GenHack 2025

Phase 4: Explanatory Modelling

Wrap-Up

ChefAI

Introduction

This presentation summarizes the analyses carried out over the past weeks on the Urban Heat Island (UHI) effect.

The work focused on computing a **pixel-based UHI index** using NDVI-derived rural references and other geomorphological data to **evaluate the accuracy of ERA5 temperature fields against ECA ground-station measurements**.

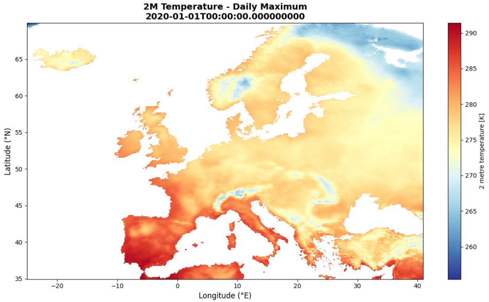
Through spatial discrepancy analysis, NDVI gradients, and distance-dependent validation, we assessed the ability of ERA5's coarse grid to capture micro-scale urban thermal variability.

The results highlight **systematic biases**, interpolation uncertainties, and **the need for denser observational networks** for reliable urban climate assessment.

Week 1: Data Exploration

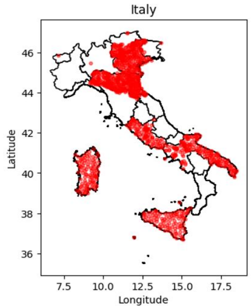
ChefAI

Exploration of ERA5-Land and ECA Blend Datasets:



The analysis began with two main datasets:

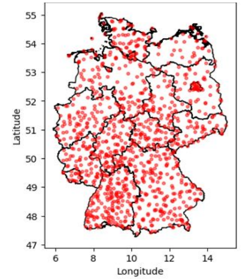
- ERA5-Land, a high-resolution 0.1° reanalysis product whose grid cell size varies with latitude,
- ECA Blend, a station-based dataset enriched with SYNOP reports to fill gaps.



The ECA station density analysis showed a **highly uneven distribution across countries**, with Italy, Germany, and Spain having the most stations.

This non-uniformity limits territorial analyses, especially in very small countries or those with too few stations to be statistically meaningful.

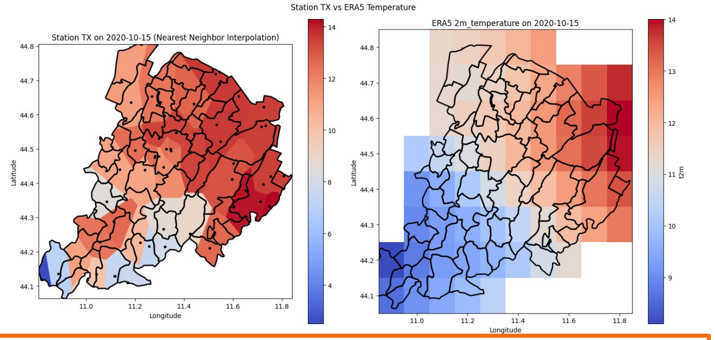
Although **some records date back to the 1760s**, these early observations are too sparse and distant in time to be useful for modern atmospheric analysis.



Week 1: Data Exploration

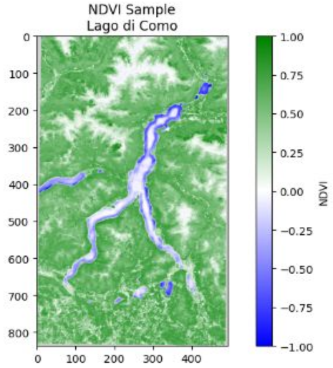
NDVI Dataset Exploration and ECA data interpolation:

A temperature grid was generated by interpolating ECA station data using a nearest-neighbor approach and compared with ERA5 2m temperatures over Bologna.



The comparison highlights how interpolation preserves sharper local temperature variations captured by stations, whereas ERA5, due to its coarse resolution, smooths these patterns.

This illustrates the fundamental difference between station-based interpolated surfaces and gridded reanalysis products.



The exploration also included the NDVI (Normalized Difference Vegetation Index) dataset.

