Genhack 4 - Hackathon for generative modelling

Authors: Edoardo Olivieri[1], Giacomo Fullin[2], Luca Iaria[2], Paolo Portanova[2] and Francesca Verna[3]

[1]*Politecnico di Milano, Master's Degree in Mathematical Engineering*
[2]*University of Milano-Bicocca, Master's Degree in Statistical and Economic Sciences*
[3]*University of Milano-Bicocca, Master's Degree in Data Science*

Abstract

GenHack 2025 is a four-week hackathon on Urban Heat Islands (UHI) and climate downscaling, based on combining ERA5-Land, Sentinel-2 NDVI and ECAD station data. Our team studied when and where ERA5-Land underestimates summer daily maximum temperatures relative to ground stations, and which environmental variables control this bias.
We defined a simple UHI proxy as the difference between station and ERA5-Land daily maximum temperature and built a harmonised summer dataset (June-August, 2020-2023) for several thousand European stations. Using a regression model, we quantified how precipitation, wind speed, vegetation, time within the summer season, geographic coordinates and altitude shape this temperature bias.
ERA5-Land generally underestimates urban summer maxima, with the largest positive biases on hot, dry, low-wind days. A regression model explains about 55 % of the variance at European scale, with wind speed as the main cooling factor and precipitation, vegetation, season progression and latitude tending to increase the bias. Residuals display strong spatial structure, especially in mountainous regions, and stratifying the model by elevation (stations below and above 250 m) slightly improves performance and reveals coefficient changes between low- and high-altitude regimes. This highlights both the potential and limitations of linear models and points to the value of richer environmental descriptors and spatial or fixed-effects approaches.

Keywords: *Urban Heat Island, ERA5-Land, ECAD, NDVI, Climate Downscaling, Bias, Europe*

1. Introduction

Urban Heat Islands (UHIs) arise because cities modify the surface energy balance through dense construction, reduced vegetation, anthropogenic heat emissions and altered radiative properties. As a result, air temperatures in urban areas can be several degrees higher than in their rural surroundings, particularly during summer heatwaves and at night. Understanding and quantifying this effect is crucial for impact assessment, public health and climate adaptation planning.

The GenHack 2025 hackathon challenges participants to explore the UHI effect at European scale using three complementary data sources: ERA5-Land reanalysis fields, Sentinel-2 NDVI maps and ECAD ground-station observations. Teams are evaluated over four weekly periods, and the final deliverables include a comprehensive report, a slide deck and a public code repository.

Our team set out to answer the following questions:

- How large is the discrepancy between ERA5-Land daily maximum temperatures and ground-station observations across Europe?
- Can this discrepancy be interpreted as a proxy for UHI intensity in urban and peri-urban environments?
- Which environmental and geographical variables most strongly control the bias?
- To what extent can a global linear model explain these differences, and where does it fail?
- How does model performance change when we account for geographic heterogeneity, for example through elevation-based stratification?

The remainder of this report is organised as follows. Section 2 describes the datasets and the construction of our UHI proxy. Section 3 outlines the methodological pipeline, from station selection and data extraction to regression modelling. Section 4 presents the main empirical findings, including case studies, global models and elevation-stratified analysis. Section 5 discusses limitations and implications, and Section 6 outlines possible extensions.

2. Data and Study Area

2.1. Data sources

All datasets were provided through the GenHack shared infrastructure and documented in the Kayrros data user guide. The four core data sources are:

- Administrative boundaries (GADM).
- Meteorological reanalysis (ERA5-Land).
- Vegetation index (Sentinel-2 NDVI).
- Ground stations (ECAD).

We used the GADM 4.1 database of global administrative areas, restricted to Europe. The data are provided as a GeoPackage file containing national and subnational polygon boundaries, which we used for mapping, spatial context and country-level subsetting.

ERA5-Land is a high-resolution land surface reanalysis produced by ECMWF. The hackathon dataset consists of daily fields over Europe (approximately 9 km resolution) from January 2020 to October 2025. For each day we use:

- daily maximum of the hourly 2-metre air temperature (K);
- daily mean total precipitation;
- daily mean wind components (ms^{-1}).

