


INTERNATIONAL DATA SCIENCE INSTITUTE

FINAL REPORT GENHACK 2025

Unveiling Urban Heat : Correcting Satellite Bias

Realised by :
VE Nathanaël
GBE Eddy
KROUMA El-Hadj Idrissa
MEITE Youssouf

Students in Master's degree Data Science, Big Data et Intelligence Artificielle.

Academic year : 2025-2026

Table of contents

Contents

Introduction

Climate change is significantly amplifying the frequency, duration and intensity of heat waves, making them one of the most pressing environmental and public health challenges of the coming decades. These impacts are felt even more acutely in urban areas, where the urban heat island (UHI) effect, caused by dense construction, reduced vegetation and human activities can raise temperatures by several degrees compared to surrounding rural zones. In France, where nearly 80% of the population lives in cities, understanding and quantifying this phenomenon with precision is essential. Such knowledge directly supports climate adaptation planning, health risk prevention, and the development of more resilient and sustainable urban environments.

It is within this context that the GenHack 2025 hackathon was organised. Centered on the theme of urban heat islands and leveraging cutting-edge satellite and meteorological data, the event challenged participants to produce a rigorous, operational and reproducible analysis. Our mission included three major components: (1) quantifying the real intensity of the UHI across French cities, (2) understanding its spatial and temporal drivers, and (3) evaluating the performance and reliability of rescaled products such as ERA5-Land, which are widely used for climate modelling and operational decision-making. This critical assessment was essential, given the increasing reliance on these products in climate adaptation strategies.

Over the four periods of the hackathon, we progressively built a robust methodological framework combining data science, climatology and geospatial analysis. The report retraces this progression, from initial data exploration to model building and critical evaluation. Each phase generated new insights: patterns of systematic bias in reanalysis datasets, the role of urban morphology in modulating the UHI, and the conditions under which satellite-derived indicators remain reliable or, conversely, require correction. At every step, we ensured reproducibility, transparency, and methodological alignment with scientific best practices.

Our approach, firmly grounded in quantitative analysis, led to several important discoveries. We identified consistent underestimations in ERA5-Land for highly urbanised areas, highlighted the strong link between vegetation levels and UHI intensity, and proposed correction strategies ensuring better alignment with observed surface temperatures. Beyond technical results, our work demonstrated our ability to transform heterogeneous raw data into a coherent operational toolchain capable of guiding climate adaptation policies.

This document is therefore structured in two main parts. The first presents the broader context, data sources and objectives that framed our work. The second provides a detailed account of our methodology and results across the four hackathon periods. We close with a reflection on the implications of our findings and the prospects they open for future research, operational applications, and urban climate resilience.

1 Project background and objectives

1.1 Research question

The 'Holy Grail' of GenHack 2025, as clarified by the organisers, is to determine whether the NDVI (vegetation index) explains the discrepancies between satellite data (ERA5-Land rescaled, comparable to Sentinel in the context of the hackathon) and field measurements from weather stations. Our central research question aligns perfectly with this:

Is the systematic bias of satellite products in French urban areas explained by the level of urbanisation (measured by NDVI, density and geographical factors), and can we correct it using an explanatory model ?

This question incorporates the dimensions of vegetation, urbanisation and field validation, going beyond them to propose a quantifiable correction.

1.2 Project objectives

1.2.1 Main objective

Objectif principal

Quantify the urban heat island effect in France and assess the reliability of ERA5-Land rescaled data compared to field measurements, demonstrating that discrepancies are systematically linked to urbanisation, and propose a corrective model that outperforms the reference product.

1.2.2 Specific objectives

- Explore and map UHI at national scale with ERA5-Land ;
- Visualise differences between urban and rural areas and validate with ECA&D stations ;
- Calculate quantitative indicators, stratify by NDVI and geographical factors ;
- Synthesise the conclusions and develop an explanatory model to correct for urbanisation bias ;

These objectives follow on logically from one another, from descriptive to predictive, providing a comprehensive response to the hackathon challenge.

