ESC403.1 Introduction to Data Science

Solar Irradiance Prediction

Elia Torre, Martha Rösler,
Cyrill Albrecht, Mario Baumgartner

University of Zürich

Institute for Computational Science, UZH

University of Zurich[UZH]

Data set

https://power.larc.nasa.gov/data-access-viewer/

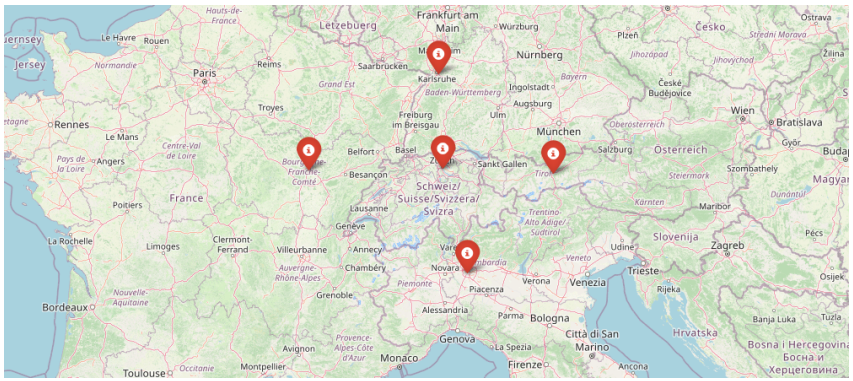
Zurich [Latitude: 47.35, Longitude: 8.55]
hourly measurements from from 01.01.2013 until 31.12.2022

- The radiance $[W/m^2]$
- Temperature $[C^\circ]$
- Humidity $[g/kg]$
- Precipitation $[mm/h]$
- Wind speed $[m/s]$
- Wind direction $[^\circ]$.
- Frost point $[C^\circ]$
- Wet bulb temperature $[C^\circ]$
- Surface Pressure $[kPa]$

Additional data sets

Same data set for different locations

- Milan [Latitude: 45.47, Longitude: 9.18]
- Dijon [Latitude: 47.32, Longitude: 5.06]
- Innsbruck [Latitude: 47.27, Longitude: 11.43]
- Karlsruhe [Latitude: 48.99, Longitude: 8.43]

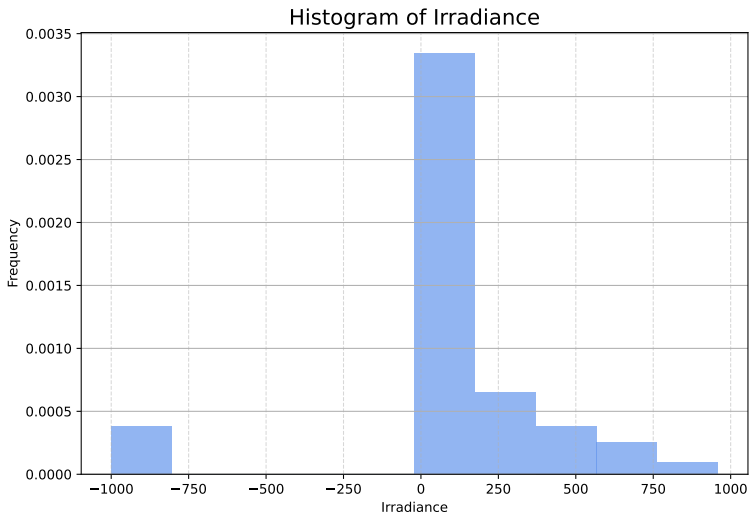


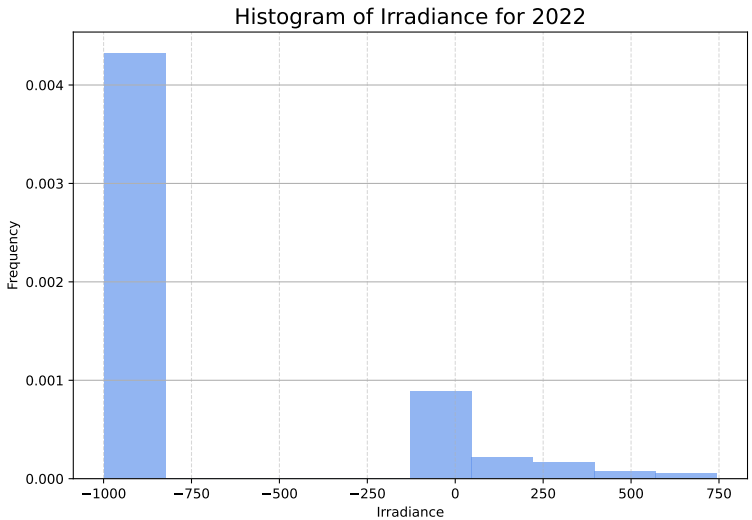
Research questions

- What relations exist in our data set?
- Can we develop an accurate machine learning model to predict solar radiation using our data?
- Is it possible to estimate solar radiation in Zurich by averaging over four neighboring locations?

