GenHack 2025 : Analyse des Îlots de Chaleur Urbains (UHI)

Équipe : NoName (Semaine 1)

Basé sur l'analyse approfondie des notebooks

17 novembre 2025

Contexte et Objectif

Contexte du Challenge (Semaine 1)

- Explorer l'effet d'Îlot de Chaleur Urbain (UHI).
- Utiliser les données de température ERA5, les cartes NDVI de Sentinel-2 et les observations des stations au sol.

Objectif Principal

- Démontrer que l'ajout du NDVI (végétation) améliore la prédiction de la température au sol (`T_Station`).
- Prouver la capacité à **généraliser** un modèle local (downscaling).

Sources de Données Utilisées

Ce notebook fusionne quatre sources de données distinctes :

1. **Température au Sol (Vérité Terrain)**
   - Source : `ECA_blend_tx/` (Données ECA&D)
   - Fichiers : `stations.txt` et `TX_...txt`
2. **Occupation du Sol (Végétation)**
   - Source : `sentinel2_ndvi/` (Données Sentinel-2)
   - Fichiers : Images GeoTIFF (`.tif`)
3. **Température Globale (Météo)**
   - Source : `derived-era5-land-daily.../` (Données ERA5)
4. **Géométrie (Frontières)**
   - Source : `gadm_410_europe.gpkg`

Définition de la Zone d'Intérêt (AoI) : La France

L'analyse est explicitement filtrée pour se concentrer sur la France.

- **1. Stations (Vérité Terrain) :**
  - Le fichier global `stations.txt` est filtré : `stations_df['CN'] == 'FR'`
  - Résultat : L'analyse se limite à **44 stations** situées en France.
- **2. Données Météo (ERA5) :**
  - Le fichier NetCDF chargé est déjà pré-filtré pour la France (`..._FRANCE.nc`).
- **3. Données NDVI (Végétation) :**
  - L'échantillonnage ne se fait **qu'aux coordonnées** des 44 stations françaises.

Méthodologie de Fusion et d'Analyse

1. **Fusion Spatiale (Fonction `create_snapshot`) :**
   - Crée un DataFrame unique (`T_Station`, `T_ERA5_C`, `NDVI`) pour les 44 stations à des dates clés.

2. **Analyse de Corrélation (Preuve UHI) :**
   - Vérification de la corrélation négative entre `NDVI` et `Temperature_C`.

3. **Modélisation par Cluster (Downscaling) :**
   - Entraînement d'un Modèle B (`T_ERA5 + NDVI`) sur un cluster local (ex : Marseille).
   - Test de sa capacité à "généraliser" sur une station non vue (ex : Toulouse).

4. **Analyse de performance par Altitude :**
   - Segmentation des stations en "Basse" $(\leqslant 500m)et"Haute"(> 500m)Altitude$.

Difficultés et Points Méthodologiques Clés

- **Gestion des Données (Drive) :**
  - Navigation dans un Google Drive partagé (chemins et filtres `glob.glob`).
- **Complexité Géospatiale :**
  - Gestion de Systèmes de Coordonnées de Référence (CRS) différents (EPSG :4326 vs EPSG :3035).
- **Hétérogénéité des Données :**
  - Fusion de données tabulaires (CSV), matricielles NetCDF (ERA5) et matricielles GeoTIFF (NDVI).

Analyse Approfondie - Preuve de l'UHI

Concerne : Cellule 20 (Nuage de points)

Objectif : C'est le graphique clé pour l'hypothèse.

- Vise à prouver visuellement l'effet UHI dans le snapshot du 15 juillet 2022.

Analyse Détaillée :

- Le graphique trace `Temperature_C` (Axe Y) en fonction du `NDVI` (Axe X).
- Il montre une **corrélation négative** évidente.
- **Interprétation :**
  - **NDVI faible (gauche)** → Peu de végétation → Températures **élevées**.
  - **NDVI élevé (droite)** → Beaucoup de végétation → Températures **basses**.
- **Coefficient de régression (Modèle B) : -2.203 °C.**

Résultats de la Modélisation (Downscaling)

Validation de l'ajout du NDVI et de la généralisation

1. **Amélioration du Modèle (Cluster Marseille - Été) :**
   - **Modèle A** (`T_ERA5` seul) : $R^2 = 0.632$
   - **Modèle B** (`T_ERA5 + NDVI`) : $R^2 = \mathbf{0.643}$
   - **Conclusion :** L'ajout du NDVI (facteur local) améliore la prédiction.

2. **Validation Externe (Robustesse) :**
   - Le modèle B (entraîné sur Marseille) est utilisé pour prédire une station non vue (Toulouse).
   - Résultat : Erreur moyenne (RMSE) d'environ **1.5 °C** sur les quatre saisons.
   - **Conclusion :** Le modèle B réussit à **généraliser**.

Conclusion Stratégique - L'Effet de l'Altitude

Le besoin de correction (downscaling) dépend de l'altitude

L'altitude est le meilleur prédicteur de la faiblesse de l'ERA5 :

Basse Altitude ($\leqslant 500m$) :

- **Le Modèle ERA5 est déjà excellent.**
- **R^2 (ERA5 seul) $\approx$ 0.900**
- **Besoin : Le NDVI (correction locale) est peu utile.**

Haute Altitude ($>$ 500m) :

- Le Modèle ERA5 est **faible**.
- R^2 (ERA5 seul) ≈ 0.35
- **Besoin :** Le NDVI (correction locale) est **crucial** pour capturer les microclimats.

Prochaines Étapes (Semaine 2)

1. **Feature Engineering (Priorité 1) :**
   - Introduire l'**altitude** (`HGHT`) comme variable explicative.
   - Objectif : Cibler directement la faiblesse de l'ERA5 et réduire l'erreur de prédiction dans les zones montagneuses.

2. **Modèles Avancés :**
   - Tester des modèles non linéaires (Random Forest, Gradient Boosting).
   - Objectif : Capturer des relations plus complexes (ex : interaction entre NDVI, altitude et ensoleillement).

Analyse des Îlots de Chaleur Urbains (UHI)

Modélisation des variations locales par couplage ERA5 et Sentinel-2

Équipe NoName

École Polytechnique

24 Novembre 2025

1. Intégration des Données et Périmètre

Objectif : Comprendre et modéliser les variations climatiques locales en France.

Sources de Données :

- **Stations (Cible) :** 44 stations fiables (ECA&D).
- **Météo Globale :** Données ERA5 (Température max 2m).
- **Occupation du Sol :** NDVI dérivé de Sentinel-2.