The data is provided in GeoTIFF format, a standard for storing geospatial raster data.

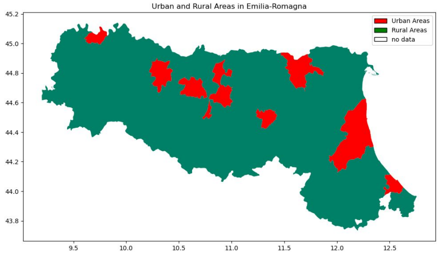
In this format, each pixel corresponds to a specific geographic location and contains a value representing the NDVI at that location.

Week 2: Visualization & Communication

Integration of DEGURBA dataset:

The second phase of the analysis incorporated the DEGURBA (Degree of Urbanisation) dataset to improve urban area identification, classifying regions as Cities, Towns and Suburbs, or Rural Areas.

NDVI alone is insufficient to capture urbanization for UHI studies, so DEGURBA was integrated, and ECA&D data were cleaned of outliers to ensure reliable meteorological inputs.



The meteorological analysis focused on daily maximum temperatures from ECA&D. Data were aggregated into weekly averages using a rolling window to reduce high-frequency noise and highlight underlying thermal trends. The UHI effect was quantified using two indices:

UHICI

normalized index that quantifies UHI intensity by integrating thermal, vegetation, and urbanization data.

$$\mathrm{UHICI} = N\big(w_1 \cdot N(T_{\max}) + w_2 \cdot (1 - N(\mathrm{NDVI})) + w_3 \cdot (1 - N(\mathrm{DEGURBA}))\big)$$

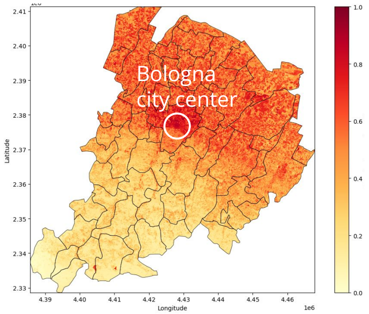
UHIDI

compares each pixel’s temperature to a rural mean baseline defined using DEGURBA Type 3 and NDVI > 0.5 vegetated areas.

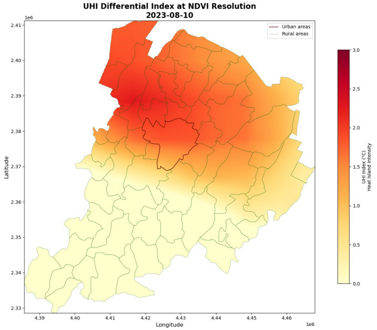
$$\mathrm{UHI}(p,d) = T_{\max}(p,d) - T_{\mathrm{rural}}(d)$$

Week 2: Visualization & Communication

UHICI



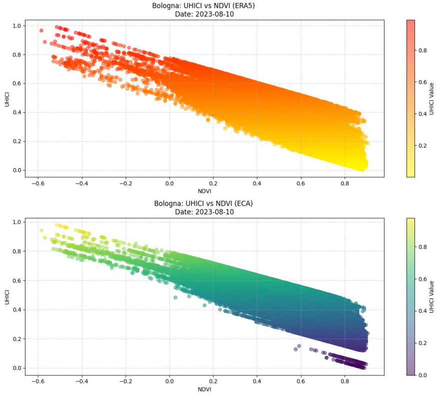
UHIDI


UHI Differential Index at NDVI Resolution
2023-08-10

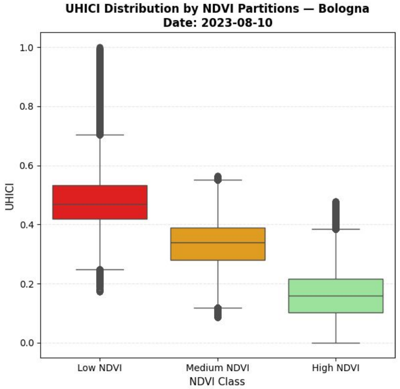
A final visualization stage was dedicated to mapping the defined UHI indicator. The urban temperature field was contrasted against the established rural baseline to spatially illustrate the thermal anomaly.