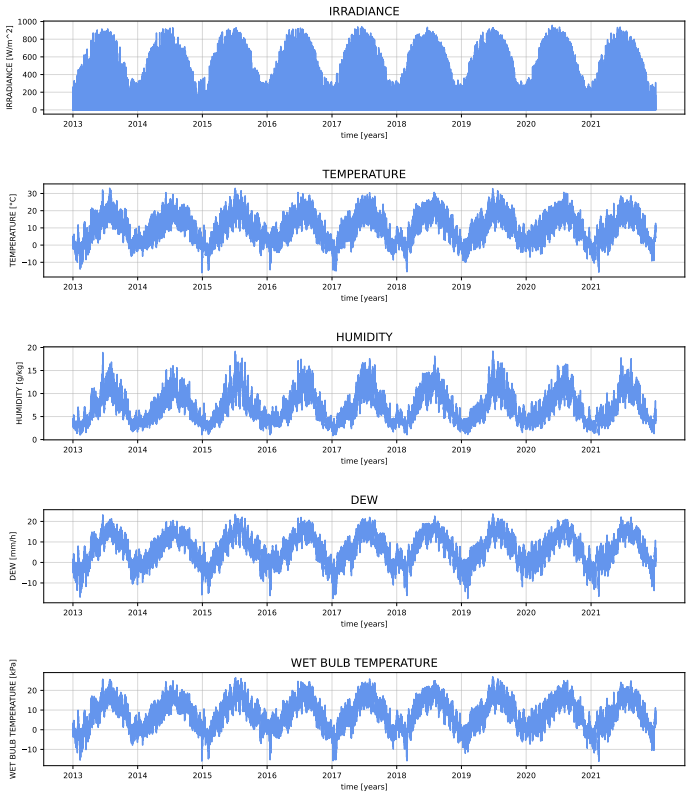
Data Cleaning & Augmentation

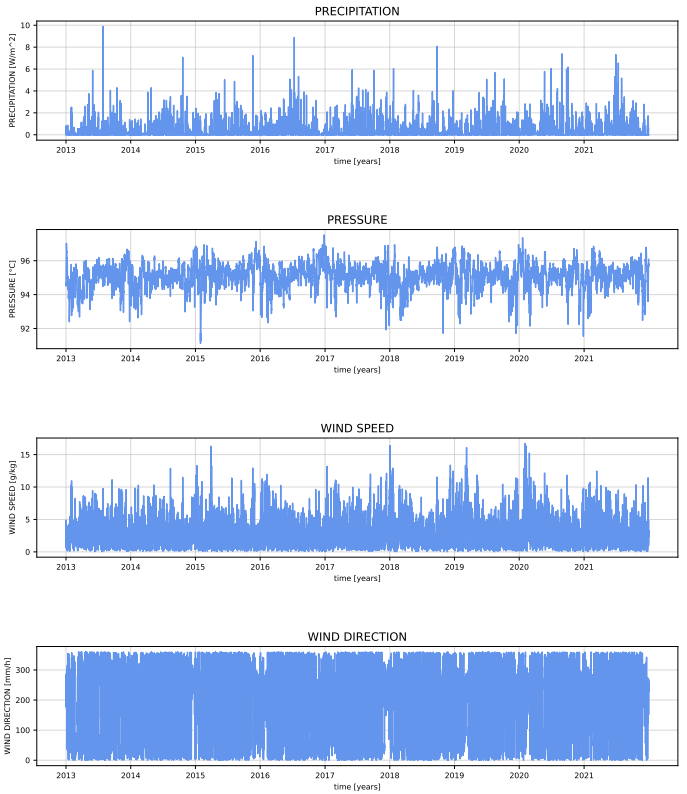
- Limit data set to items till 31. December 2021
- Add sunrise and sunset time
- Make a binary variable light
- Check for NaN values