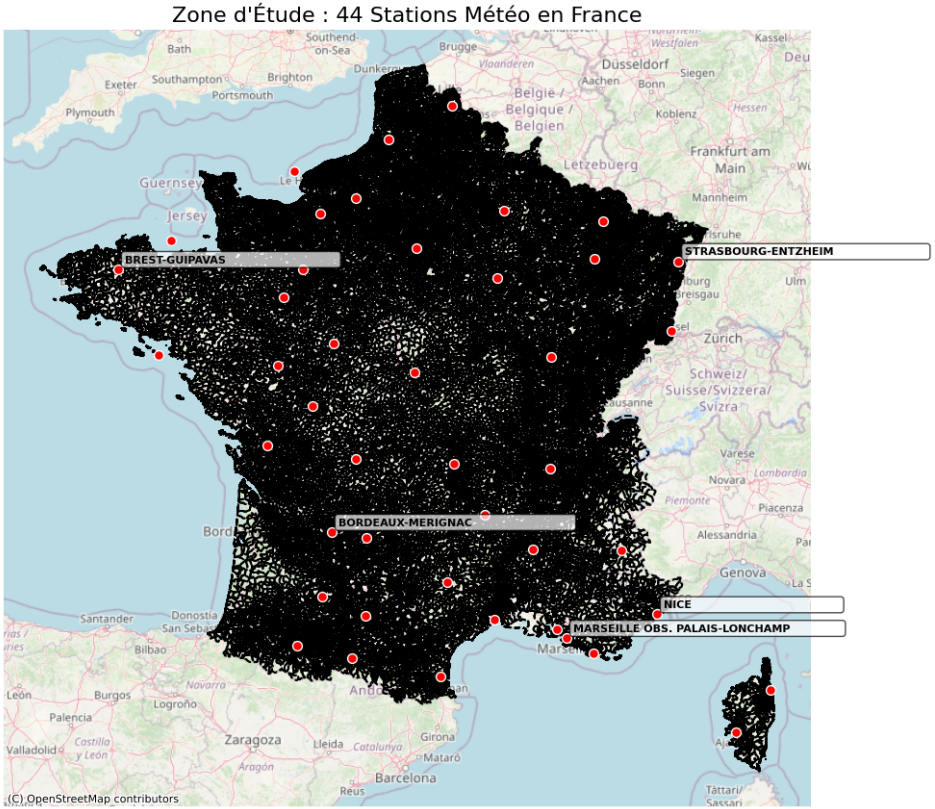


Figure – Répartition des 44 stations météorologiques utilisées en France.

2. Compréhension du Phénomène UHI

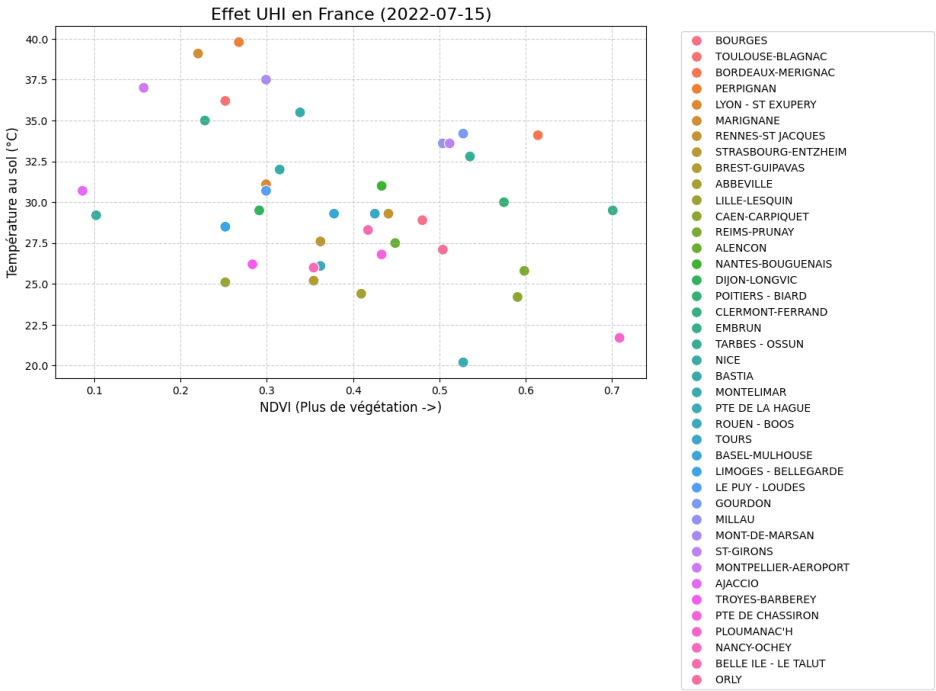


Figure – Corrélation entre Végétation et Température (Été 2022).

Observation Clé :

- Il existe une **corrélation négative** nette.
- Plus le NDVI est élevé (Végétation), plus la température au sol est basse.

Conclusion Physique :

- Les zones à faible NDVI (Urbaines/Béton) subissent une surchauffe locale.
- Le coefficient de régression indique une baisse de $\approx 2.2°C$ par point de NDVI.

3. Modélisation et Validation

Stratégie de Downscaling : Comparaison de deux modèles :

1. Modèle A : ERA5 Seul.
2. Modèle B : ERA5 + NDVI.

Résultat : Le modèle incluant le NDVI surpasse systématiquement le modèle global, particulièrement dans les zones complexes.

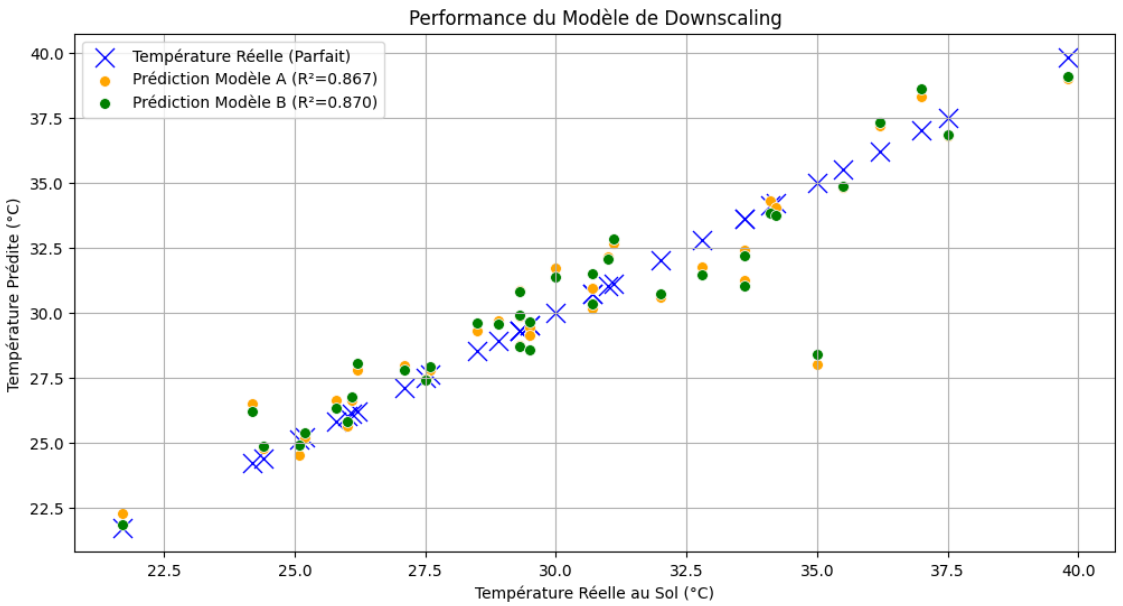


Figure – Comparaison des scores R^2 : Gain significatif grâce au NDVI.

