A Framework for Explainable AI in Weather Forecasting:

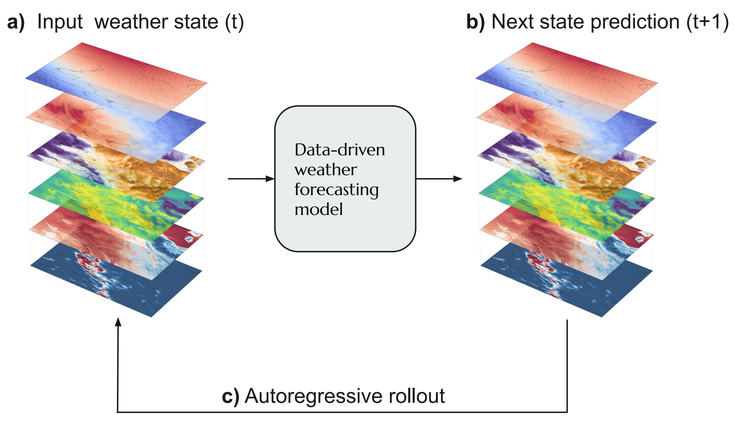
Diagnosing Deep Learning Models via Gradient-Based Attributions

Younes Essafouri[1] Corentin Seznec[2] Luciano Drozda[3]
Laure Raynaud[2] Laurent Risser[1]

[1]Univ. Toulouse, INSA-Toulouse, CNRS UMR 5219, IMT, Toulouse, F-31077, France
[2]Météo-France, CNRS, Univ. Toulouse, CNRM, Toulouse, France [3]Cerfacs, Univ. Toulouse, CNRS/Cerfacs/IRD, CECI

1. Deep Learning in Meteorology

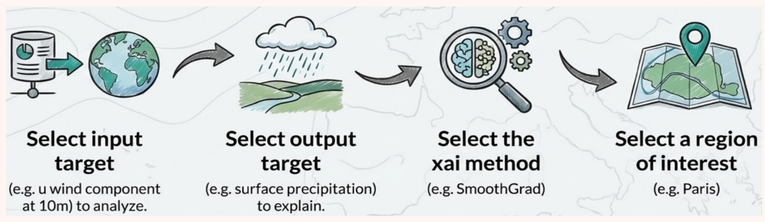
- **Efficiency:** DL models are increasingly competing with Numerical Weather Prediction (NWP) solvers, offering orders-of-magnitude faster inference [1].
- **Opacity:** Their inherent opaque **"black-box" nature** prevents operational trust [2].
- **Stakes:** Without interpretability, forecasters cannot verify *why* a prediction was made: a critical barrier for life-safety decisions such as extreme weather warnings.



2. Explainable AI to the Rescue

Explainable AI (XAI) seeks to open the DL "black box" by **attributing predictions to specific input features**.

Our Proposed Framework Pipeline:



For high-dimensional atmospheric grids, XAI methods based on game theory and on perturbations of the model's inputs (e.g., SHAP) face severe computational bottlenecks [3]. We therefore rely on **gradient-based methods** for their scalability to large spatial grids:

Method	Formulation
BaseGrad (Saliency)	$A(x) = \nabla_x f(x)$
Input × Gradient	$A(x) = x \odot \nabla_x f(x)$
SmoothGrad	$A_{SG}(x) = \frac{1}{N}\sum_{i=1}^{N} \nabla_x f(x + \epsilon_i)$

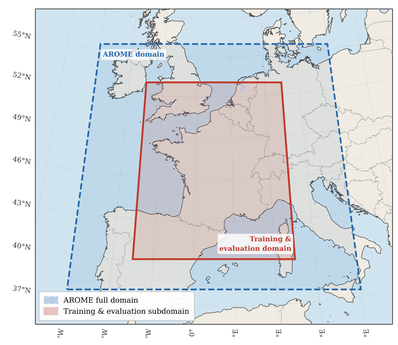
3. Numerical Setup

A. Model and Dataset

- **Data-driven model for regional high-resolution forecasting:** UNETR++ (hybrid transformer-convolutional)
- **Domain:** TITAN dataset based on the **AROME** analysis [4](Météo-France), 2.5 km resolution grid (512×640).
- **State Variables (21 Channels):**
  - *Surface:* Temp, Humidity, Zonal/Meridional Wind, Precipitation.
  - *Upper-Air (4 levels):* Temp, Winds, Geopotential at 250, 500, 700, 850 hPa.

B. The XAI Framework

- **The Target (Effect):** Surface precipitation.
- **The Input Space (Cause):** Zonal wind component.