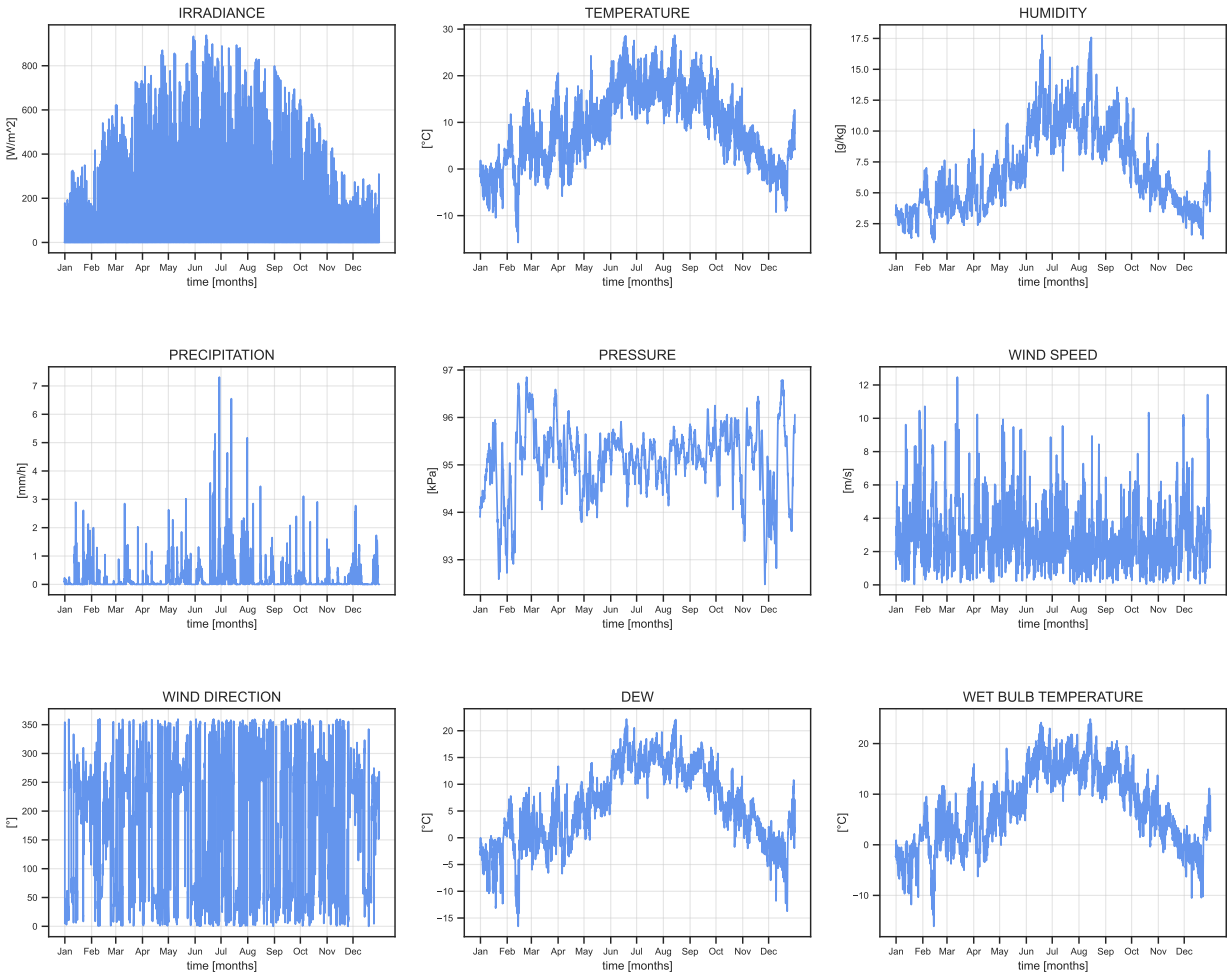


Change over time for all features

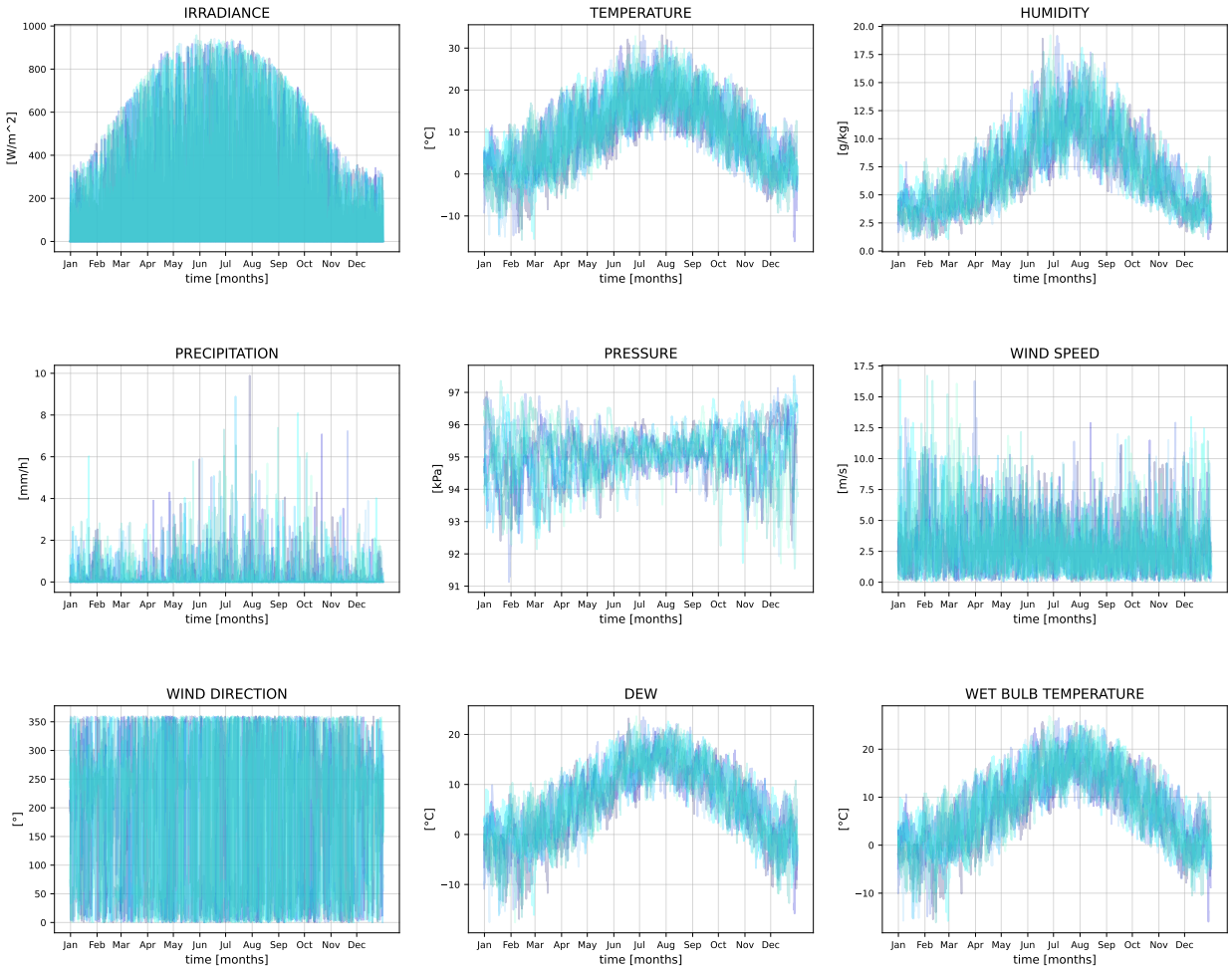




