Period 04

GENHACK CHALLENGE:
FINAL PRESENTATION

Project Goal
summarize your findings and propose explanatory models or adjustments to existing data.

Team Bat_TEAM:
Tito Nicola Drugman
Asia Montico
Beshoy Guirges

BAT_TEAM

THE PROBLEM

- **ERA5-Land Data**: Provides continuous coverage of Europe but at a coarse resolution (9km x 9km). It "averages" the temperature of forests, lakes and concrete into a single value.
- **Ground Stations (ECA)**: Provide precise "Ground Truth", but are sparse and unevenly distributed.
- **The Gap**: When a station is located in a dense city, ERA5 (which sees a mix of city and field) often underestimates the maximum temperature. This is the Urban Heat Island (UHI) bias.

BAT_TEAM

THE TASK

- We cannot install stations everywhere. We need a way to correct ERA5 data mathematically.
- **Goal:** Build a Machine Learning model that uses ERA, ECA and sentinel data to predict the Δ between the Model (ERA5) and Reality (Station).
- By applying this predicted Δ to the ERA5 baseline, we can effectively downscale climate data, generating hyper-local temperature estimates for any location, even where no sensors exist.

BAT_TEAM

Dataset

1. Covering Southern & Western Europe (IT, ES, FR, PT). 1,891 unique weather stations representing diverse micro-climates.

2. Rigorous temporal alignment of Ground Truth (ECA) and (ERA5). More than 3 million data points.

3. Used a Station-Based Split (not random shuffle). To ensure the model learns physical laws (urbanism effects) rather than memorizing specific station IDs.

BAT_TEAM

Features

Target Variable

- ***delta_temp***: The prediction target. We are predicting the bias, not the absolute temperature.

Vegetation & Land Cover (The "Green Contrast")

- ***ndvi_local***: Vegetation density at the specific station point.
- ***ndvi_global***: Average vegetation density of the surrounding 9km grid.
- ***delta_ndvi***: The difference (Local - Global).
- ***perc_urban***, ***perc_suburban***, ***perc_forest***: The fraction of land cover types within the grid cell.

Atmospheric Dynamics (ERA5)

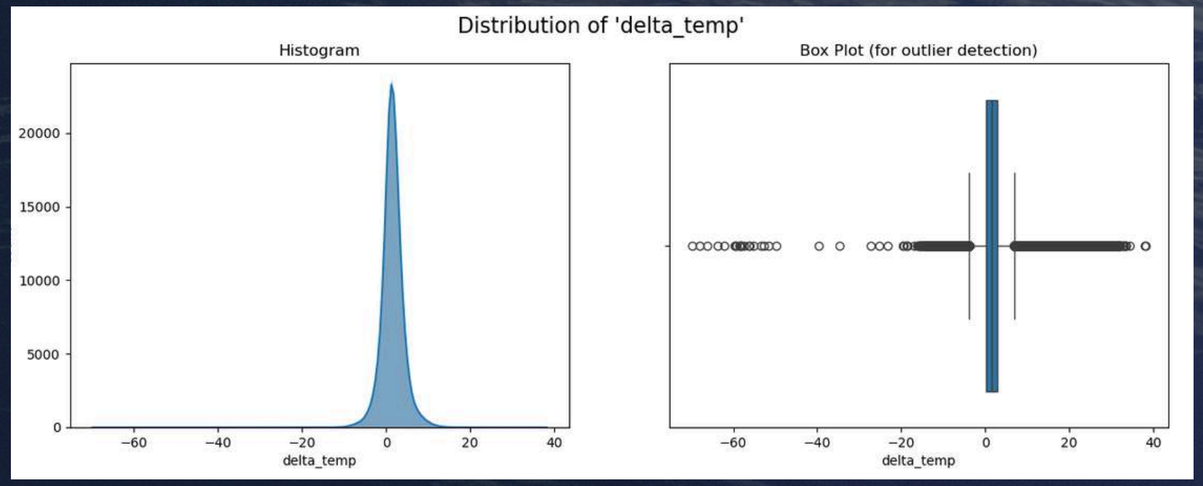
- ***wind_speed***, ***era5_u10*** and ***era5_v10***: magnitude and direction of the wind
- ***era5_precip***: Daily precipitation volume.
- ***rain_7day_avg***: 7-day rolling average of rain.