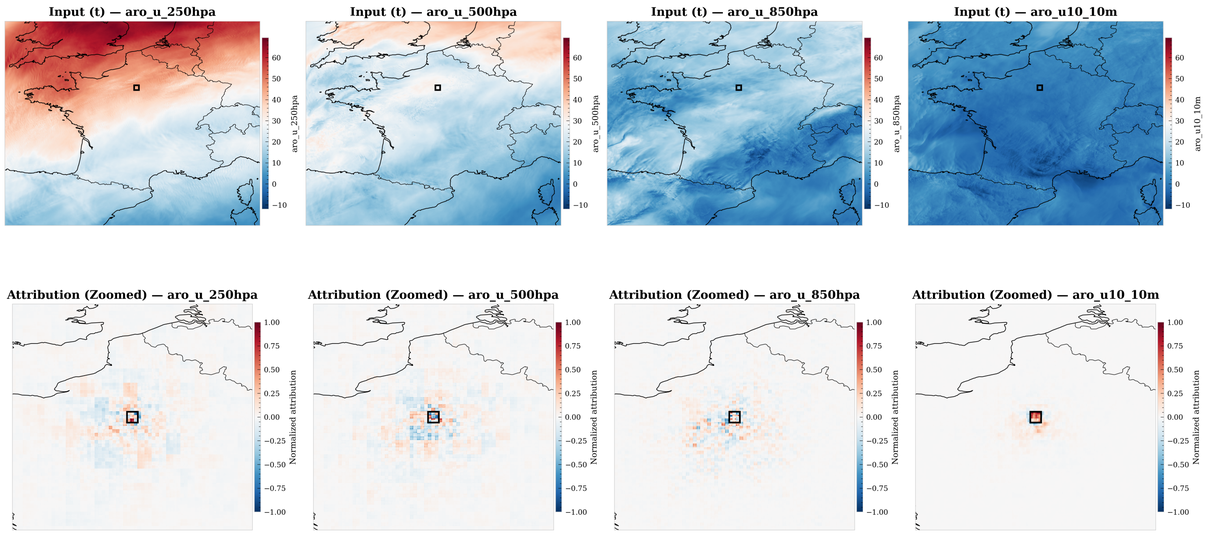
4. Physical Validation of XAI Maps

A. Vertical Hierarchy of Atmospheric Variables

The Diagnostic Setup: We run the attribution for a **single forecast step** ($T+1$), with no autoregressive roll-out.

- **Method Applied:** SmoothGrad ($N = 50$).
- **Interpretation:** Positive attribution values (red) indicate regions where an increase in the input wind directly intensifies the target precipitation.

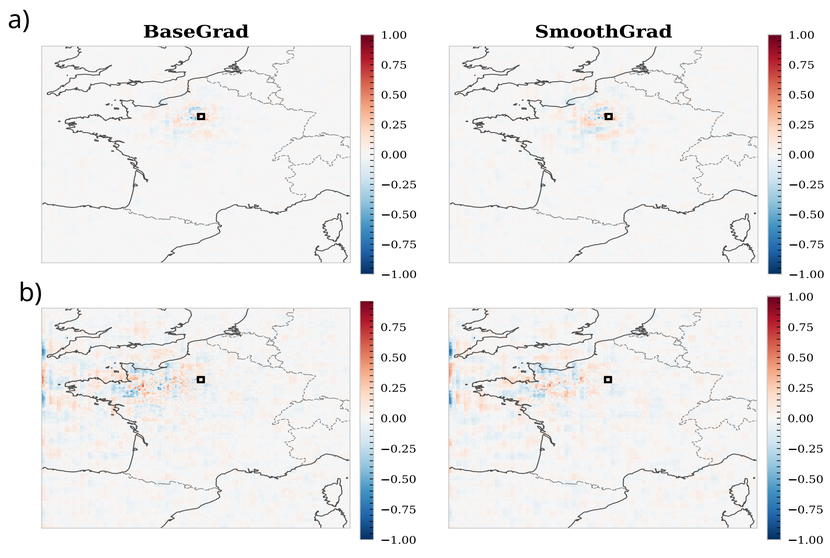
The framework verifies that the DL model correctly learns atmospheric vertical dynamics. **Effective receptive fields expand with increasing altitude**, matching the physical reality of upper-level atmospheric features having broader spatial influence.



B. Limitations of Gradient-Based Attribution Maps

When applied to continuous atmospheric fields, standard gradient methods produce attribution maps that are difficult to interpret physically, for three reasons:

- **High spatial variability:** Saliency maps exhibit strong pixel-level fluctuations, reflecting sensitivity to local input variations rather than large-scale meteorological structures.
- **Spatial dispersion across lead times:** As the model iterates over successive forecast steps: a) $T{=}1 \rightarrow$ b) $T{=}5$, the non-zero gradient values spread across an increasingly wide spatial region.
- **Smoothing Fails:** Standard techniques like *SmoothGrad* merely blur the noise without recovering the physical structure of gradient attributions.