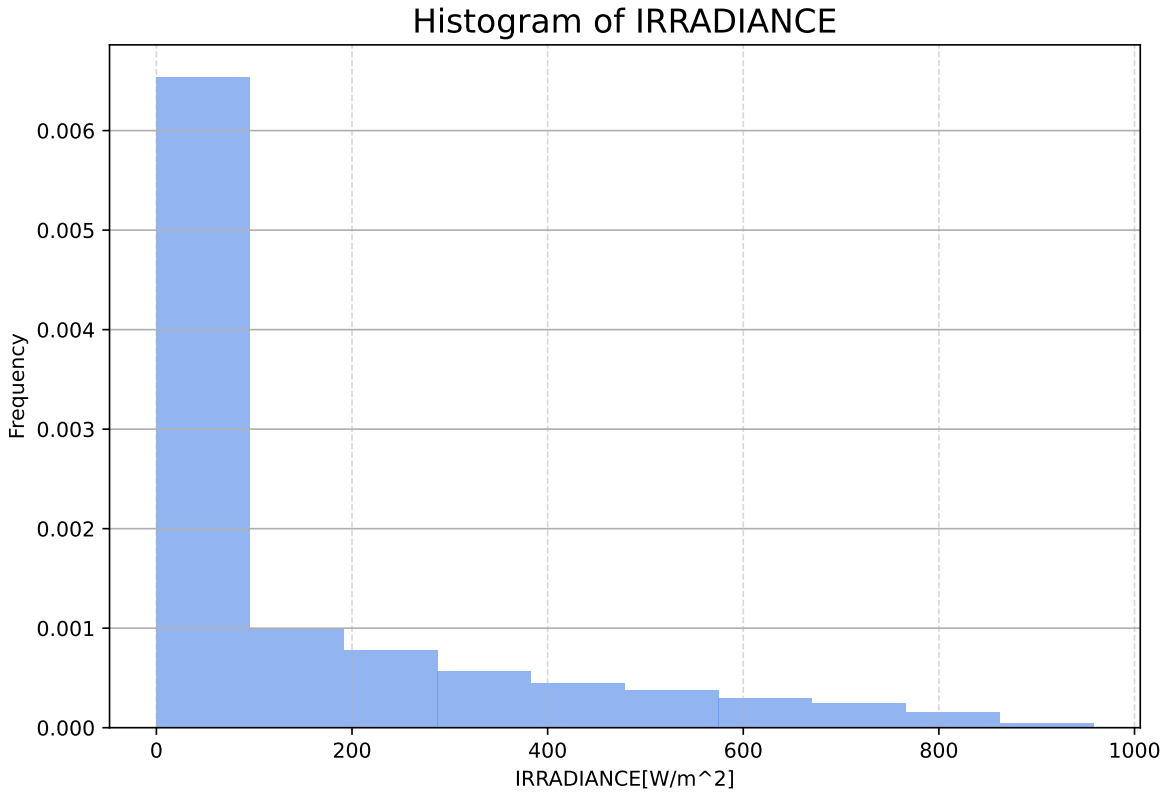
Change over 2021 for all features



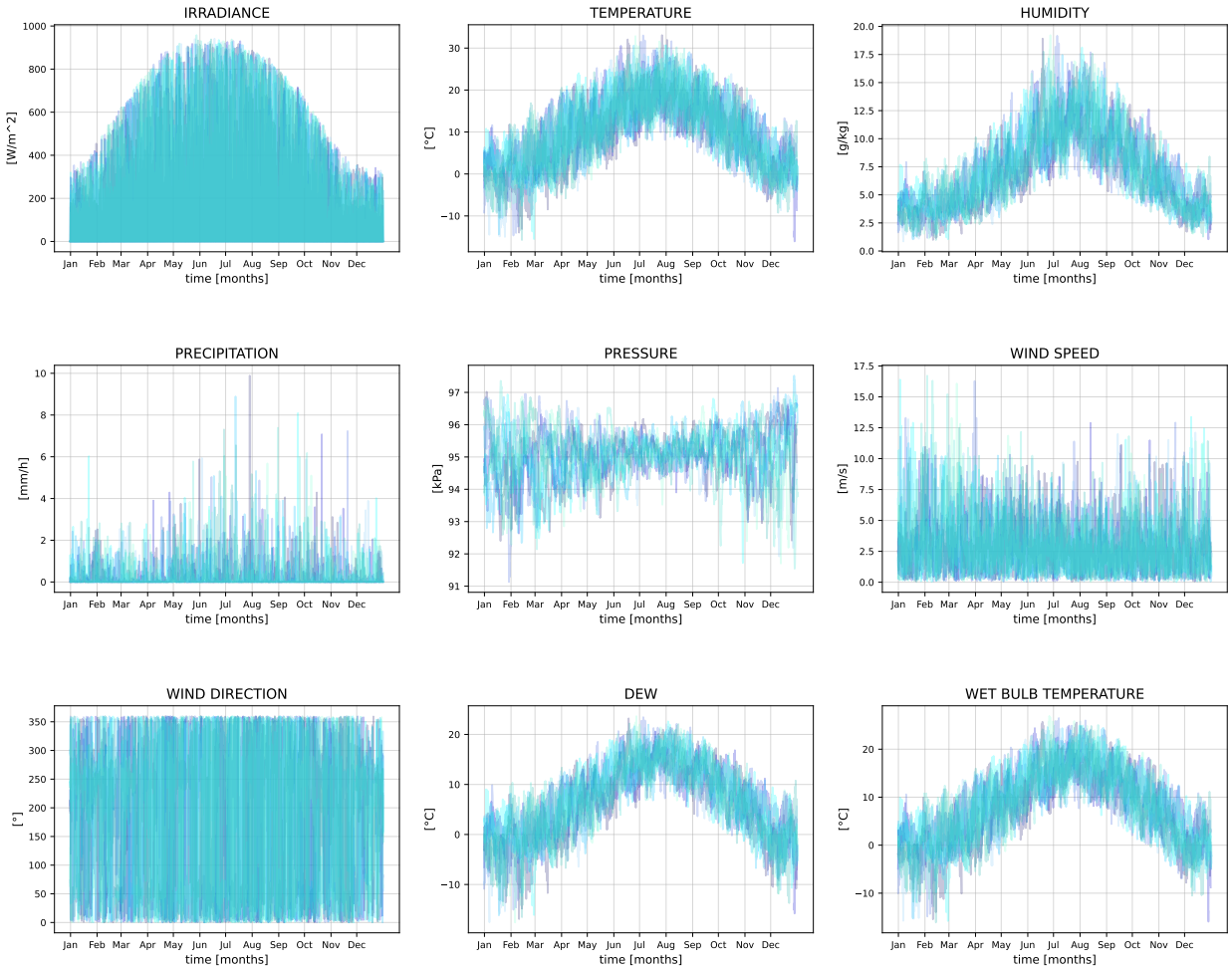
Change over the year, 2013-2021 overlaid