4. Où le Downscaling est-il critique ?

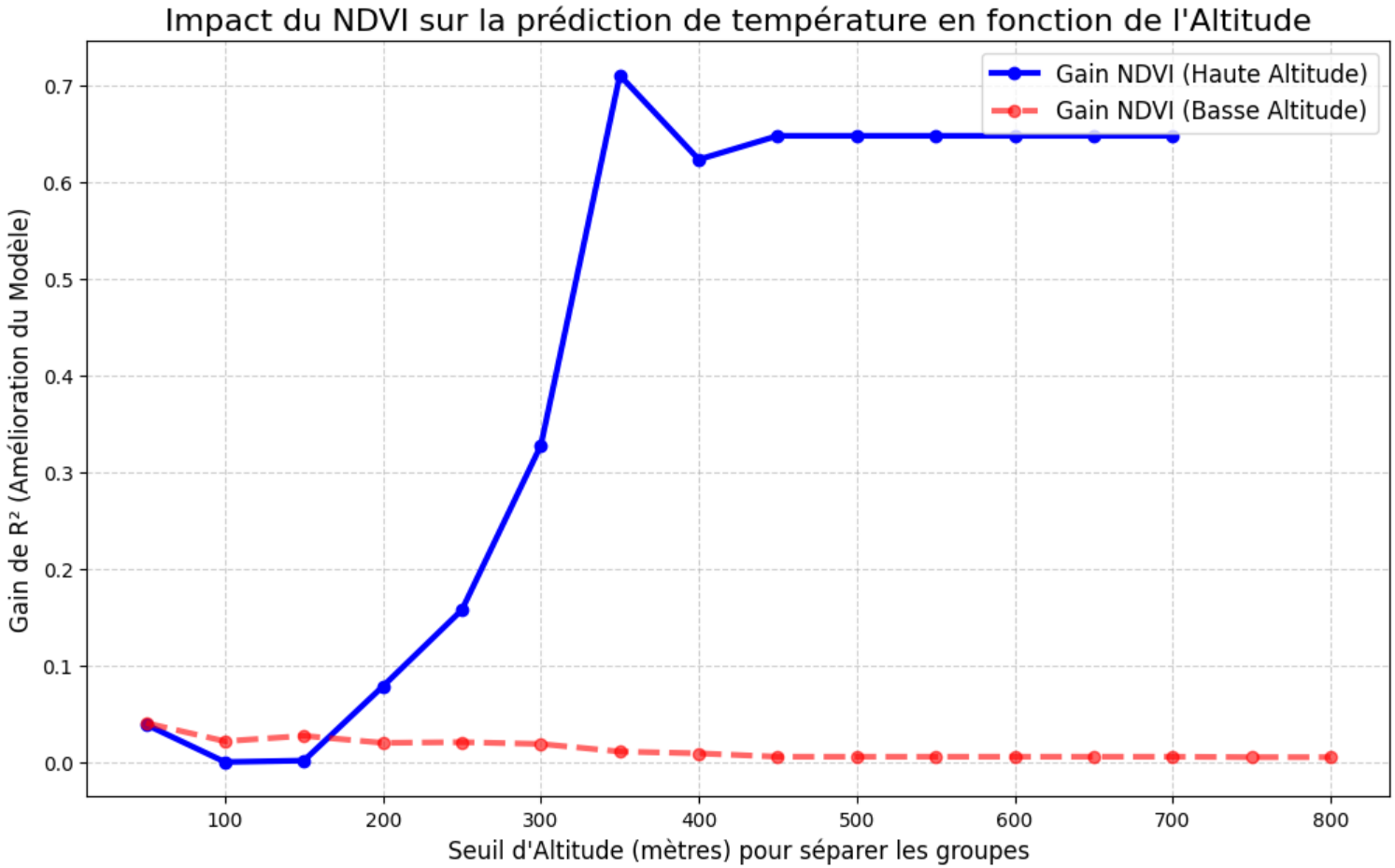


Figure – Sensibilité du Gain de Performance en fonction de l'Altitude.

Analyse Stratégique :