From these, we derived the scalar mean wind speed as

$$\text{Wind\_Speed} = \sqrt{u^2 + v^2}.$$

Vegetation conditions are characterised by the Normalised Difference Vegetation Index (NDVI), derived from Sentinel-2

satellite imagery at 80 mt resolution. NDVI values are stored as `int8` in quarterly GeoTIFF files and require a linear rescaling from the integer range $[0, 254]$ to the physical range $[-1, 1]$, with 255 reserved for missing data. For GenHack 2025, NDVI is available for Europe from 2020 to 2023 with quarterly temporal resolution.

The ECAD project provides daily observational series from thousands of European weather stations. The hackathon focuses on daily maximum temperature (TX), stored in separate text files, and accompanied by a metadata file, listing station IDs, names, coordinates and altitude above sea level.

2.2. Study period and spatial domain

We focused on the recent period from 2020–2023, for which all four datasets overlap, and restricted attention to boreal summer (June–August) when UHI effects on maximum temperature are strongest. Spatially, we considered all ECAD stations within the European GADM polygons. In practice this resulted in a dense network of several thousand stations spread across Western, Central and Southern Europe, Scandinavia and parts of Eastern Europe.

2.3. Definition of the UHI proxy

Following the hackathon guidelines and our own physical intuition, we defined a simple proxy for UHI intensity as the discrepancy between local ground-station data and the co-located ERA5-Land cell:

$$\mathrm{Bias}_{i,t} = T_{\mathrm{Station},i,t} - T_{\mathrm{ERA5},i,t}, \tag{1}$$

where $T_{\mathrm{Station},i,t}$ is the daily maximum temperature recorded at station i on day t, and $T_{\mathrm{ERA5},i,t}$ is the daily maximum temperature from ERA5-Land at the corresponding location and date. Positive values indicate that the station is warmer than the reanalysis, which we interpret as a manifestation of UHI intensity.

3. Methodology

3.1. Overall pipeline

Our workflow, developed iteratively over the four hackathon periods, can be summarised in three main stages:

1. **Station selection and metadata construction.** We parsed all ECAD station files, identified valid stations with continuous data in the 2020–2023 period, and combined these with station coordinates and altitude.
2. **Vectorised extraction and dataset integration.** For each selected station we extracted co-located ERA5-Land variables and Sentinel-2 NDVI, then constructed a harmonised summer dataset with one row per station-day.
3. **Exploration and modelling.** We performed exploratory visualisations for selected cities, built European-scale regression models, inspected residual patterns and finally refined the model through elevation-based stratification.

All steps are implemented in a Jupyter notebook.

3.2. Station selection and coverage analysis

We began by reading the ECAD `stations.txt` metadata file and converting station coordinates from degrees-minutes-seconds (DMS) to decimal degrees. We constructed a GeoDataFrame of all stations and intersected it with the European GADM layer to retain only stations inside our study region. This also allowed us to associate each station with a country code for later stratified analysis.

For each station, we then parsed the corresponding file to identify the first and last year with valid daily maximum temperature measurements. Using this metadata, we created a binary *coverage flag* indicating whether the station provides continuous data from January 2020 to December 2023. Only stations with full coverage were retained for our downstream analysis, ensuring a consistent and balanced dataset over time.

Finally, we produced diagnostic summaries (using the `skimpy` package) and spatial plots of station density across Europe, which confirmed a good coverage in Western and Central Europe and somewhat sparser networks in mountainous and northern regions.

3.3. Construction of the summer bias dataset

3.3.1. Merging station data with ERA5-Land

For all stations with full coverage, we extracted daily ERA5-Land fields at the station coordinates. Using `xarray`, we loaded the relevant NetCDF files for the 2020–2023 period and interpolated the gridded fields to the station locations along the spatial dimensions. We then merged the interpolated ERA5-Land temperatures, precipitation and wind components with the station TX series, aligning by station ID and date.

At this stage we computed:

- daily mean wind speed from the u and v components;
- the temperature bias $\mathrm{Bias}_{i,t}$ as defined in Eq. (1);
- the calendar month and other calendar features.

We restricted the dataset to summer months (June–August), as UHI effects on maximum temperature are most pronounced in this season and our slides and analyses are focused on this interval.