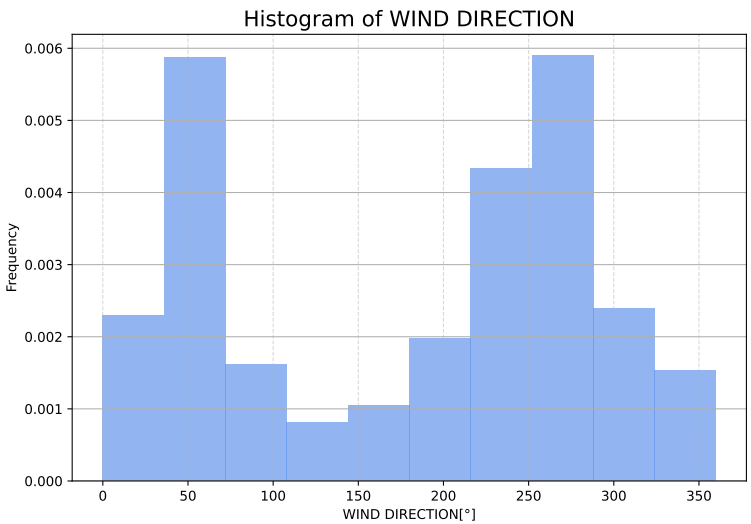
Closer look at irradiance

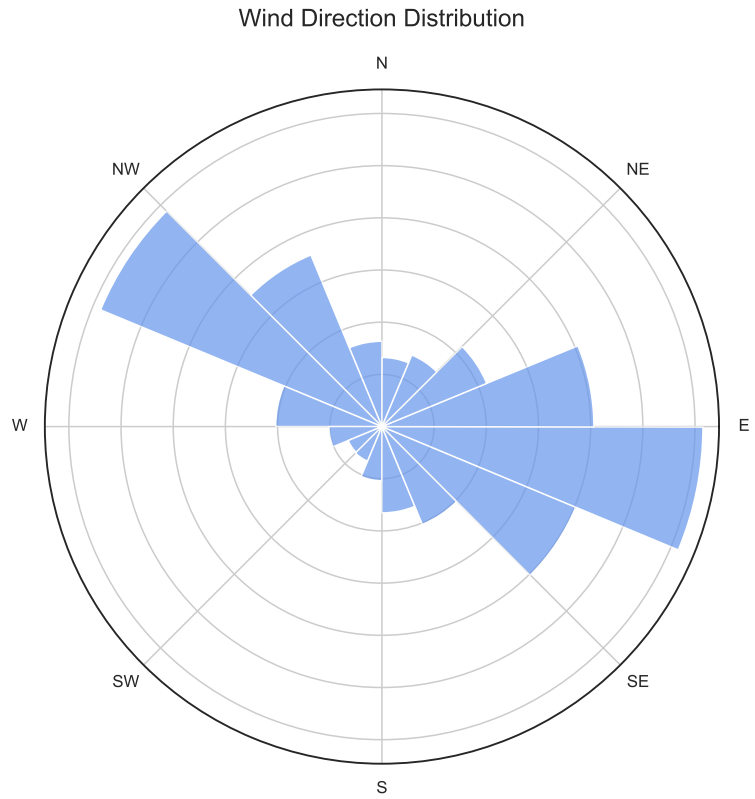


Change over the year, 2013-2021 overlaid

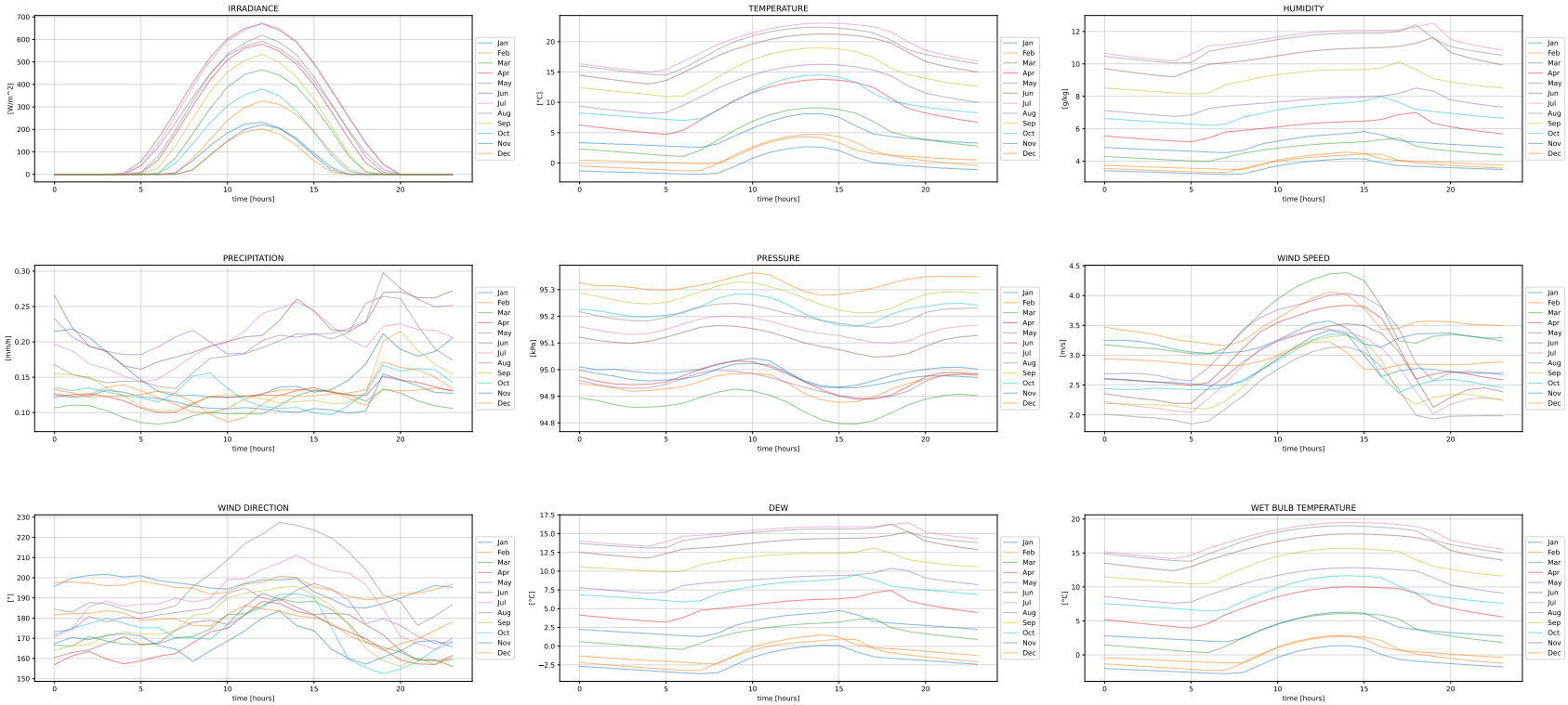


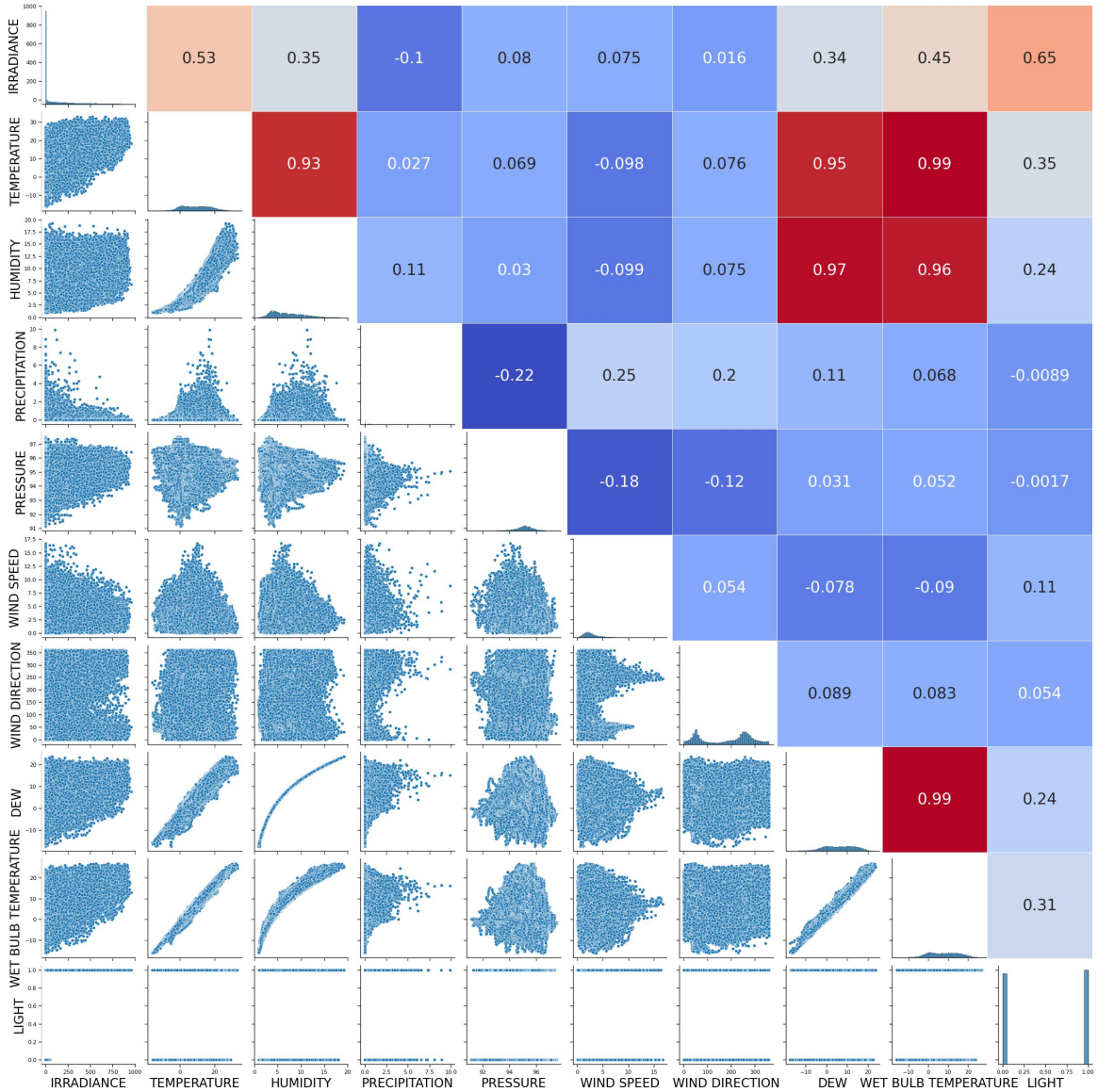
Closer look at the wind direction





Mean over the months




IRRADIANCE
TEMPERATURE
HUMIDITY
PRECIPITATION
PRESSURE
WIND SPEED
WIND DIRECTION
DEW
WET BULB TEMPERATURE
LIGHT
0.53 0.35 -0.1 0.08 0.075 0.016 0.34 0.45 0.65
0.93 0.027 0.069 -0.098 0.076 0.95 0.99 0.35
0.11 0.03 -0.099 0.075 0.97 0.96 0.24
-0.22 0.25 0.2 0.11 0.068 -0.0089
-0.18 -0.12 0.031 0.052 -0.0017
0.054 -0.078 -0.09 0.11
0.089 0.083 0.054
0.99 0.24
0.31

Prediction Pipeline

1. Dataset splitting with 5-Kfold Cross-Validation
2. Initial Models Evaluation
3. Tuning of Models