- **Basse Altitude (< 250m) :** Le modèle global ERA5 est suffisant

5. Synthèse Visuelle : Température et Végétation

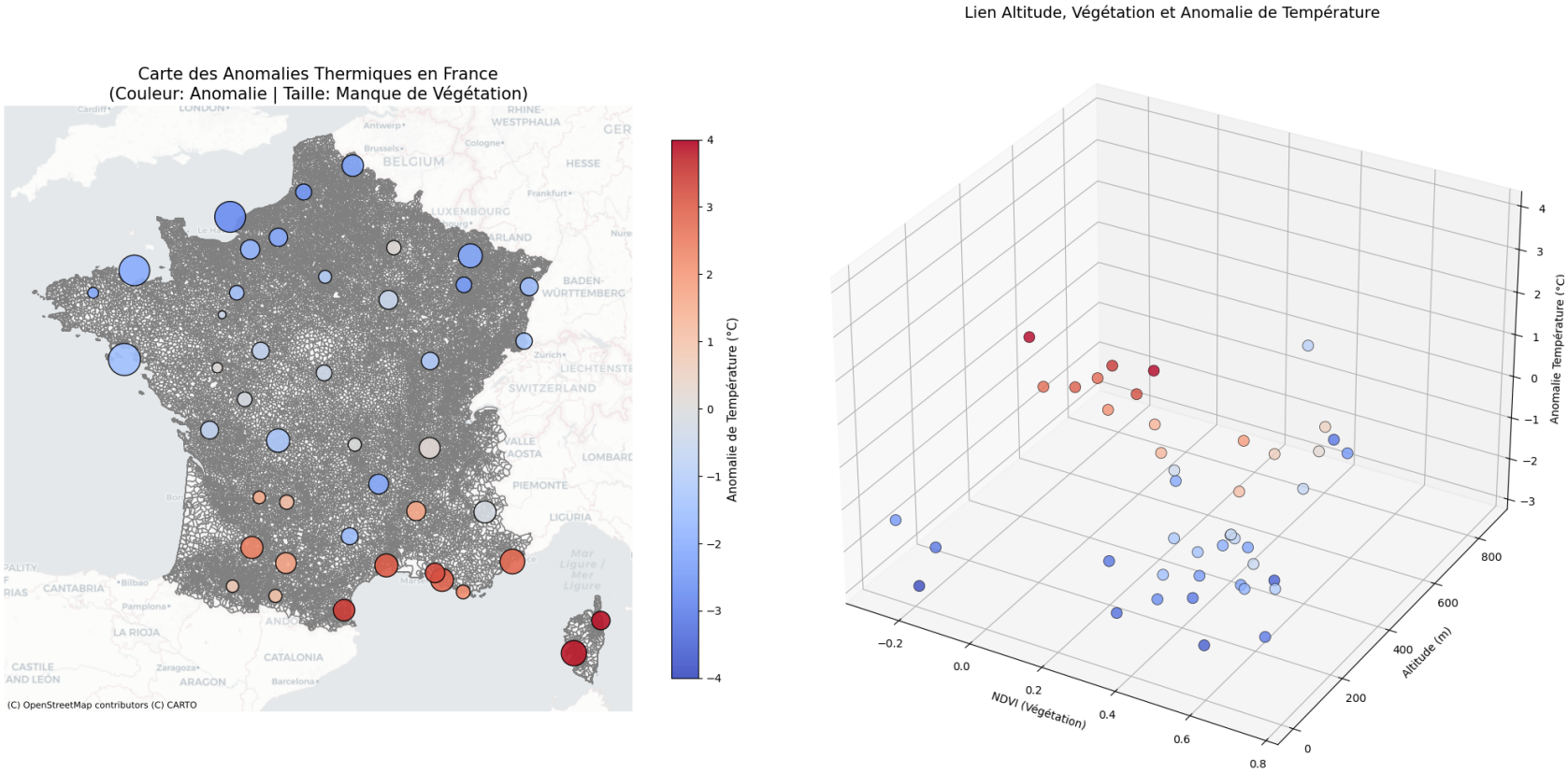


Figure – Carte de synthèse : Température (Couleur) et Lignes de niveau NDVI.

Conclusion et Analyse des Résultats (Semaine 1)

Objectif : Valider l'effet UHI par couplage **ERA5** (Global) + **NDVI** (Local).

8.1. Analyse des Résultats

- **Limites d'ERA5 :** Tendance globale correcte (R^2 élevé) mais **surestimation locale** (erreur $\approx 1.6^\circ C$ à Clermont-Fd).
- **Correction NDVI :** Gain variable selon la zone :
  - *Urbain :* Gain faible, ERA5 suffit.
  - *Végétalisé :* Gain fort. À Clermont, l'erreur chute de **1.6°C à 0.9°C**.

8.2. Synthèse Scientifique

- **Preuve UHI :** Corrélation négative confirmée ($-2.2^\circ C$ / point NDVI).
- **Validation :** Le downscaling réduit l'erreur (RMSE) quand la géographie locale diffère du climat régional.
- **Facteur Clé :** L'altitude et la végétation sont les déterminants de l'efficacité du modèle.

Conclusion Finale

L'information locale (NDVI) est le complément indispensable à l'information globale (ERA5) pour modéliser finement le climat urbain, particulièrement en zone complexe.

Team 23 NoName

Ayoub Olulad Ali
Maria Cristina Centorame
Yuxin Li

Discrepancies between Datasets

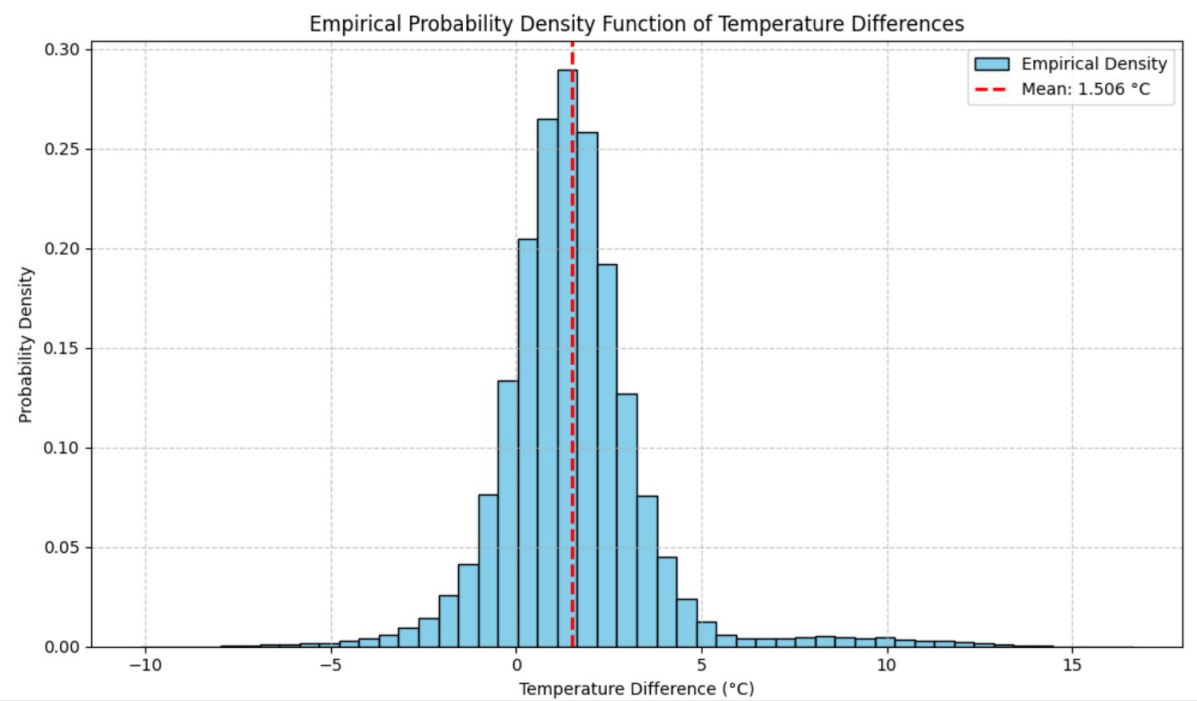
Comparing ERA5 and Weather Station Data

ERA5 vs Ground Truth

After matching the weather station temperature data with the nearest available ERA5 temperature data available in France, we found a difference:

Weather station data reports a higher temperature.

The mean of difference (weather station data - ERA5 data) ≈ 1.51, standard deviation is 2.01.