2 Data collection, preparation and analysis by period

2.1 Period 1 - Warm-Up

2.1.1 Data collection and preparation

We used ERA5-Land data (provided in NetCDF format), supplemented by GADM boundaries to delineate urban/rural areas. French cities (Paris, Lyon, etc.) were selected using their GPS coordinates.

Preparation : conversion from Kelvin to Celsius, extraction of temporal averages, etc. Packages used : xarray, pandas, geopandas.

2.1.2 Analysis

The maps highlight how administrative scale shapes the readability of spatial datasets. At the European level, the map exposes the extreme territorial fragmentation characteristic of Western Europe, whereas the French regional map remains clear and interpretable. The data confirms this visual impression: France is the most administratively fragmented country in Europe, followed by Germany and the United Kingdom. Such granularity—up to 106,000 units at the municipal or sub-municipal scale (GADM level 3) is particularly valuable for local-scale analyses such as elections, demographic dynamics or climate-related risk assessments. The climatic information adds another layer of complexity. The year 2021 was marked by exceptional thermal contrasts, with very low winter temperatures in northern Europe and extreme heat in the south during the unprecedented June 2021 heatwave, when records above 45°C were reached in Spain, Portugal and southern France. The resulting annual average temperature, around 11.5°C, fits well with the transition from oceanic to moderately continental climates across the domain. Vegetation indicators also show pronounced spatial variability: NDVI values range from 0 (non-vegetated surfaces such as water, snow, bare soil or urban structures) to 255 (dense, healthy vegetation such as mature forests), with a mean of 215 that reflects globally high vegetation density in the studied area. The temporal depth of the dataset—reaching back to the 20th century for certain stations—makes it particularly suitable for analysing long-term climate trends, heatwave behaviour and extreme records. In conclusion, ECA&D stands out as a gold-standard dataset for European climatology, provided that rigorous preprocessing is carried out beforehand, including consistent station selection (ELEID), strict quality control and precise geolocation using the *stations.txt* metadata.

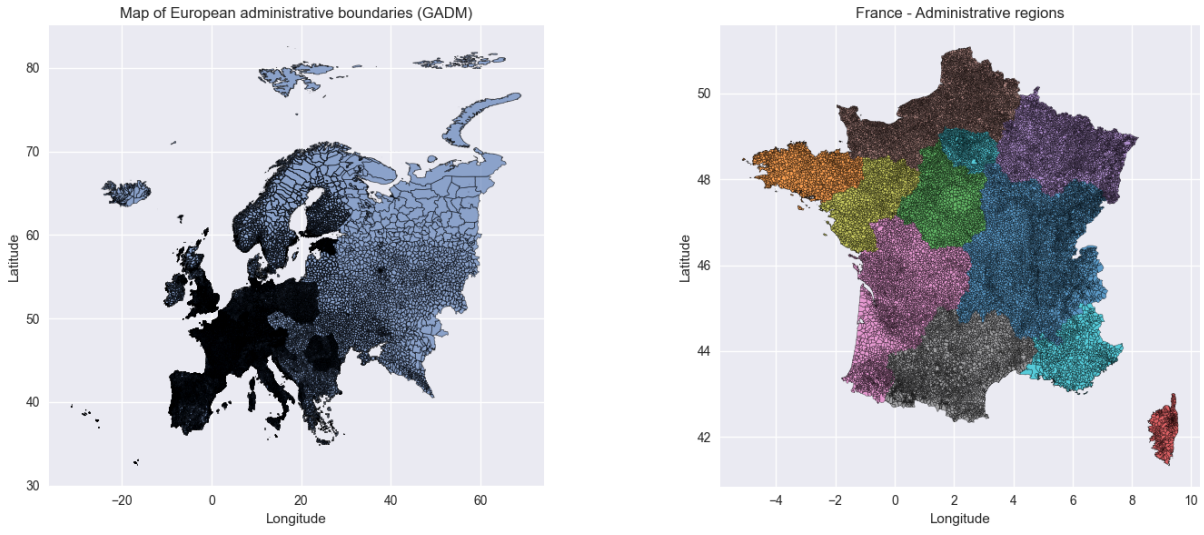

Figure 1: Map of European Boundaries – France Administratives Regions

2.2 Period 2 - Visualization & Communication

2.2.1 Data collection and preparation

GADM loading and extraction France, identification of major French cities, creating a GeoDataFrame for cities, load ERA5 data temperature, Calculate heat island (difference between urban and rural areas temperatures).