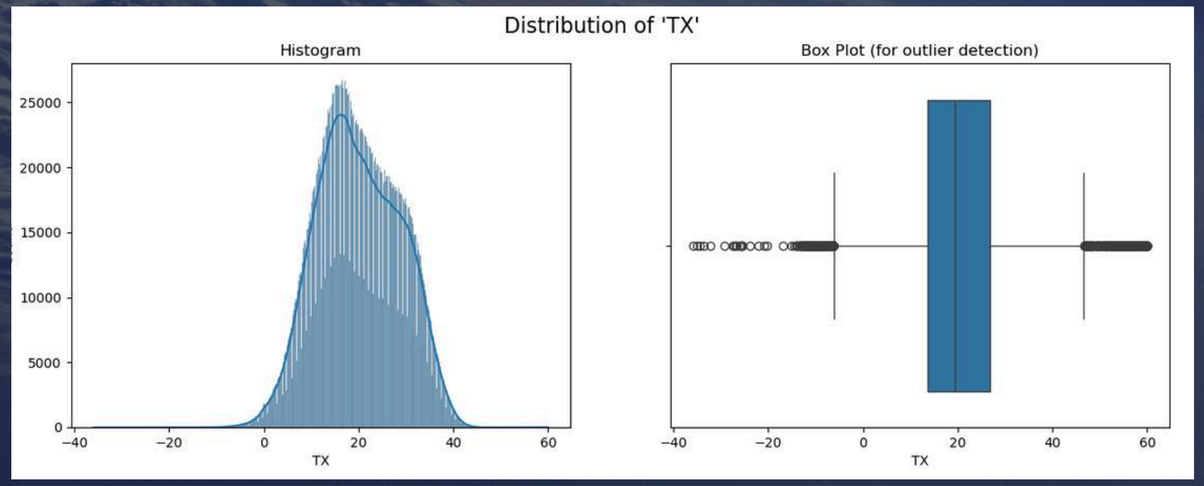
Spatio-Temporal Context

- ***Elevation***: Station altitude (meters).
- ***Latitude***, ***longitude***: Geographic coordinates.
- ***sin_day***, ***cos_day***: Cyclical time features to capture seasonal solar angles.

BAT_TEAM

Data Cleaning





Remove datapoints where |Δ|>20 OR daily TX value is outside [-45, 50].
Removed 588 rows.

BAT_TEAM

Some remarks

1. Elevation is a key driver of temperature, available for all our training stations. However, for full deployment on arbitrary 80m×80m pixels (where no stations exist), the model would require integrating an external Digital Elevation Model (DEM) to provide accurate altitude data.

2. Land-use features (%Urban, %Suburban, %Forest, %Water) were derived by aggregating Sentinel-2 NDVI data over the summer months. We applied specific biophysical thresholds to classify the underlying surface of the grid cells.

3. Temporal Continuity: We encoded the date using Sine and Cosine transformations. This ensures the model understands the cyclical nature of seasons, mathematically connecting December 31st to January 1st without a discontinuity.

