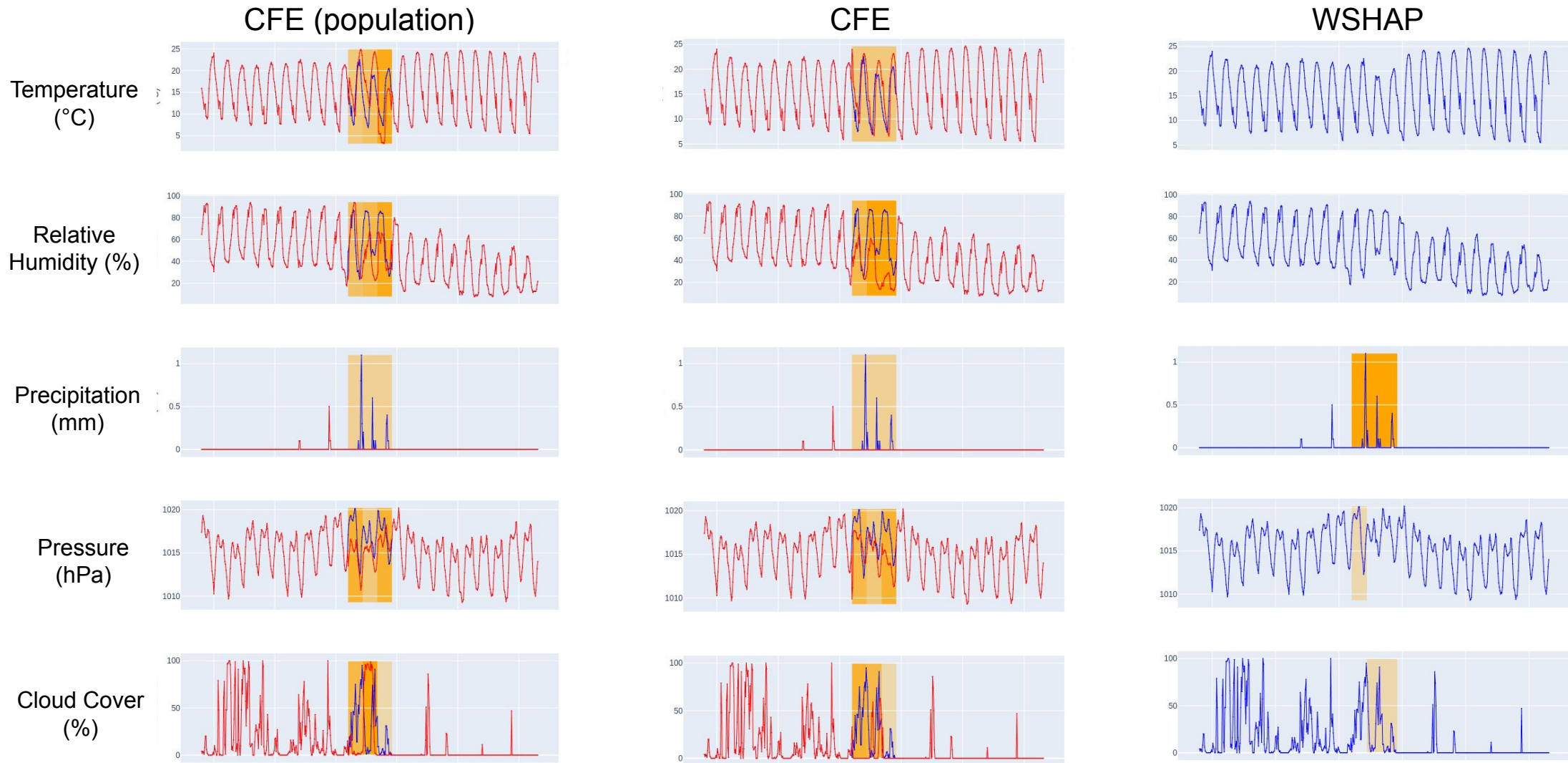


Cloud Cover
(%)



Blue: original data
Red: counterfactual

Details Analysis (GMM, Anomaly 3)



Blue: original data
Red: counterfactual

The background of the slide is a high-resolution image of a satellite in orbit. The satellite has a central body with various instruments and two long, rectangular solar panel arrays extending outwards. It is positioned over a view of Earth showing a coastline, likely the British Isles, with green land and blue water. The curvature of the Earth and the thin blue atmosphere are visible at the top of the frame.

Results: SaWT

Results: Secure Water Treatment Systems Dataset (SWaT)



Dataset:

- 51 variables,
- Some anomalies are results of simulated attacks,
- Attack points are seen as the true explanations to the anomalies,
- Labels extended over long time periods.