Our statistical evaluation confirms that higher UHICI values strongly correlate with lower NDVI.

This aligns with expectations: areas lacking vegetation show higher heat retention and more intense UHI effects.


Bologna: UHICI vs NDVI (ERA5)
Date: 2023-08-10

Bologna: UHICI vs NDVI (ECA)
Date: 2023-08-10

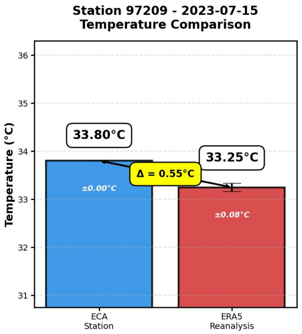

UHICI Distribution by NDVI Partitions — Bologna
Date: 2023-08-10

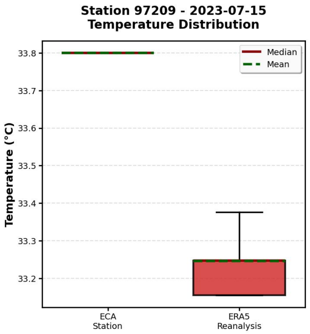
Week 3: Metrics & Quantitative Insight

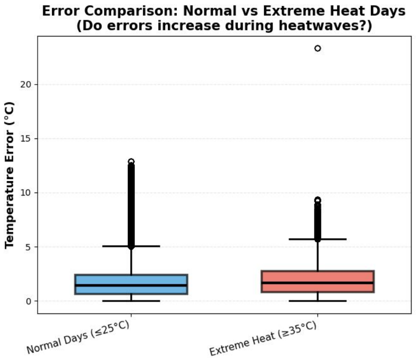
Discrepancies are systematic, not random.

Temperature bias

- Satellite systems have critical biases during summer heat waves, when accurate data is most crucial for public health decisions;
- ERA5 underestimates the temperature;
- Largest temperature errors occur in areas with NDVI < 0.5, corresponding to dense urban zones.


Station 97209 - 2023-07-15
Temperature Comparison


Station 97209 - 2023-07-15
Temperature Distribution


Error Comparison: Normal vs Extreme Heat Days
(Do errors increase during heatwaves?)

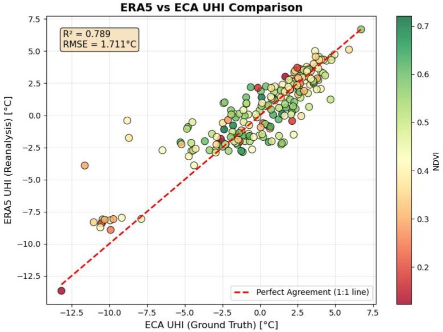
Week 3: Metrics & Quantitative Insight

UHI bias

$$\Delta_{\mathrm{UHI}} = |\mathrm{UHI}_{\mathrm{ERA5}} - \mathrm{UHI}_{\mathrm{ECA}}|$$

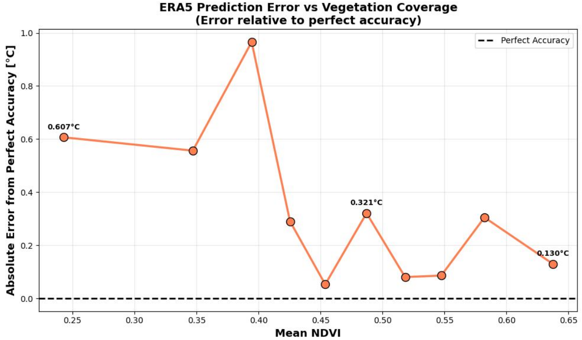
Errors are highly dependent on land cover, with urban stations showing much larger deviations than rural sites.

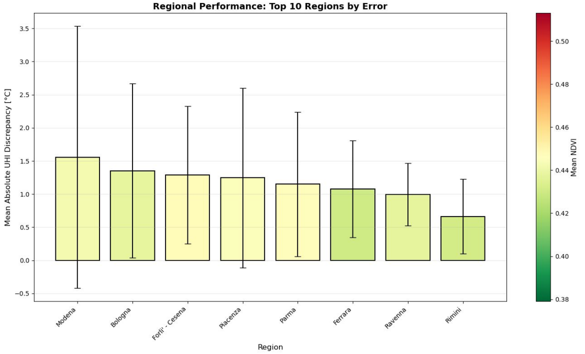
UHI Discrepancy	
Mean	1.20 °C
Standard Deviation	1.22 °C
Max	8.40 °C



Conclusions

While ERA5 remains useful for regional-scale climate analysis, it cannot replace high-resolution ground station networks in local urban heat monitoring.