4. Training of Tuned Models
5. Ensemble of Best Performing Models
6. Neural Networks Architectures Training & Hyperparameters Optimization
7. Tuned Models Evaluation

Models

- Linear Regression
- K-Nearest Neighbor Regression (KNN)
- Decision Tree Regression
- Random Forest Regression (RF)
- Extreme Gradient Boosting (XGB)
- AdaBoost (ADB)
- CatBoost
- LightGBM (LGBM)
- Voting Regressor (rf, xgb, adb)
- Stacking Regressor (Ridge) (rf, xgb, adb)
- Stacking Regressor (ElasticNet) (rf, xgb, adb)
- Feed-Forward Neural Network (FFNN)
- Long Short-Term Memory Neural Network (LSTM)

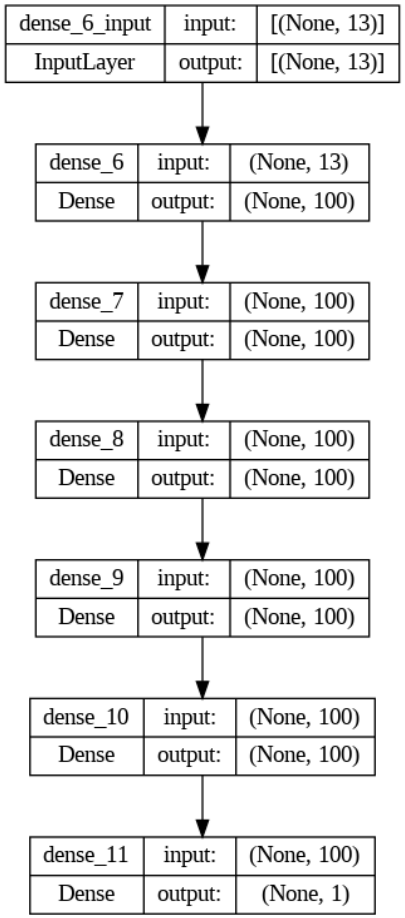
Hyperparameter Tuning

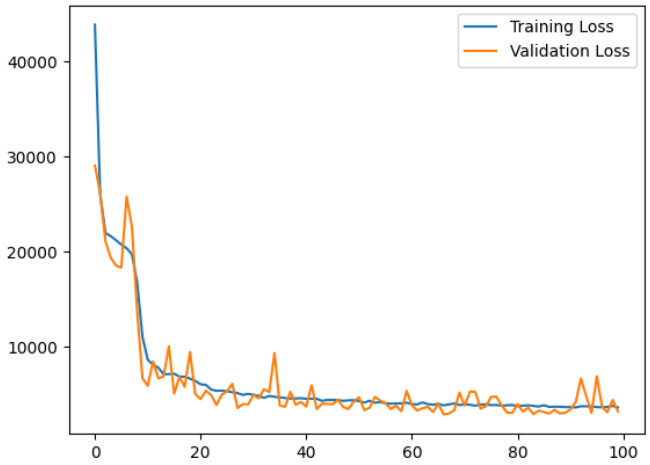
- RandomizedSearchCV (5-Folds)
- 20 Iterations (approx. 3h per model)
- Negative Mean Squared Error