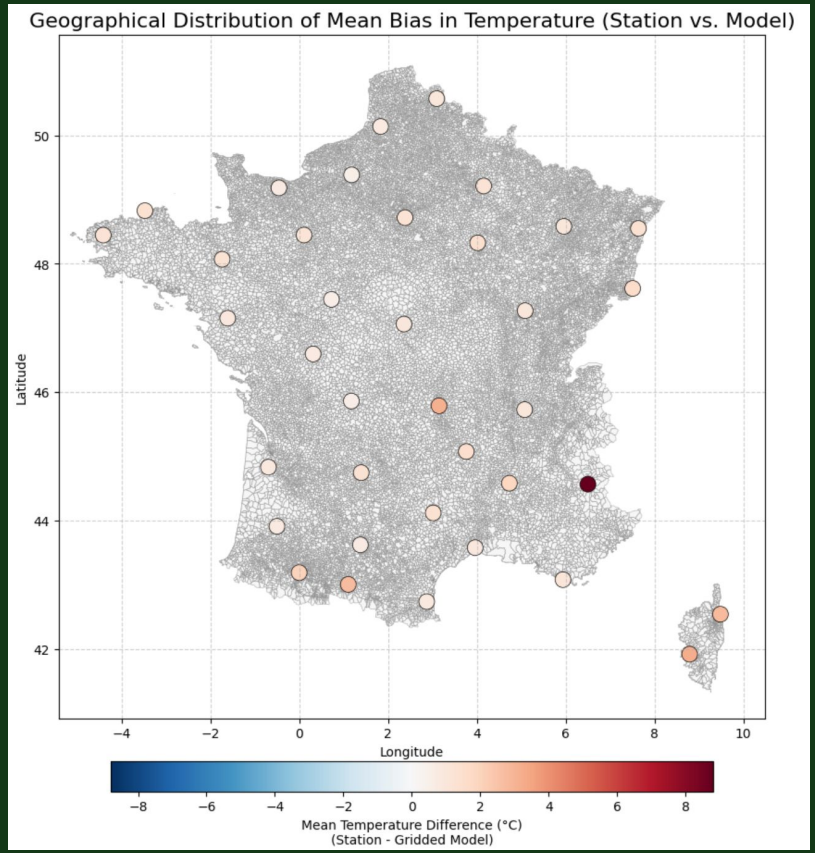
Geographical Differences:

By looking at temperature difference in all 44 weather stations in France, we found out the differences to be significant across all regions.

ID	Weather Station Name	Mean Temperature Diff	Significance
755	EMBRUN	8.80	Significant ($p < 0.05$)
2209	AJACCIO	3.23	
750	CLERMONT-FERRAND	3.10	
758	BASTIA	2.78	
2205	ST-GIRONS	2.73	

Top 5 Most Different Weather Stations

As shown on the right graph, the differences are more severe in the south-eastern region. Extreme outliers (e.g., Embrun +8.8° C) are driven by altitude/relief.
STRATEGIC DECISION: To capture the pure Urban Heat Island effect, we filtered the dataset to low-altitude stations (< 500m).