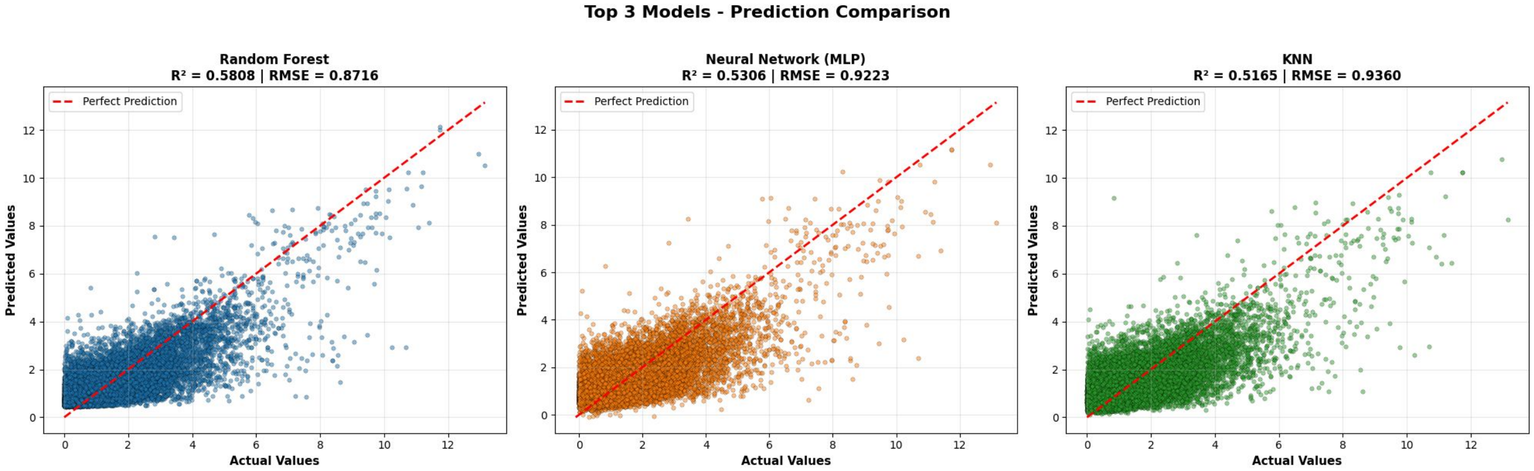
Week 4: Explanatory Modeling

We examined two methods to enhance the utility of ERA5 data:

1. Using machine-learning models to quantify the discrepancy between the ERA5-based UHI and observed UHI.
2. Replacing the observed UHI with an index that accounts for additional variables, including DEGURBA categories and ECA data.

Week 4: Explanatory Modeling

1. **Using machine-learning models to quantify the discrepancy between the ERA5-based UHI and observed UHI.**