3.3.2. Sampling NDVI at station locations

To characterise land-cover and vegetation, we sampled Sentinel-2 NDVI at each station location. For each year from 2020 to 2023 we loaded the corresponding June–September NDVI composite GeoTIFF, reprojected station coordinates into the raster CRS if necessary, and sampled the NDVI pixel value at each station.

The NDVI values were stored as integers; we converted them back to physical units using the linear transformation

$$\mathrm{NDVI} = \frac{2}{254}\,\mathrm{NDVI}_{\mathrm{raw}} - 1,$$

treating the special value 255 as missing. The resulting annual NDVI value was then merged with the summer dataset based on station ID and year, providing a simple measure of local vegetation greenness around each station.

3.3.3. Final feature set

The final dataset used for modelling contains the following variables for each station-day:

- **Response:** Bias (station minus ERA5-Land daily maximum temperature in celsius);
- **Predictors:**
  - Precipitation (daily mean);

- Wind_Speed (daily mean);
- NDVI (annual);
- Month (6, 7 or 8);
- Latitude and longitude (decimal degrees);
- HGHT (station altitude above sea level, in metres).

We removed records with missing values in any of the predictors or the response. Outliers in the temperature bias were diagnosed via boxplots and histograms; we truncated the bias distribution to remove extreme values while keeping the bulk of the observations, thereby increasing the robustness of the regression fits.

3.4. Period 2: Visualisation and city case studies

The second hackathon period was dedicated to visualisation and communication. We selected two coastal cities, Barcelona (Spain) and Rimini (Italy), as illustrative case studies. For each city, we identified the closest ECAD station within the corresponding administrative polygon and extracted both the station TX series and the co-located ERA5-Land daily maximum temperatures.

We produced daily time series plots for 2020–2023, showing that ERA5-Land largely captures the seasonal cycle and interannual variability but systematically underestimates the observed peaks. Particularly in Barcelona, a persistent positive offset is visible in summer and during heatwaves, with the station repeatedly recording maximum temperatures a few degrees higher than ERA5-Land. Rimini exhibits similar systematic bias but with larger day-to-day variability and more pronounced spikes, likely influenced by its coastal and mesoscale environment.

To explicitly visualise the UHI proxy, we plotted the daily bias time series for both stations. The values cluster predominantly above zero, confirming the presence of a persistent UHI signal. Barcelona exhibits a relatively stable positive bias of about 2-3 °C, whereas Rimini shows large swings, from strongly positive values (up to roughly +6 °C) to occasional negative excursions.

We then explored how meteorological conditions modulate the bias in Barcelona. Scatter plots of bias against daily mean wind speed revealed a clear "triangular" pattern: large positive biases only occur at very low wind speeds ($< 2\ \mathrm{ms}^{-1}$), while for wind speeds above roughly $4\ \mathrm{ms}^{-1}$ the bias collapses towards zero. In other words, strong winds act as an efficient mixer that ventilates the urban boundary layer and suppresses the UHI effect.

A similar analysis for daily precipitation showed that the largest positive biases occur almost exclusively on dry days, while days with substantial rainfall tend to have smaller and more tightly clustered biases. This is consistent with the role of evaporative cooling and enhanced atmospheric mixing during rain events.

These visual analyses provided a physically interpretable story: ERA5-Land underestimates urban maximum temperatures mainly on hot, dry, stagnant days, and meteorological conditions such as wind and precipitation are key modulators of UHI intensity.

3.5. Period 3: Metrics and European-scale regression

In the third period we moved from visual insights to quantitative modelling. Our goal was to estimate how much of the bias variance could be explained by the available predictors and to quantify their marginal effects.

3.5.1. Correlation structure

We first examined the correlation matrix among the predictor variables. Pairwise correlations were generally low in magnitude, indicating that the features carry relatively independent information about the environment and geography. However, this also suggested that each individual predictor might only explain a modest fraction of the variability in the bias, and that high R^2 values would be difficult to achieve with a simple linear model.,

3.5.2. Baseline linear model

Our baseline model was a multivariate linear regression without intercept:

$$\begin{aligned}\text{Bias} = {} & \beta_1 \cdot \text{Precipitation} + \beta_2 \cdot \text{Wind\_Speed} + \beta_3 \cdot \text{NDVI} \\ & + \beta_4 \cdot \text{Month} + \beta_5 \cdot \text{Latitude} + \beta_6 \cdot \text{Longitude} \\ & + \beta_7 \cdot \text{HGHT} + \varepsilon. \end{aligned} \tag{2}$$

Fitting this model over the full European summer dataset yields an R^2 of approximately 0.55, meaning that the selected predictors jointly explain about 55% of the variance in the temperature bias. While far from perfect, this confirms that the bias is not random but systematically related to environmental and geographical variables.