5. From Quantification to Attribution Interpretation

To recover physical meaning from noisy attributions, we model the spatial distribution of attribution mass A_i at pixels $\mathbf{P}_i$ using **95% confidence ellipses**.

A. Centre of Mass (Weighted Barycenter)

$$\bar{\mathbf{P}} = \frac{\sum_{i=1}^{M} A_i \mathbf{P}_i}{\sum_{i=1}^{M} A_i}$$

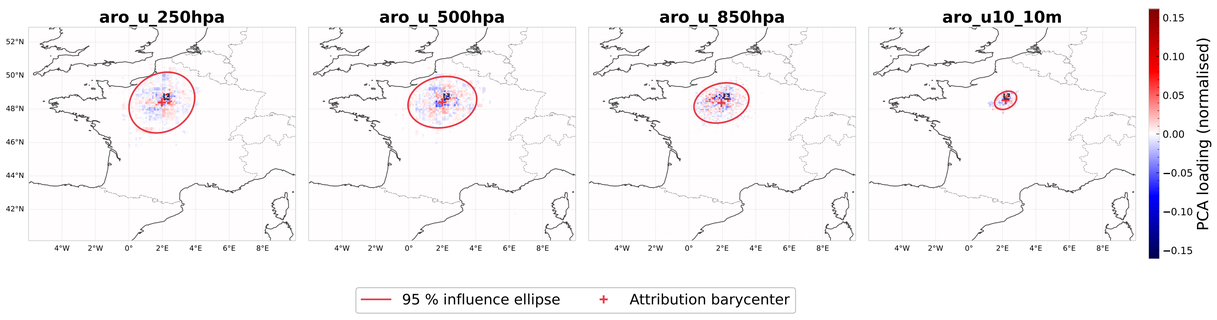
B. Spatial Spread (Weighted Covariance)

$$\mathbf{S}_w = \frac{1}{\sum_i A_i - 1} \sum_{i=1}^{M} A_i (\mathbf{P}_i - \bar{\mathbf{P}})(\mathbf{P}_i - \bar{\mathbf{P}})^T$$

C. Directional Axis & Shape

Extracted from the eigenvectors $(\mathbf{v}_1, \mathbf{v}_2)$ and eigenvalues (λ_1, λ_2) of the weighted covariance matrix $\mathbf{S}_w$, scaled by $\chi^2_2(0.95) = 5.991$:

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \bar{\mathbf{P}} + \begin{bmatrix} v_{1x} & v_{2x} \\ v_{1y} & v_{2y} \end{bmatrix} \begin{bmatrix} \sqrt{5.991\lambda_1}\cos t \\ \sqrt{5.991\lambda_2}\sin t \end{bmatrix}$$

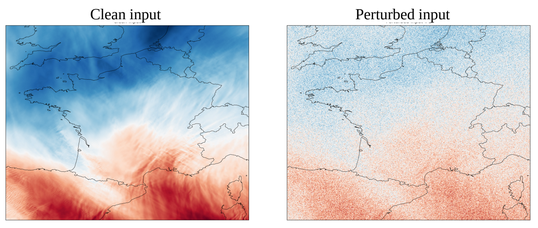


Variable	Centre (Lon, Lat)	Area (Spread)	Main Axis	Anisotropy
	Paris: (2.35°, 48.86°)			
Zonal Wind 250 hPa	(1.87°, 48.51°)	0.37	37.3°	1.13
Zonal Wind 500 hPa	(1.90°, 48.53°)	0.27	30.0°	1.21
Zonal Wind 850 hPa	(1.98°, 48.54°)	0.15	19.6°	1.37
Surface Wind 10m	(2.04°, 48.55°)	0.10	24.7°	1.35

6. Limitations & Future Work

Limitation: Unphysical Perturbations

Random noise applied to an atmosphere in equilibrium creates **physically impossible states** (e.g., violating mass conservation).



Path Forward: Physics-Aware XAI

- **Physics-Aware Perturbations:** Replace naive Gaussian noise with constrained, *physically admissible* perturbations derived from model climatology.
- **Concept-based XAI & Mechanistic Interpretability:** Extract physical meteorological variables directly from the model's hidden layers to build interpretable intermediate representations.

Acknowledgments & References

Acknowledgments

This work has benefitted from the AI Interdisciplinary Institute **ANITI**. ANITI funded by the France 2030 program under the Grant agreement **n° ANR-23-IACL-0002**.

References

[1] Lam, R., et al. (2023). *Learning skillful medium-range global weather forecasting.* Science.

[2] Bommer, P., et al. (2024). *Finding the right XAI method—a guide for evaluation and ranking.* AIES.

[3] Salih, A. M., et al. (2025). *A perspective on explainable AI methods: SHAP and LIME.*

[4] `huggingface.co/datasets/meteofrance/titan`