BAT_TEAM

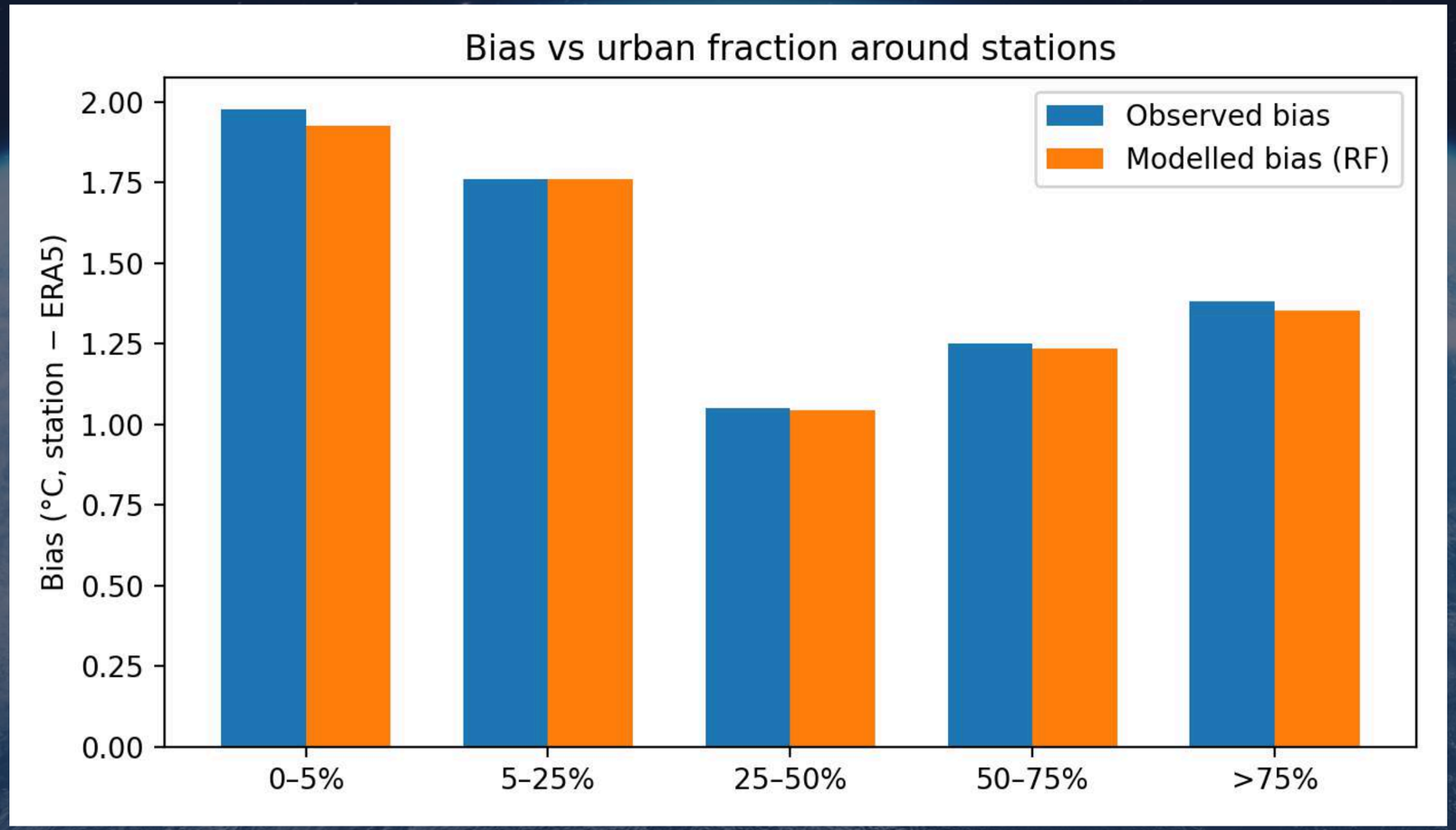
Model Selection & Optimization

1. **Random Forest Regressor** (Selected Model)
   - Why: Robust to outliers; captures non-linear interactions essential for UHI.
   - Performance: Best trade-off between RMSE (1.94C) and Explainability.
   - Key Hyperparameters:
     - n_estimators: 200
     - max_depth: 30
     - max_features: 0.5