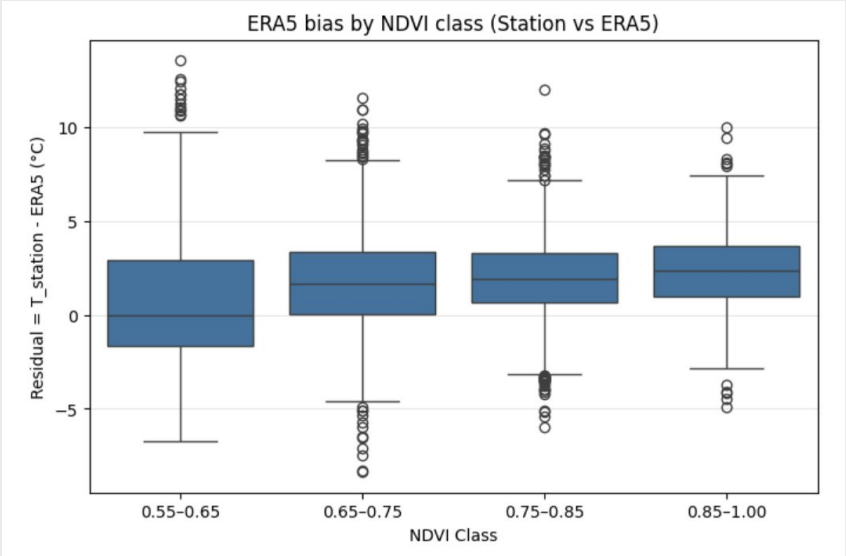
NDVI & Urbanization Effect on ERA5 Bias

Vegetation Strongly Modulates ERA5 Temperature Errors

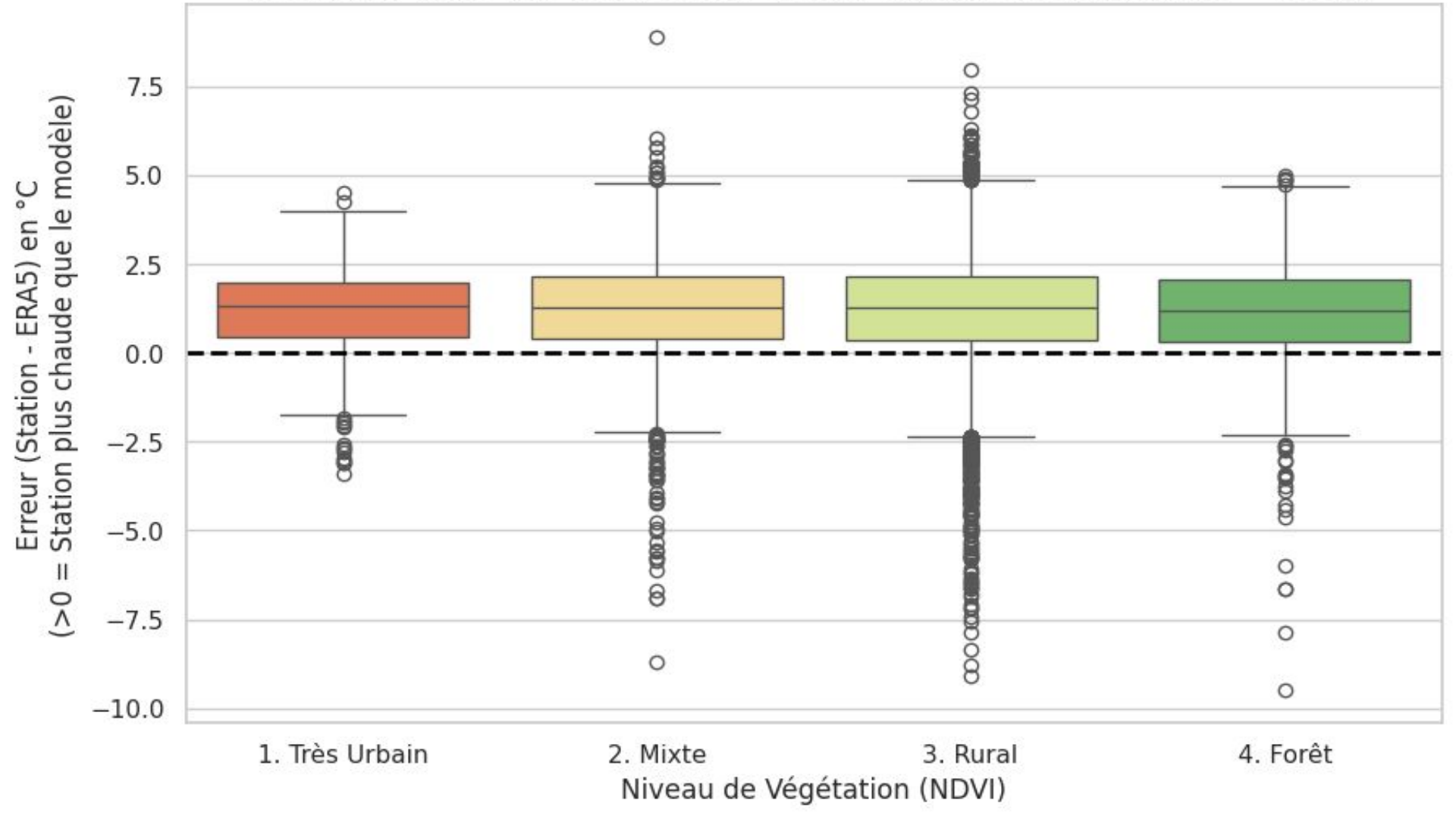
Key Results

- **ERA5 underestimates temperature in low-NDVI (urban-like) environments**:
  mean bias = **−1.04 °C**.
- **ERA5 overestimates temperature in high-NDVI (vegetated) regions**:
  mean bias = **+1.66 °C**.
- **Error variability is highest in low-NDVI areas** (std = **3.57 °C**), reflecting the strong heterogeneity of urban microclimates that ERA5 cannot resolve.
- **These patterns are consistent with an urban heat island signature**: ERA5 smooths strong urban warming and fails to capture localized heat amplification.

NDVI class	Mean bias	Std. dev.	Count
0.55–0.65	−1,04	3.57	576
0.65–0.75	−0,74	2.43	1822
0.75–0.85	−0,96	2.06	2175
0.85–1.00	+1,66	2.03	947



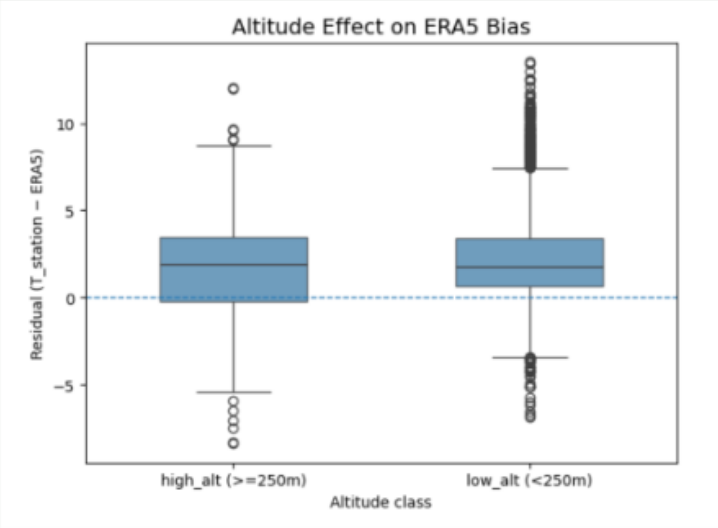
Note: Analysis performed on a curated subset of stations with complete daily data and valid NDVI sampling to ensure comparability across NDVI, season, and altitude classes.


Erreur (Station - ERA5) en °C
(>0 = Station plus chaude que le modèle)
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
1. Très Urbain
2. Mixte
3. Rural
4. Forêt
Niveau de Végétation (NDVI)

Geographic Factors: Altitude, Latitude, Coastline

Geographic Factors: Altitude, Latitude, Coastline

Region	Mean Bias (°C)	Std	RMSE	N
East / Inland	2.11	2.56	3.32	4879
West / Coastal	1.38	2.61	2.95	1095



Altitude effect:

- **Low-altitude (<250 m): mean bias = +2.29 °C**
- **High-altitude (≥250 m): mean bias = +1.66 °C**

Low-altitude stations tend to be more **urban and inland**, where ERA5 struggles the most to capture local heat amplification.

Coastline effect

ERA5 underestimation is **stronger inland** (mean = +2.11 °C) than in west/coastal regions (mean = +1.38 °C). Coastal thermal gradients make ERA5 slightly more reliable there.

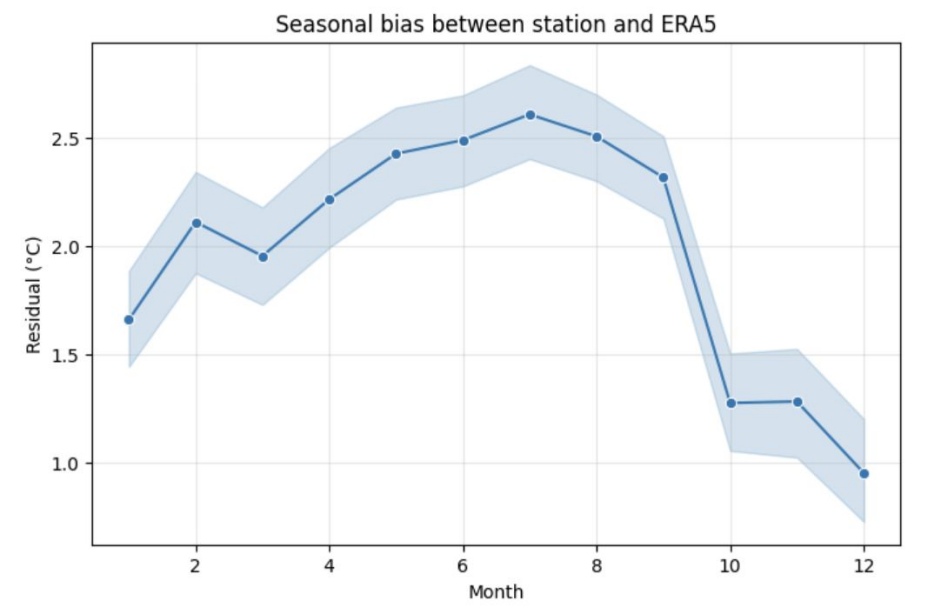
Seasonal Bias Patterns

Seasonal Patterns Reveal Missing Urban Warming in Summer

The discrepancy between station data and ERA5 is strongly seasonal:

- ERA5 underestimates temperature year-round.
- Largest bias in summer (+2.53 °C) → strongest UHI + solar forcing.
- Moderate bias in spring (+2.20 °C).
- Lower bias in autumn (+1.63 °C) and winter (+1.54 °C).

season	mean	std
autumn	1,633	2,679
spring	2,198	2,531
summer	2,534	2,368
winter	1,543	2,624



Seasonal pattern matches known UHI dynamics: warm-season heat storage is not captured by coarse-grid reanalysis.