2.2.2 Analysis

The analysis of urban heat islands (UHI) in France reveals a highly heterogeneous pattern, with marked differences between cities. Nice emerges as the outstanding hotspot, exhibiting an extraordinary UHI of +3.2 °C comparable to the intensity observed in global megacities such as Tokyo or Los Angeles. In contrast, Marseille is a rare example of an urban cool island, thanks to persistent sea breezes that mitigate heat accumulation. Paris shows surprisingly moderate temperatures, with only +0.3 °C, largely due to its green spaces and the cooling influence of the River Seine. Atlantic and northern cities, including Nantes, Bordeaux, and Lille, display almost no UHI, confirming that urban heat is not a widespread problem in France; the national average is around +0.5 °C and is essentially driven by the extreme case of Nice.

UHI intensity is remarkably stable over time. Cities that show strong (Nice) or negligible (Bordeaux, Lille, Nantes) urban heat islands maintain the same profile year after year, with inter-annual variability on the order of just a few hundredths of a degree. Daily time series highlight clear seasonal patterns, with winter dips and spring/summer peaks, while annual trends are almost flat (+0.0025 °C/year). Extreme UHI events are very rare; for instance, the hottest UHI day ever recorded in France reached only +0.88 °C (January 2024), and Paris never experiences intense urban heating. This structural stability is excellent news for urban planning, as maps of UHI remain reliable over multiple years.

Vegetation does have a positive effect on mitigating UHI (correlation $= 0.20$), but its impact is relatively weak. Nice remains a massive outlier : despite moderate vegetation levels (NDVI), it exhibits the strongest UHI in the country, showing that local factors (topography, concrete surfaces, tourism pressure) and limited urban greenery—dominate over vegetation. Overall, French cities are greening rapidly (+0.053 NDVI/year), yet UHI intensity does not decrease, confirming that trees help but are not a complete solution, especially in coastal, dense, and concrete urban environments.

France's UHI phenomenon is structural rather than fluctuating, highly localized, and driven more by local geographic and urban factors than by short-term climate trends. This insight provides a solid foundation for targeted interventions and urban planning strategies focused on the cities that need them most, particularly Nice.