3.5.3. Coefficient interpretation

The estimated coefficients are consistent with the qualitative insights from the city case studies and with physical expectations:

- **Wind speed** has a negative coefficient of relatively large magnitude, confirming its role as the main cooling factor: stronger winds reduce the ERA5-Land bias by flushing heat out of the urban canopy.
- **Precipitation** has a small positive coefficient, indicating that, on average, days with more rainfall tend to have slightly higher bias. This effect is subtle and likely reflects complex interactions between rainfall events, soil moisture and subsequent clear-sky days.
- **NDVI** enters with a positive coefficient in the baseline global model, implying that more vegetated surroundings are associated with higher station–ERA5 differences. This may partly capture the fact that vegetated areas tend to be cooler than urban cores, so a station located in a greener pixel within an otherwise urbanised region may record temperatures that diverge from the coarse ERA5 grid-box average.
- **Month** has a positive coefficient: as summer progresses from June to August, the mean bias tends to increase, consistent with stronger UHI effects during late summer and heatwaves.
- **Latitude** is positively associated with bias, suggesting that more northern locations experience slightly larger positive discrepancies, possibly due to differences in land-cover representation or snow-albedo processes in ERA5-Land.
- **Longitude** has a small negative coefficient, hinting at a mild west–east gradient in the bias pattern.

- **Altitude (HGHT)** shows a modest effect in the global model, but its role becomes clearer in the elevation-stratified analysis described below.

These findings were summarised and communicated in our period 3 slides, where we emphasised wind speed as a "killer" of the UHI effect and highlighted the need for richer, possibly non linear or spatial models to capture the residual structure.

3.5.4. Spatial pattern of residuals

To investigate where the global model performs poorly, we computed regression residuals and aggregated them by station. For each station we calculated the mean residual over all summer days and plotted these averages on a European map, overlaid on GADM country boundaries.

The resulting residual map displays strong spatial structure. Mountainous regions, such as the Alps and parts of Scandinavia, show systematic overestimation or underestimation patterns, whereas lowland areas exhibit more modest residuals. This clearly indicates that important geographical factors, particularly elevation and terrain complexity, are not fully captured by the global linear model and that a single set of coefficients is insufficient to describe the entire continent.

3.6. Period 4: Explanatory modelling and elevation stratification

The fourth period focused on explanatory modelling and model refinement. Motivated by the spatial residual patterns, we hypothesised that elevation is a key driver of model performance and that distinct regimes might exist at low and high altitudes.

3.6.1. Elevation-based stratification

Using station altitude, we examined its distribution and chose a threshold of 250mt to separate lowland from highland stations. This threshold approximately distinguishes flat or gently rolling terrain from mountainous regions and yields two reasonably balanced subsets:

- **Low-altitude group:** stations with HGHT $<$ 250mt;
- **High-altitude group:** stations with HGHT $\geq$ 250mt.

We then re-fit the same linear regression specification as in the global model separately on each subset. This approach keeps the feature space unchanged while allowing the coefficients to adapt to different elevation regimes.

3.6.2. Model comparison

The elevation-stratified models confirm that altitude strongly modulates the relationship between environmental variables and temperature bias.

For the **low-altitude** stations, the model attains an R^2 of about 0.587, slightly higher than the global model. In this regime, wind speed retains a negative effect, but its magnitude is moderate. NDVI tends to slightly reduce the bias, while month, latitude and altitude all contribute positively. This suggests that in lowlands the bias grows over the course of the summer and is generally stronger at higher latitudes and altitudes within the low-altitude range.

For the **high-altitude** stations, the model explains a slightly smaller fraction of variance ($R^2 = 0.562$), and several coefficients change markedly. Wind speed becomes a much stronger cooling factor, NDVI switches sign and now increases the bias, and the coefficient of HGHT turns negative, indicating that, within the high-altitude regime, higher stations tend to have smaller ERA5-Land biases. These shifts emphasise that the physical mechanisms and representation errors at high elevation differ substantially from those in the lowlands, possibly due to complex topography, valley effects, snow cover and local circulation patterns.