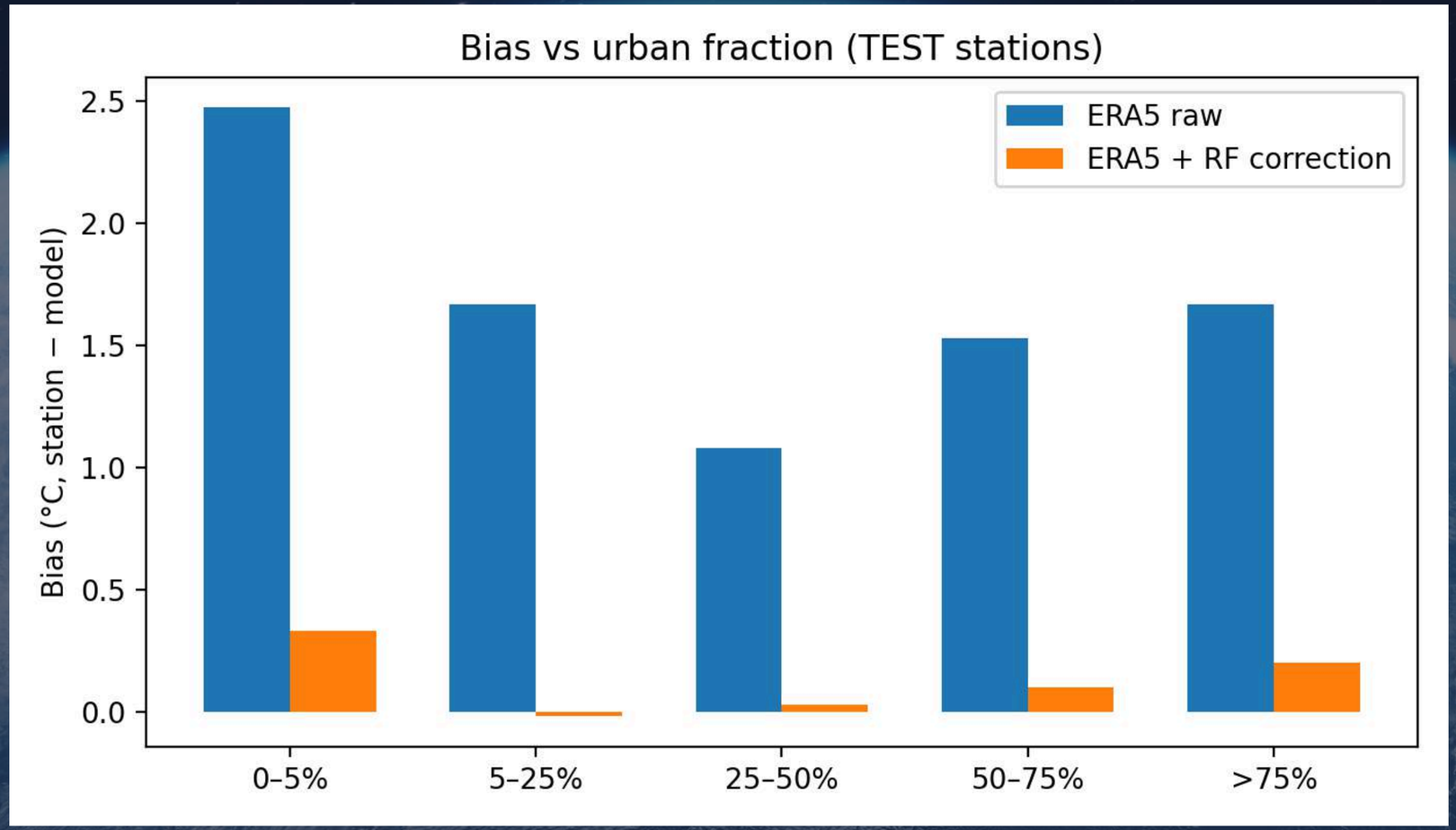
2. **HistGradientBoosting** (Challenger)
   - Native handling of large datasets (>3M rows); fast convergence.
   - Performance: Slightly higher error (2.06C) but faster training time.
   - Key Hyperparameters:
     - learning_rate: 0.05
     - max_iter: 600
     - l2_regularization: 0.5