AD Model:

- CBLOF: F1 = 0.749, PR-VUS = 0.374.
- Find 7 true anomalies, 4 are attacks.

	WSHAP	CFE	CFE (population)	CFE: • 100 candidates. • 100 generations.
Metric 1	0.978	0.975	0.970	
Metric 2	1	1	1	
Runtime	112:10	42:53	10:05	

Results: Secure Water Treatment Systems Dataset (SWaT)

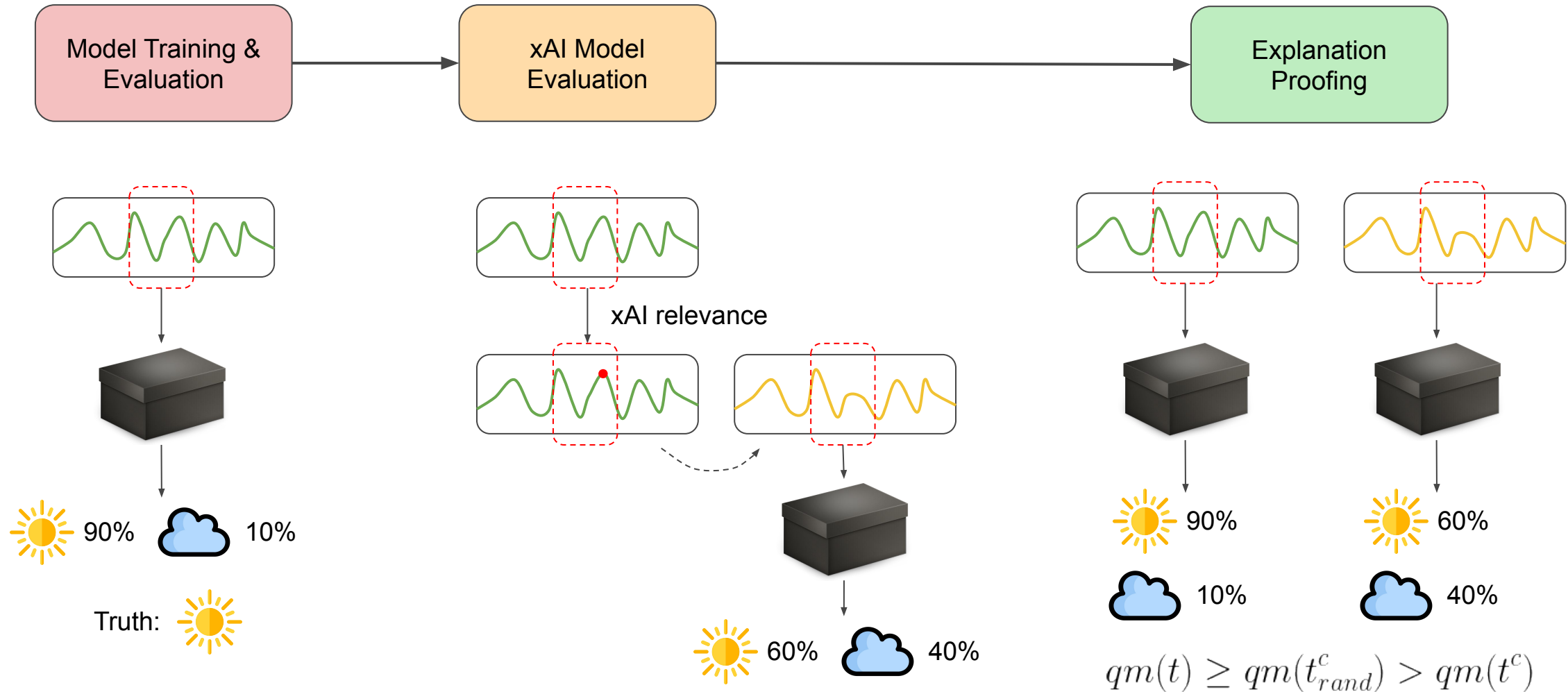


- Rank of the attacked variables:

	Attack	WSHAP	CFE	CFE (population)
Anomaly 1	AIT202	1	1	1
Anomaly 2	FIT401	2	1	1
Anomaly 3	AIT504	1	1	1
Anomaly 4	UV401	1	1	1
Anomaly 4	AIT502	4	13	14
Anomaly 4	P501	2	3	2

Alternative Methodology

Explanation Quality: Benchmarking Methodology



Thema: Improving Fault Diagnosis in Complex Systems: Investigating
xAI Techniques for Abnormal Sensor Isolation

Datum: 19.05.2025

Autor: Dr. Baptiste Lambert

Institut: DLR Institute of Data Science

Bildquellen: All pictures „DLR (CC BY-NC-ND 3.0)“,
unless otherwise stated