Model	**Parameters**	**Model**	**Parameters**
XGBRegressor	subsample: 1.0 n_estimators: 362 min_samples_split: 0.001 min_samples_leaf: 1 max_depth: 7 learning_rate: 0.06	**Decision Tree Tuning**	splitter: best min_weight_fraction_leaf: 0.2 min_samples_leaf: 7 max_leaf_nodes: 50 max_depth: 11
Random Forest Tuning	n_estimators: 50 max_features: 5 max_depth: 20	**AdaBoost Tuning**	n_estimators: 250 learning_rate: 1.0 base_estimator: DecisionTree

Feed-Forward Neural Network

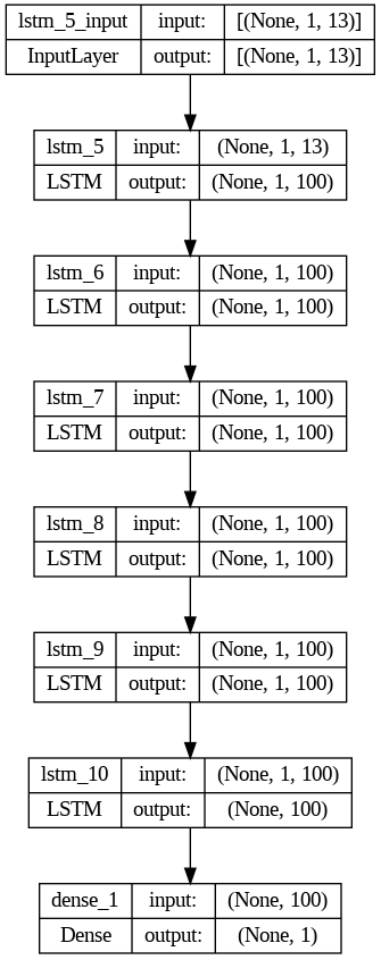
- Min-Max Scaling
- 6 FC Layers
- ELU Activation
- ADAM Optimizer
- 32 Batch Size
- 100 Epochs
- Colab GPU (40 mins)

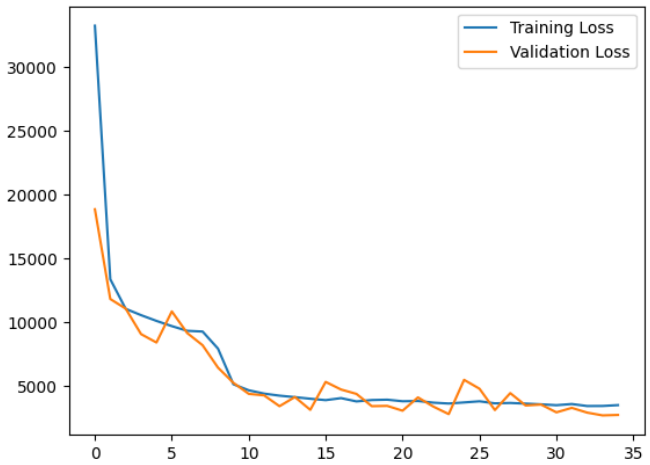




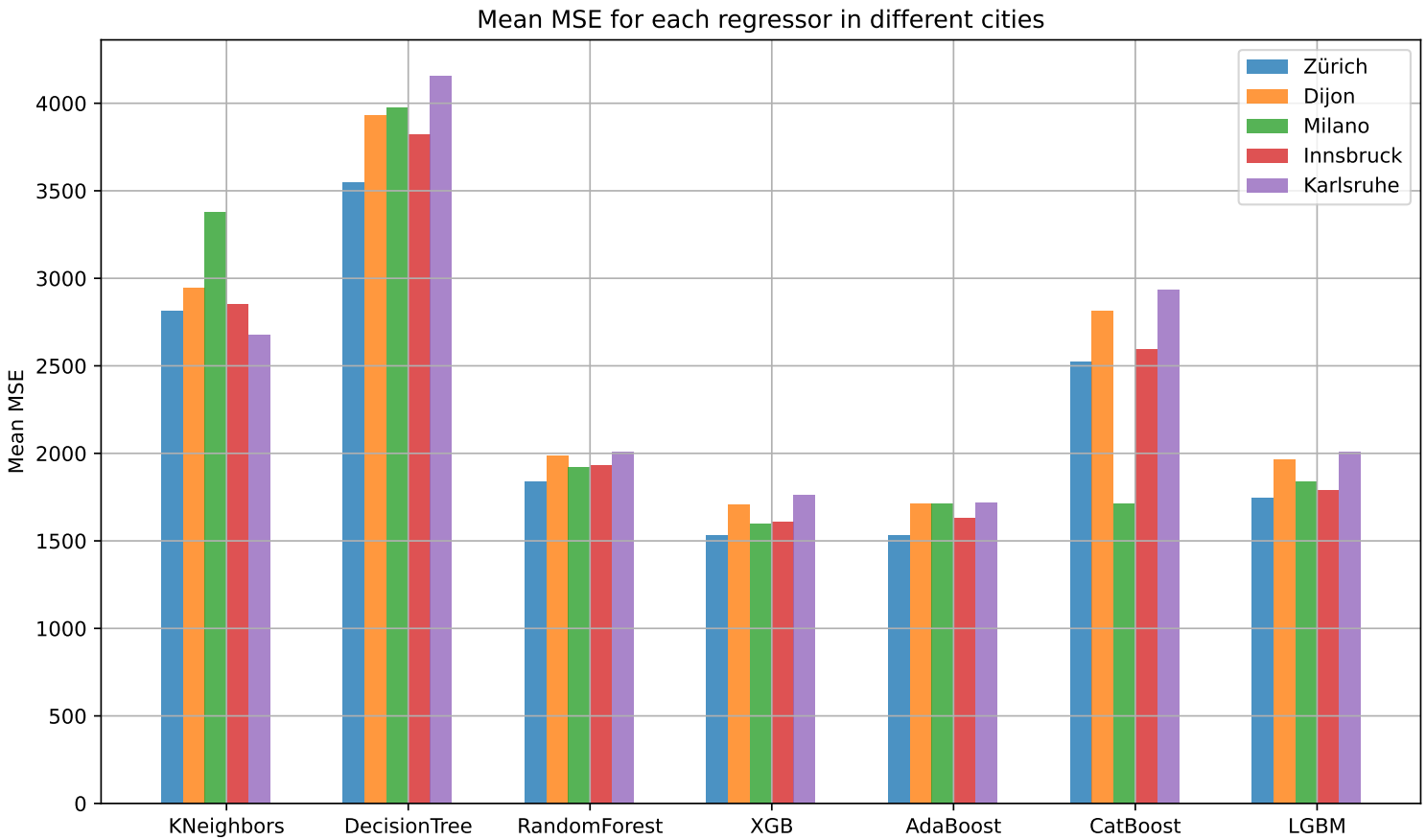
Long Short-Term Memory Neural Network (LSTM)