RESULTS

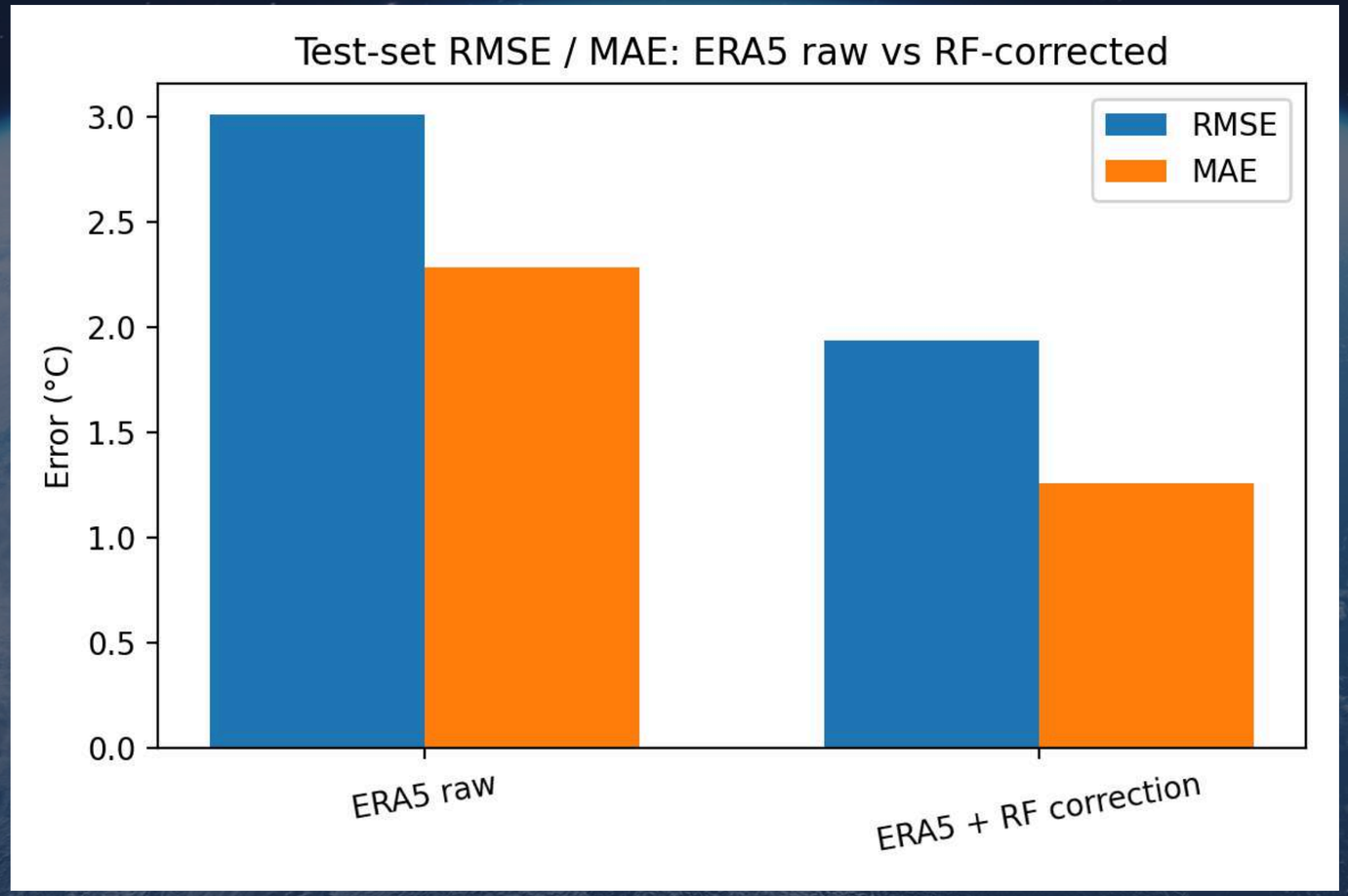
BAT_TEAM


Bias vs urban fraction around stations
Observed bias
Modelled bias (RF)
Bias (°C, station − ERA5)
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0–5%
5–25%
25–50%
50–75%
>75%