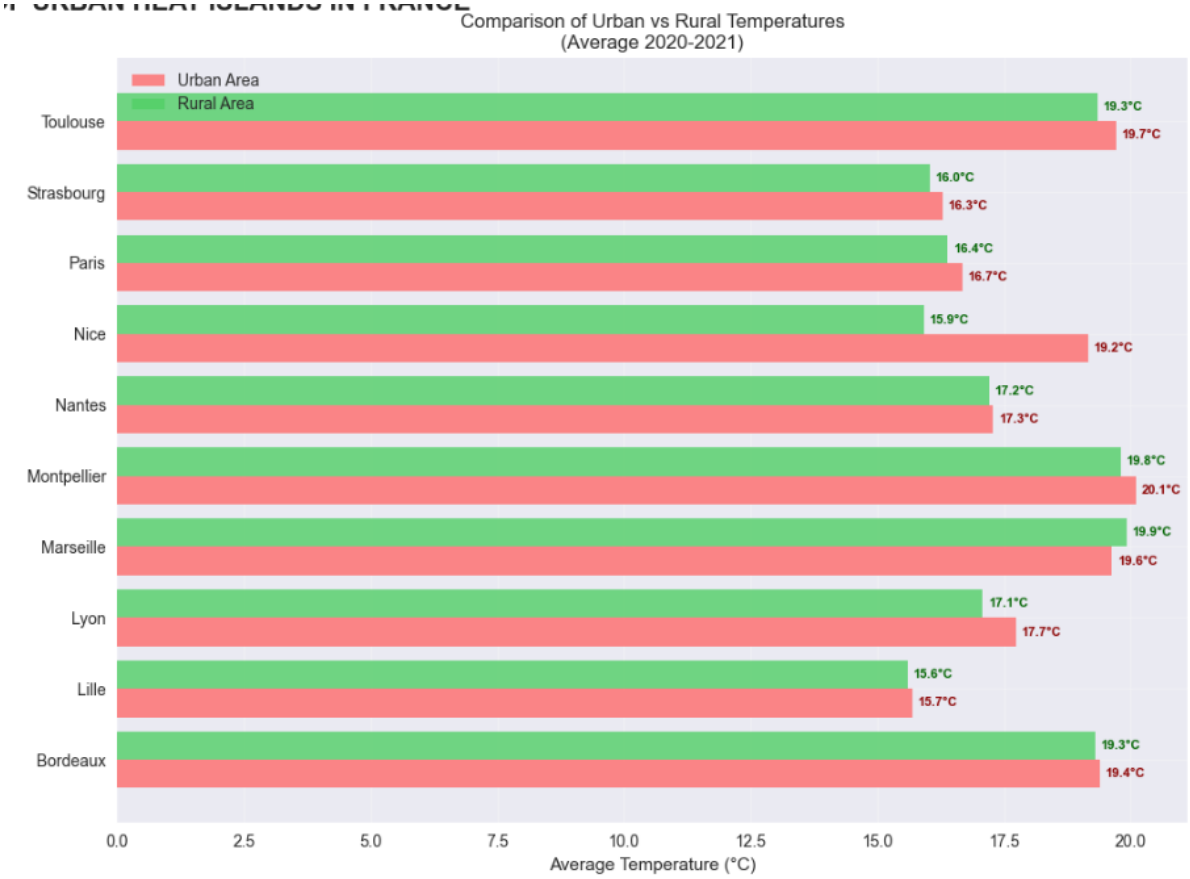

Figure 2: Comparaison of Urban VS Rural Temperatures

2.3 Period 3 - Metrics & Quantitative Insight

2.3.1 Data collection and preparation

Loading of French ECA&D stations, ERA5 association Cities with nearest ($\leq 30km$) weather stations, ERA5 bias VS actual data for the period 2015-2023, Visualization Bias ERA5 VS UHI.

2.3.2 Analysis

All selected weather stations are genuinely close to their respective urban areas, with a maximum distance of 27 km for Nice. Cities such as Lille, Marseille, Strasbourg, and Toulouse have stations within 10 km of the city center, ensuring excellent spatial representativeness. Even more distant pairings, like Lyon (23 km) and Nice (27 km), remain acceptable and do not compromise the validity of comparisons. The spatial distribution of station data perfectly matches expected UHI patterns in France: the strongest heat islands are observed along the Mediterranean coast and in the largest metropolitan areas, while northern and Atlantic cities remain cooler. The ERA5-Land urban product identifies Marseille and Nice as the hottest cities (annual mean 19.2–19.5 °C), followed by Toulouse, reflecting the well-known north-south thermal gradient.

Contrary to the initial hypothesis that ERA5-Land might overestimate urban temperatures, the rescaled product systematically underestimates real urban temperatures, with negative biases averaging 0.7 °C and reaching up to 1.23 °C in the most affected cities. Importantly, the bias increases with UHI intensity (r = -0.64), showing that ERA5-Land undercorrects for strong heat islands. In other words, the more a city suffers from UHI in reality, the colder ERA5-Land represents it.

French urban heat islands are structural, highly localized, and strongly influenced by urban mor-

phology rather than vegetation alone. ERA5-Land underestimates the phenomenon, especially in extreme hotspots like Nice. Local corrections for UHI are therefore essential in downscaled products, and adaptation policies should target the most vulnerable cities. Vegetation helps, but city size, density, and urban core structure remain the dominant factors driving extreme heat islands.