Bar plots of the estimated coefficients for both altitude groups further illustrate these differences, with wind speed and altitude showing the largest regime dependence. The comparison underscores that a single European linear model cannot generalise well across such diverse terrain and that stratification, whether by altitude, climate zone or land-cover type, is a promising direction for improving explanatory power.

3.6.3. Residual maps by altitude

We repeated the residual mapping procedure for each altitude group separately. In both cases, the mean residuals exhibit less pronounced large-scale structure than in the global model, supporting the idea that elevation stratification removes part of the systematic geographical bias. However, regional patterns remain, especially in complex mountainous areas, highlighting room for further model refinement.

3.7. Country-level models

As an intermediate step between a single European model and fine-grained station-level models, we also experimented with country-level regressions. In particular, we compared Norway, representing a predominantly mountainous country, and Germany, representing mostly lowland terrain. For each country we extracted the corresponding station subset, assembled the same set of features and fitted a linear regression model.

Although we do not report all coefficients here, the exercise revealed that both the mean bias and the sensitivity to environmental variables differ substantially between countries. Norway exhibits stronger altitude effects and more pronounced spatial residuals, whereas Germany's bias structure is more homogeneous. This supports the idea that geographical stratification, by country, region or topographic class, is essential for capturing local characteristics that are smeared out in a continental model.

4. Results

In this section we summarise the main empirical results emerging from our analysis, complementing the methodological description above.

4.1. Evidence for systematic ERA5-Land underestimation

Across Europe, the majority of stations exhibit a positive mean bias in summer, with station maximum temperatures exceeding ERA5-Land values by typically 1-3 °C. This pattern is particularly pronounced in urban or densely populated regions, lending support to the interpretation of the bias as a UHI proxy.

The Barcelona and Rimini case studies provide illustrative examples: in both cities, daily time series show persistent positive offsets during summer, and the distributions of the bias are heavily skewed towards positive values. These qualitative patterns match the quantitative regression findings, where

the intercept-free model nonetheless yields a predominantly positive predicted bias under typical summer conditions.

4.2. Drivers of the bias

The European-scale regression model attributes a substantial fraction of the bias variance to a few key drivers:

- **Wind speed** is the dominant cooling factor; large positive biases are only achievable under weak-wind conditions.
- **Precipitation** and **vegetation** exert subtler but non-negligible influences, modulating the heat storage and release in the urban fabric and surrounding landscape.
- **Temporal progression within summer** (month) captures seasonal evolution in both meteorology and human activity, with larger biases in late summer.
- **Geographical coordinates** encode residual large-scale gradients not fully explainable by the local predictors alone.
- **Altitude** plays a complex role, as revealed by the elevation-stratified models.

4.3. Geographical heterogeneity and elevation effects

The spatial structure of residuals and the different behaviour of low- and high-altitude models highlight strong geographical heterogeneity:

- Mountainous regions are systematically more difficult to model, reflecting unresolved topographic detail and local circulations in the ERA5-Land grid.
- High-altitude stations show stronger sensitivity to wind and opposite-sign relationships for altitude, pointing to different physical regimes.
- Even within lowlands, regional patterns remain, suggesting that land-use and station siting (e.g., urban core vs. rural surroundings) also matter.

Overall, the results confirm that while a simple linear model can capture major tendencies, local and regional factors, many of which are not explicitly included in our feature set, remain important.

5. Discussion

Our analysis demonstrates that the discrepancy between ERA5-Land and ECAD station temperatures is structured and interpretable. The UHI proxy we defined aligns with physical expectations: it increases on hot, dry, stagnant days and is modulated by vegetation and geographical context.