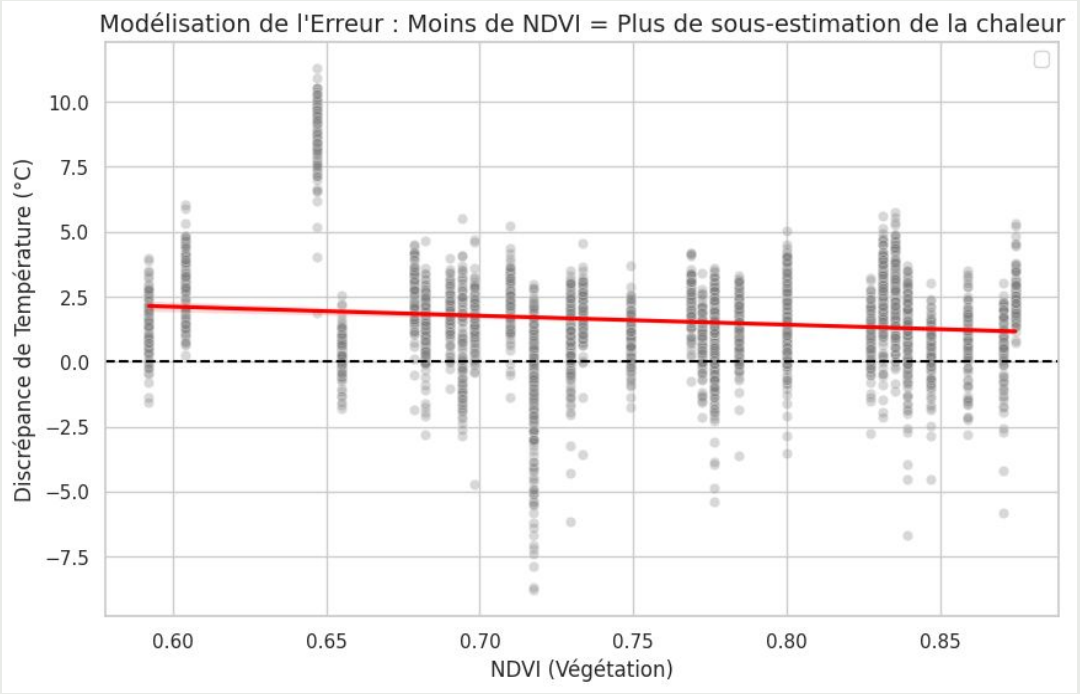
Impact of Vegetation on Temperature Prediction Error

Observations:

- Slight negative trend: higher NDVI → smaller error.
- Low vegetation areas (low NDVI): model often underestimates heat.
- High dispersion, especially around NDVI ≈ 0.65–0.70.

Interpretation:

- Densely vegetated areas: errors close to 0°C.
- Low vegetation / urban areas: temperature underestimated.
- Suggestion: account for vegetation to improve the model.
-



NoName

GENHACK WEEK 4

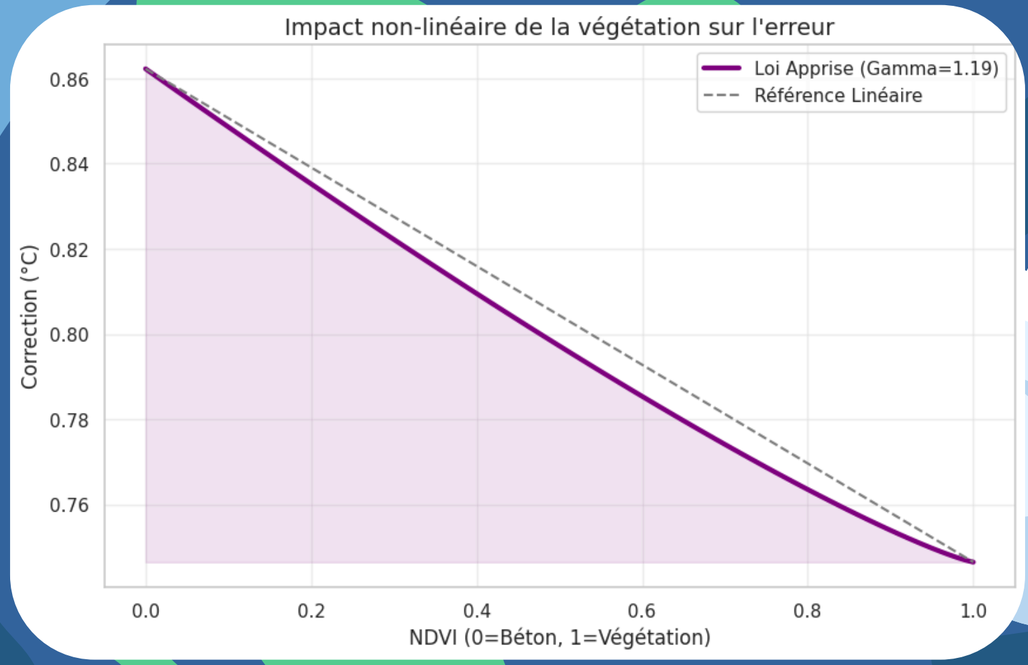
OULAD ALI AYOUB
YUXIN LI
MARIE C. CENTORAME

Objective : Correct the bias

Approach

Framing the task as a Residual Learning problem

- Interpretable Linear Methods (OLS)
- Physics-Informed Neural Network (PINN)
- Non-linear Ensemble Learning (Random Forest - RF)
- SOTA Deep Learning: Graph Neural Network (GNN)



The Nonlinear Role of NDVI

$$\Delta T = \beta_0 + \beta_1\,\mathrm{NDVI} + \beta_2\,\mathrm{Altitude} + \beta_3\,\mathrm{SummerFlag} + \beta_4\,\mathrm{NDVI}^2 + \varepsilon.$$

Predictive Correction Models

LINEAR REGRESSION WITH WITH A CUSTOM LOSS FUNCTION

$$\hat{\epsilon} = \beta_0 + \beta_1 \cdot \text{Altitude} + \beta_2 \cdot \text{NDVI}.$$

Predictive Correction Models

PHYSICS-INFORMED NEURAL NETWORK

$$\mathcal{L} = \text{MSE} + \lambda \frac{1}{N} \sum_{i=1}^{N} \left(|\hat{\epsilon}_i - \epsilon_i| \cdot \frac{e^{-\sqrt{H_i}}}{\text{NDVI}_i + \delta} \cdot I_{\text{summer},i} \right)$$