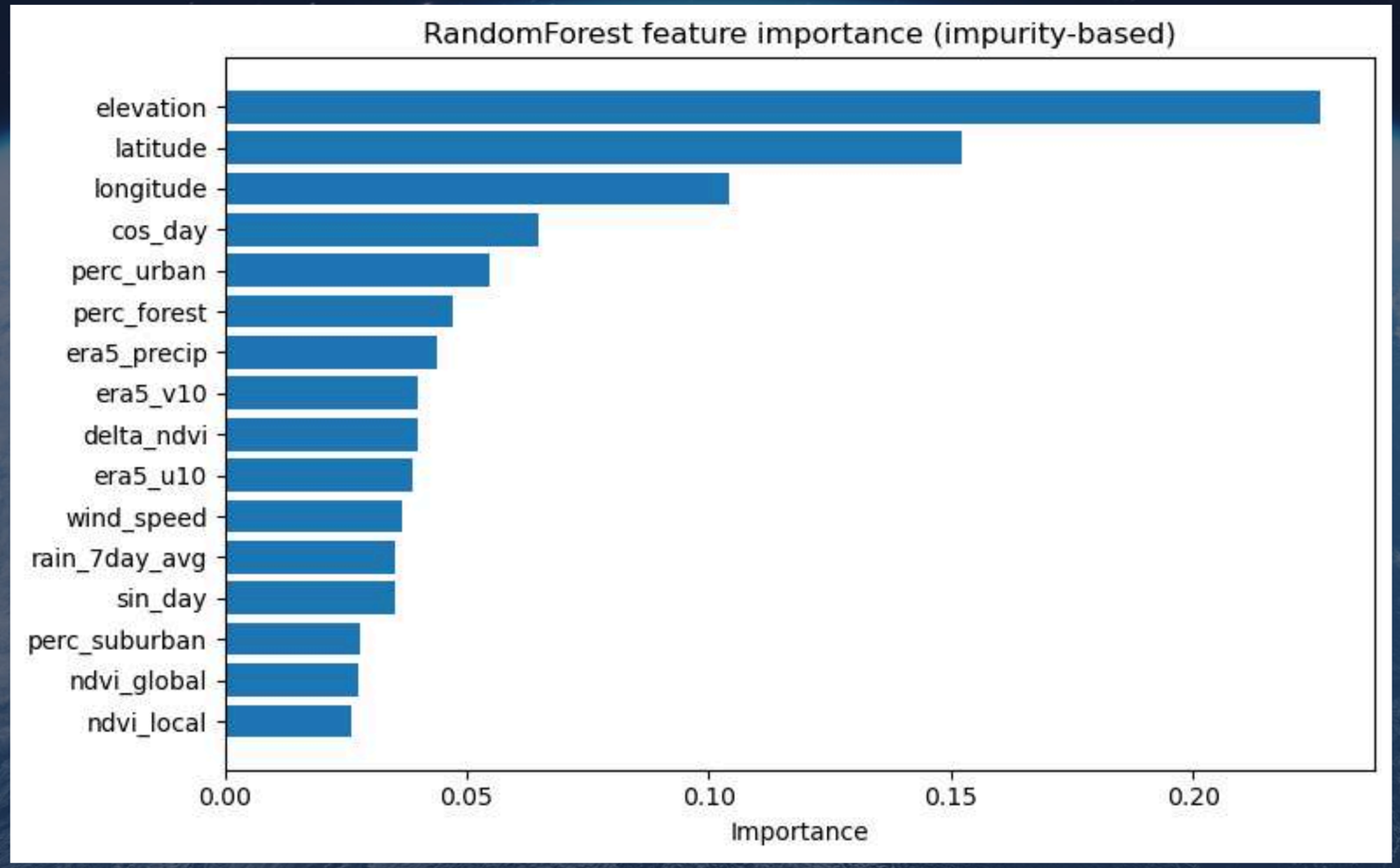
BAT_TEAM


Bias vs urban fraction (TEST stations)
ERA5 raw
ERA5 + RF correction
Bias (°C, station − model)
2.5
2.0
1.5
1.0
0.5
0.0
0–5%
5–25%
25–50%
50–75%
>75%

BAT_TEAM


Test-set RMSE / MAE: ERA5 raw vs RF-corrected
RMSE
MAE
Error (°C)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
ERA5 raw
ERA5 + RF correction

BAT_TEAM


RandomForest feature importance (impurity-based)
elevation
latitude
longitude
cos_day
perc_urban
perc_forest
era5_precip
era5_v10
delta_ndvi
era5_u10
wind_speed
rain_7day_avg
sin_day
perc_suburban
ndvi_global
ndvi_local
0.00
0.05
0.10
0.15
0.20
Importance

BAT_TEAM

TECHNICAL CONCLUSIONS

- Elevation is critical: Future deployment requires integrating external Digital Elevation Models (DEM) for locations without stations.
- Model is viable: Random Forest reduced RMSE by ~35% and successfully corrected the systematic ERA5 bias.
- UHI Signal: The Urban Heat Island effect is detectable but partially overpowered by strong topographical variances.

BAT_TEAM

TEAM FEEDBACK

- Experience: We truly enjoyed the challenge of bridging satellite imagery with climate science.
- Suggestion: Please extend the duration between Period 3 and 4 to allow for deeper final reporting.
- Thank You: Great organization. we look forward to the next GenHack edition!