At the same time, several limitations must be acknowledged:

- **Resolution mismatch.** ERA5-Land operates at $\sim$ 9,km resolution, whereas ECAD stations sample very local conditions. A single grid cell may contain heterogeneous land uses (urban, rural, water), making it challenging to attribute discrepancies purely to UHI effects.
- **Simplified NDVI representation.** We use a single annual NDVI value per station-year, sampled at 80 mt resolution. This collapses temporal variability within the summer and ignores directional effects such as upwind land cover.
- **Linear model limitations.** The linear specification cannot capture potential non-linearities and interaction effects (e.g. the combined impact of low wind and high NDVI, or month-dependent wind effects). Our period 3 slides already highlighted the need for non-linear or spatial models.
- **Missing covariates.** Important factors such as urban fraction, building height, distance to coast, orographic exposure and station siting (urban core vs neighbourhood parks) are not explicitly included.
- **Measurement and metadata uncertainties.** Both ERA5-Land and station data are subject to measurement errors, quality-control issues and metadata inaccuracies, particularly regarding station relocation or instrument changes.

Despite these limitations, the models provide useful quantitative insight into how environmental variables influence ERA5-Land bias and where the reanalysis struggles the most.

6. Future Work

Building on the insights gained during the hackathon, we identify several avenues for future improvement, many of which were already outlined in our period 4 slides:,

- **Richer environmental descriptors.** Incorporate additional covariates, such as land-use categories, urban fraction, distance to coastline, topographic indices (slope, aspect), and possibly night-time lights as a proxy for urbanisation.
- **Fixed-effects models.** Introduce station-specific and temporal fixed effects (e.g. by station ID and year or month) to control for unobserved heterogeneity and focus on within-station variations.
- **Spatial and non-linear models.** Explore geographically weighted regression, random forests, gradient boosting or spatial autoregressive models to capture non-linearities and spatial dependence in the bias.
- **Finer stratification.** Beyond altitude, experiment with clustering stations by climate zone, Köppen classification or land-cover type, and fit separate models to each cluster.
- **Towards a UHI susceptibility index.** Building on our period 2 idea, develop a composite index that combines bias statistics with NDVI, wind and precipitation climatologies to map where and under which meteorological conditions strong UHI effects are most likely.

7. Reproducibility and Team Workflow

All analyses were conducted in a single, well-documented Jupyter notebook, using open-source Python libraries. The workflow is fully reproducible given access to the original datasets. The notebook includes:

- explicit paths and loading routines for all data sources;
- modular functions for station metadata construction, ERA5-Land and NDVI extraction;
- clear sections for exploratory plots, regression fitting and residual analysis;
- code to generate the main figures used in the weekly slide decks and in this report.

8. Conclusion

Within the GenHack 2025 framework, we investigated why ERA5-Land daily maximum temperatures differ from ECAD station observations and how this discrepancy can serve as a proxy for the Urban Heat Island effect. By integrating large-scale reanalysis, satellite-derived vegetation data and ground-station measurements, we showed that:

- the bias has a strong physical and geographical structure;
- wind speed, precipitation, vegetation, seasonal progression and location jointly explain a substantial fraction of its variability;
- residuals reveal the limitations of a single European linear model, especially in mountainous regions;
- elevation-based stratification provides a simple yet effective way to improve explanatory power and highlight regime-dependent behaviour.

Our results contribute to a better understanding of how urban environments and complex terrain interact with regional climate reanalyses. They also illustrate how relatively simple statistical tools, when combined with careful data integration and physical reasoning, can yield actionable insight into climate downscaling and UHI assessment.

References

[1] Edoardo Olivieri, Giacomo Fullin, Luca Iaria, Paolo Portanova and Francesca Verna, "Team_Oroscopo.ipynb", [Online]. Available: https://github.com/edoardo-olivieri/GenHack4-Team1/blob/main/Team_Oroscopo.ipynb.

[2] European Climate Assessment & Dataset, *ECA&D: European climate assessment & dataset*, Royal Netherlands Meteorological Institute (KNMI), 2024. [Online]. Available: https://www.ecad.eu/dailydata/predefinedseries.php.

[3] GADM, *Database of global administrative areas, version 4.1*, GADM project, 2024. [Online]. Available: https://gadm.org/.

[4] J. Muñoz Sabater, *ERA5-Land: Monthly averaged data from 1981 to present*, Copernicus Climate Change Service (C3S) Climate Data Store, 2019. [Online]. Available: https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=overview.

[5] Sentinel Hub, *Normalized difference vegetation index (ndvi) for sentinel-2*, Sentinel Hub custom scripts and documentation, 2024. [Online]. Available: https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/ndvi/.