Predictive Correction Models

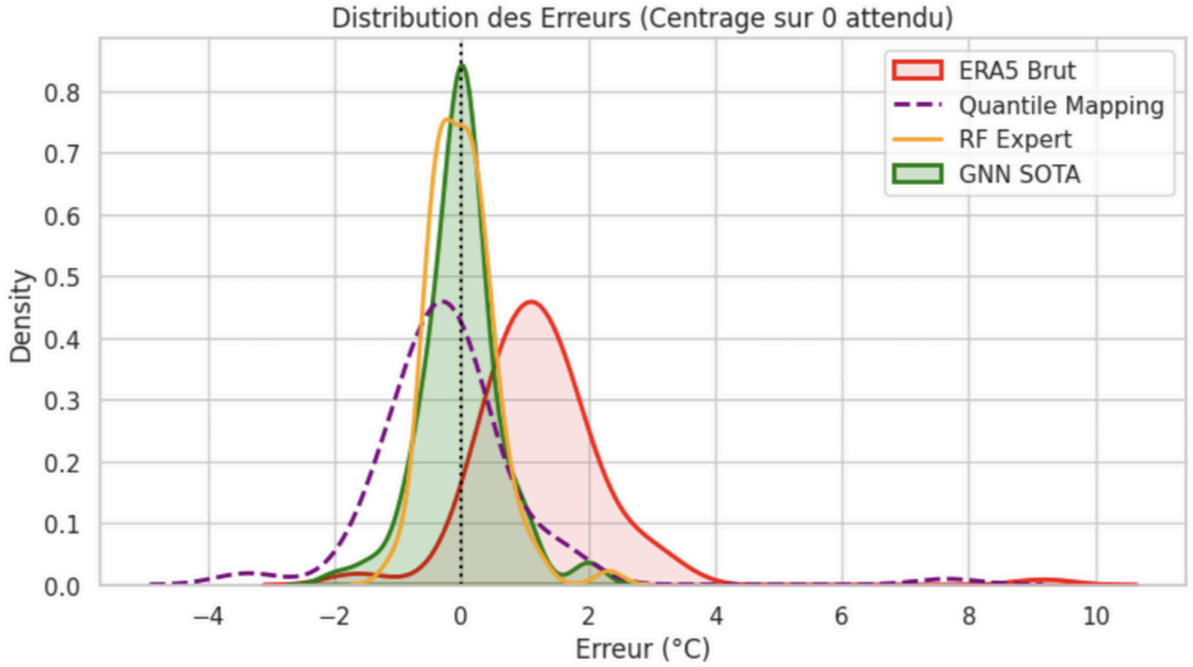
RANDOM FOREST REGRESSOR WITH FEATURE ENGINEERING

$$T_{\text{final}} = T_{\text{ERA5}} + \text{RF}(\text{Features}).$$

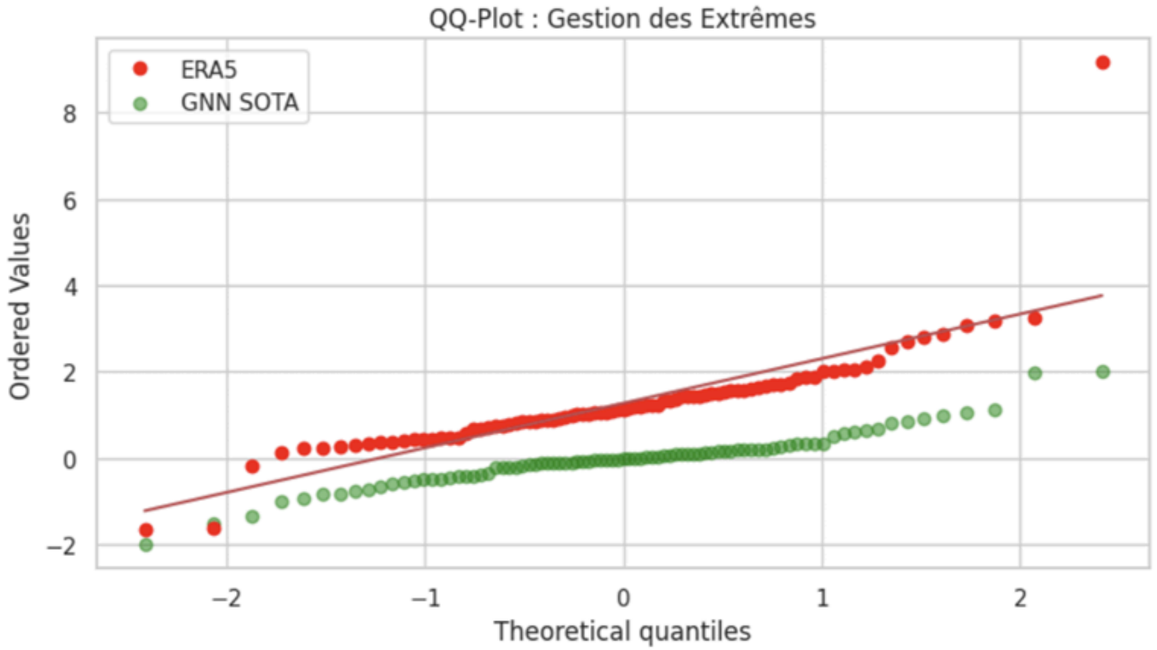
$$\text{Feat}_{\text{phys}} = \frac{\exp(-\sqrt{H})}{\text{NDVI}} \times \text{Season\_Flag}.$$

GRAPH NEURAL NETWORK : GNN

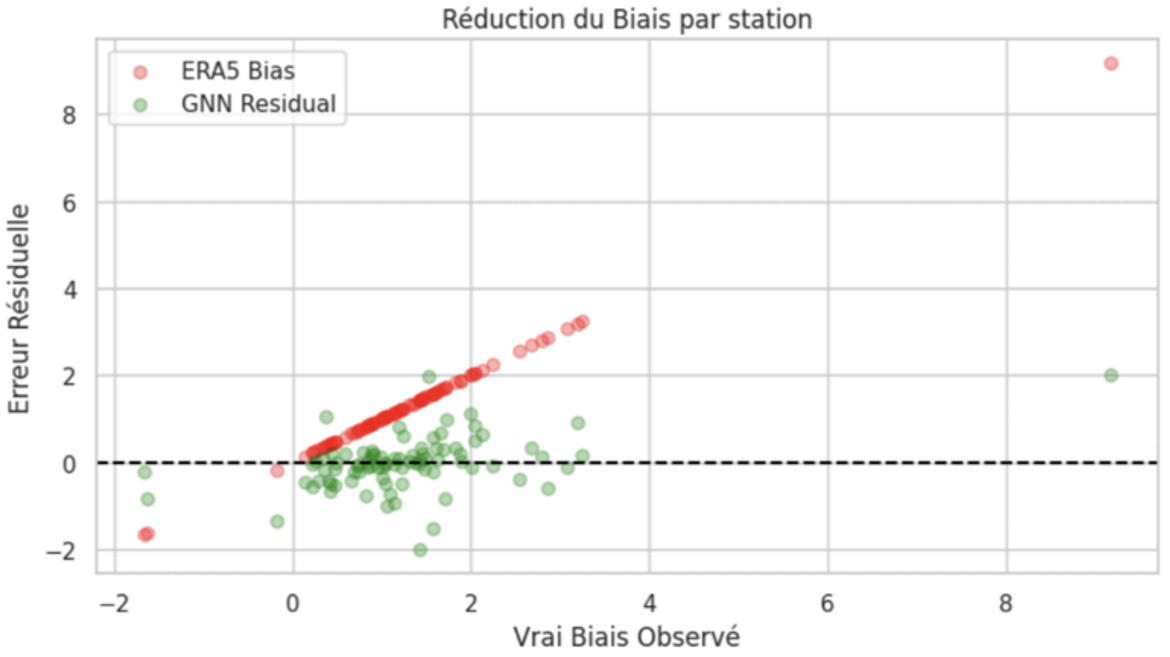
Predictive Correction Models



(b) Error distributions centered on zero



(c) QQ-Plot: extreme-value behavior



(d) Station-level bias reduction

Thank You