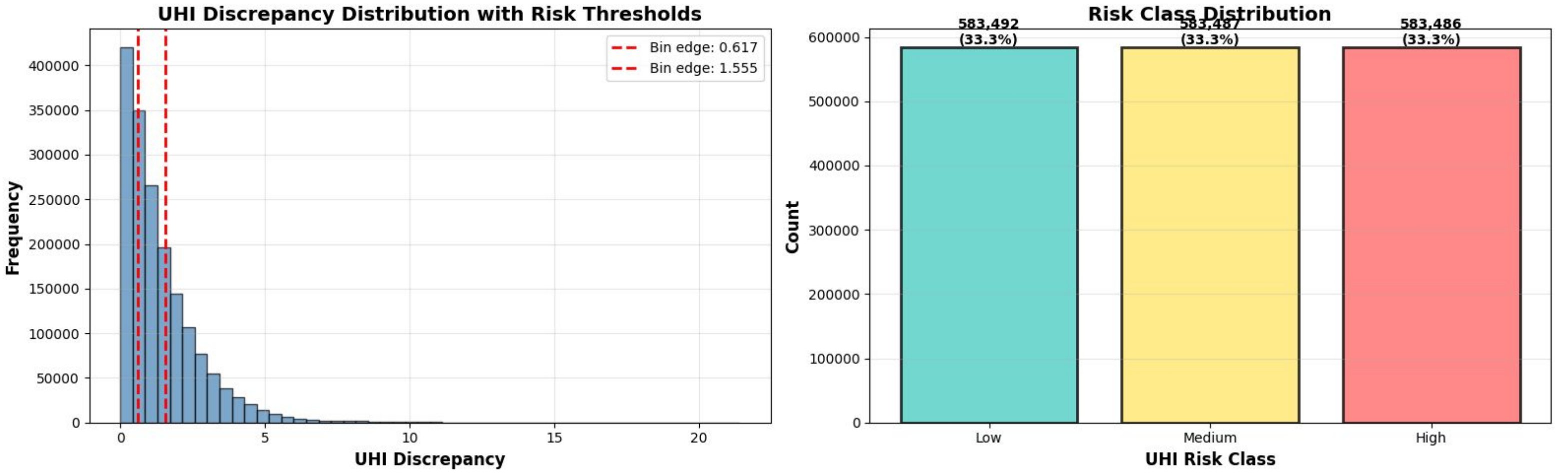


Unfortunately, the outcomes from this approach were not robust or useful enough to meet our needs.

Week 4: Explanatory Modeling

2. **Replacing the observed UHI with an index that accounts for additional variables, including DEGURBA categories and ECA data.**

To this end, we developed a model that estimates the risk of significant differences between the UHI indices, using three classes: low, medium, and high risk.



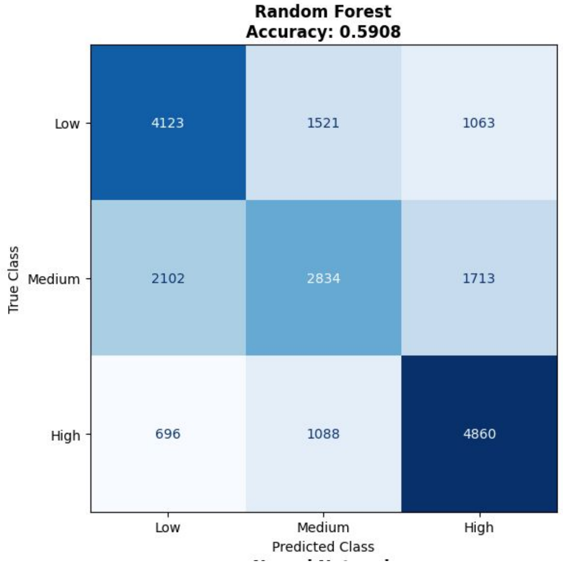
Week 4: Explanatory Modeling

2. **Replacing the observed UHI with an index that accounts for additional variables, including DEGURBA categories and ECA data.**

Our inputs included the station or regional elevation, the full set of ERA5 variables (t2m, u10, and others), the NDVI, the UHI computed from ERA5, and the DEGURBA classification.

The model outputs three risk categories, each indicating the likelihood of a substantial divergence between the UHI indices.

The model achieves good accuracy and can help identify where to prioritize correction efforts when working with ERA5 data.



Final Notes

1. Our analysis shows that **ERA5 underestimates urban temperatures**, especially in low-vegetation, highly urbanized zones.
2. Integrating **NDVI**, **DEGURBA**, and **station data** is crucial for a realistic assessment of UHI intensity.
3. Despite its limitations, ERA5 can still be used effectively when **paired with targeted correction strategies**.
4. The developed risk model helps identify where these corrections are most needed.
5. More granular observations and higher-resolution reanalysis data will be essential to advance urban climate monitoring.
6. **ECA station elevation** is one of the strongest predictors of UHI discrepancies, underscoring its importance in interpreting ERA5 biases.

So Long, and Thanks for All the Fish!

Made by Team ChefAI