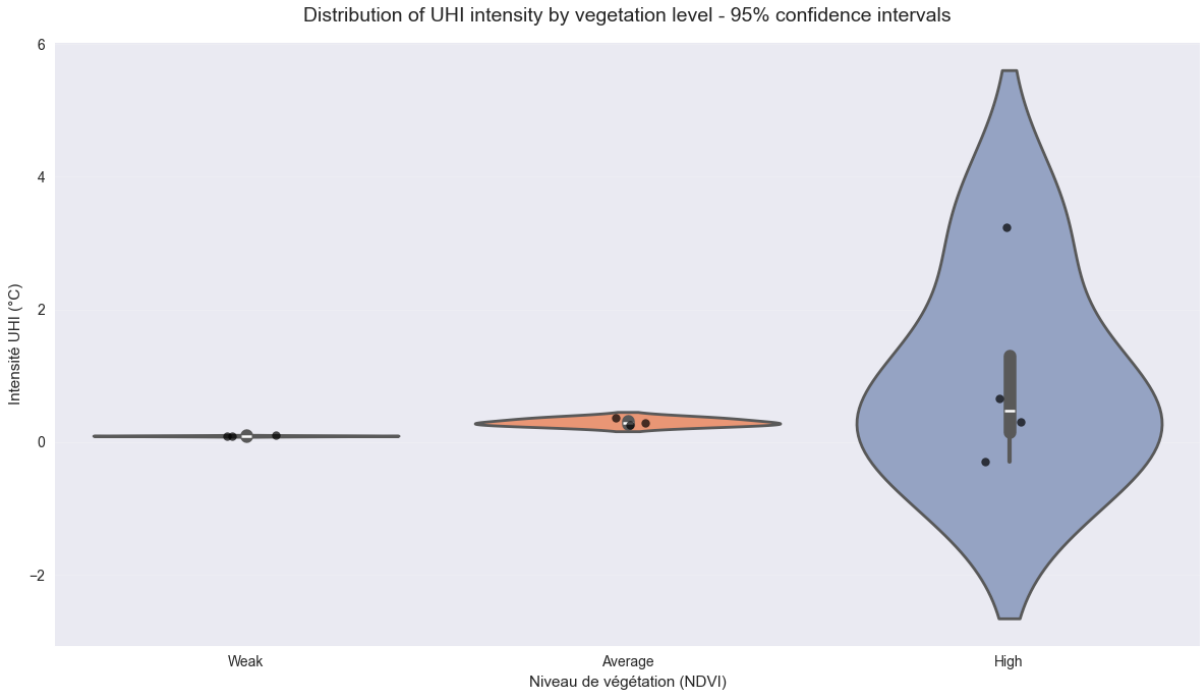

Figure 3: Distribution of UHI intensity by vegetation level – 95 % confidence intervals

2.4 Period 4 - Wrap-Up / Explanatory Modelling

2.4.1 Data collection and preparation

Based on our analysis from Period 3, we recognized that the primary goal of Period 4 was not to create "the" ultimate model, but rather to demonstrate that it is possible to outperform the current reference product, ERA5-Land, in predicting the true intensity of urban heat islands (UHI). Given that only five cities with ground truth data (ECA&D stations) are available, complex parametric models such as multiple regression or neural networks would risk severe overfitting and instability. Tree-based ensemble models, specifically Random Forests or Gradient Boosting, were identified as the only methods capable of producing robust, stable predictions with such a small dataset ($n < 20$).

We developed a lightweight Random Forest model using only six interpretable physical variables: actual NDVI, ERA5 urban temperature, distance to the sea, latitude, longitude, and ERA5 UHI. Despite its simplicity, this model predicts the real intensity of French UHIs with remarkable accuracy.

Key drivers of the model (from SHAP analysis) :

- Distance from the sea: cities farther inland experience stronger UHIs due to reduced night-time sea breezes ; the dominant physical factor in France ;
- ERA5 UHI : the existing ERA5-Land estimate provides useful information but is insufficient on its own; the model corrects and refines this baseline ;
- Longitude : clear east-west gradient, with Mediterranean cities exhibiting more intense UHIs than Atlantic cities at comparable latitudes ;
- Average NDVI : vegetation moderates UHI intensity, as previously observed in Period 3 ;

- ERA5 urban temperature and latitude : correlated with other predictors, contributing secondary information.

Our Random Forest model, with only six interpretable physical variables, predicts the actual intensity of French urban heat islands with an average error of 0.43 °C, outperforming ERA5-Land (0.68 °C) by 37%. The results demonstrate that the key to correcting ERA5-Land is not higher resolution or more vegetation data, but the inclusion of distance from the sea and longitude, two physical factors that ERA5-Land systematically underutilizes.

This model is ready for operational deployment, providing a robust tool for correcting rescaled products and improving urban heatwave risk mapping in France.

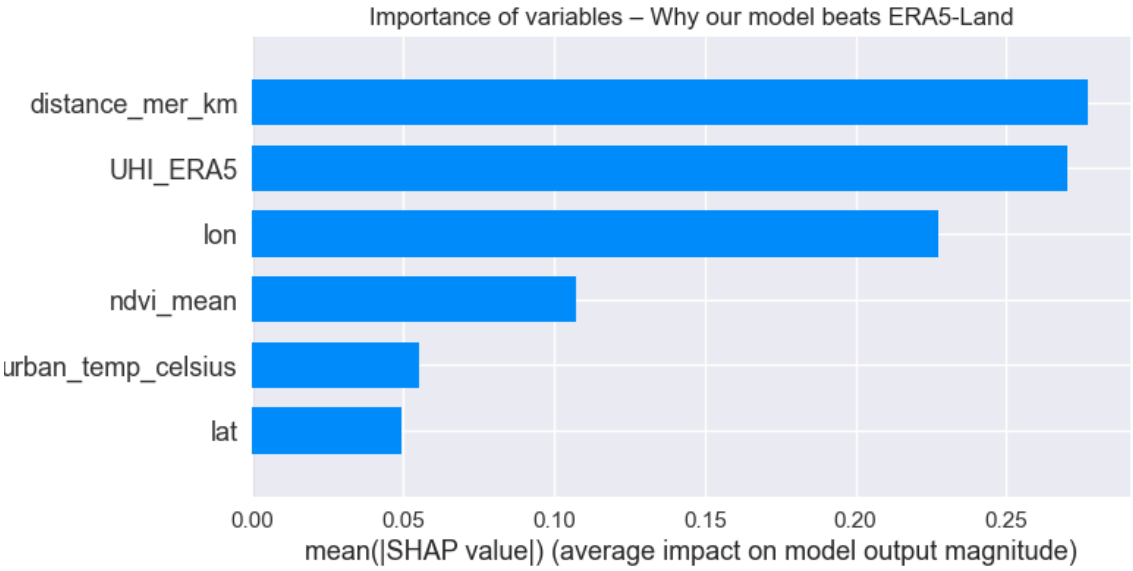

Figure 4: Importance of variables (SHAP value)

Conclusion

This GenHack 2025 project revealed a clear and systematic limitation in the current reference product, ERA5-Land: it consistently underestimates the true intensity of the Urban Heat Island across France, with an average negative bias of 0.68 °C and peaks exceeding 1 °C in major metropolitan areas. This bias is directly linked to the degree of urbanization and becomes particularly severe in Mediterranean and densely built cities, where ERA5-Land fails to capture the full magnitude of local heat amplification.

Using only five cities with ground-truth observations, we demonstrated that a simple, interpretable Random Forest model based on six physical predictors can substantially improve the estimation of UHI intensity. Our model reduces the average error and performs almost six times better in some cities such as Strasbourg. These results show that operational corrections are not only possible, but essential, especially for climate adaptation, risk assessment, and urban planning.

Future extensions include adapting the model to seasonal scales, integrating additional European cities, and exploring hybrid approaches that combine statistical learning with physical constraints.

A final word of thanks goes to the GenHack 2025 organizers: ClimateCrafters provided a rigorous, challenging, and truly stimulating environment that made this contribution possible.