- 6 LSTM Layers
- 1 Dense Layer
- ReLU Activation
- ADAM Optimizer
- 32 Batch Size
- 40 Epochs
- Colab GPU (30 mins)





Results of the different models

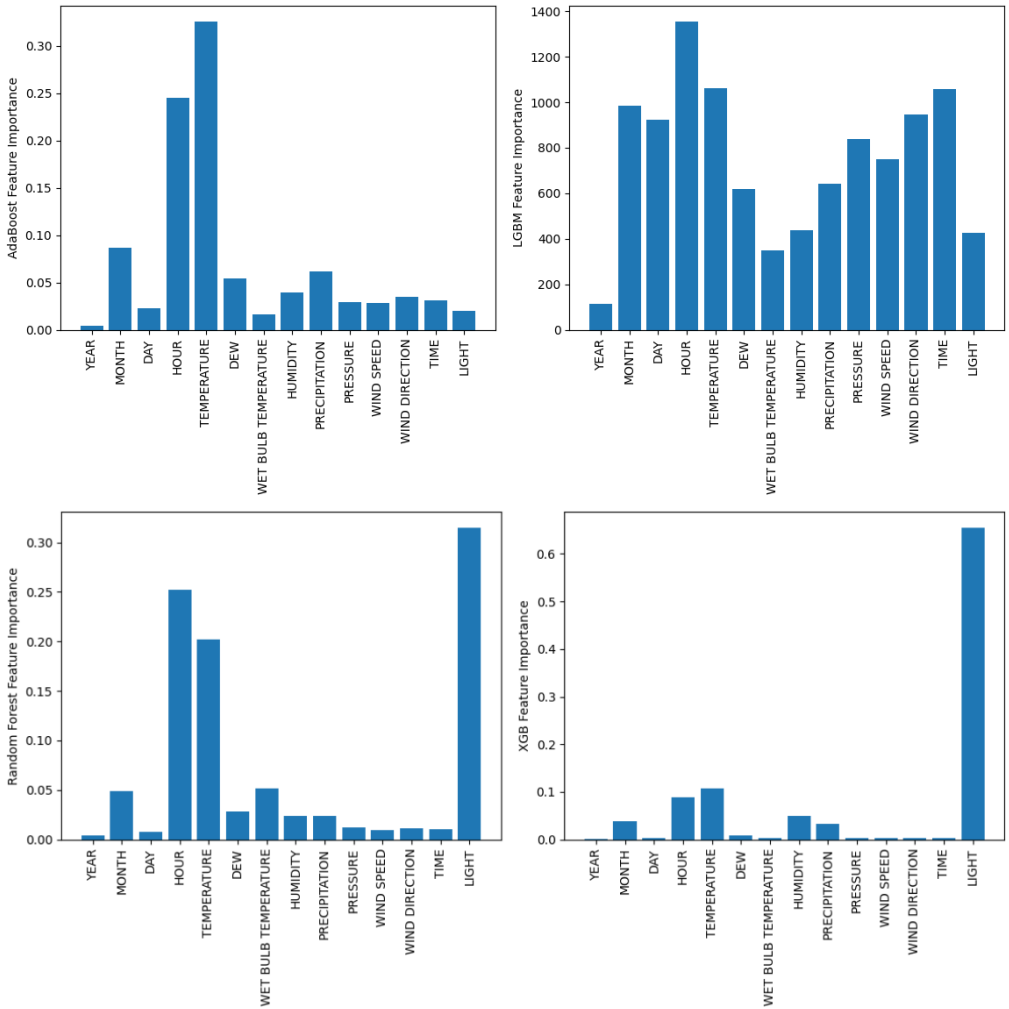


Results of the different models

Regressor	Zürich	Dijon	Milano	Innsbruck	Karlsruhe
Linear	14924.7	14536.2	17394.6	11153.7	13373.8
KNeighbors	2813.9	2942.3	3378.6	2850.6	2673.9
DecisionTree	3548.4	3932.3	3971.7	3821.7	4154.4
RandomForest	1839.9	1983.2	1920.0	1932.8	2009.1
XGB	1531.1	1707.4	1595.0	1609.7	1761.4
AdaBoost	1530.5	1710.7	1714.6	1629.9	1718.5
CatBoost	2522.5	2811.4	1711.8	2593.7	2935.1
LGBM	1743.1	1964.2	1840.5	1789.5	2008.1
Voting	1556.8	/	/	/	/
Stacking (Ridge)	1554.8	/	/	/	/
Stacking (ElasticNet)	1548.6	/	/	/	/
FFNN	3204.3	/	/	/	/
LSTM	2872.8	/	/	/	/

Table: Mean MSE for each regressor in different cities

Feature Importances



"Quadrangulation" of Zurich Location

- Predict `IRRADIANCE` in Zürich
- Ensemble of 4 Locations Data

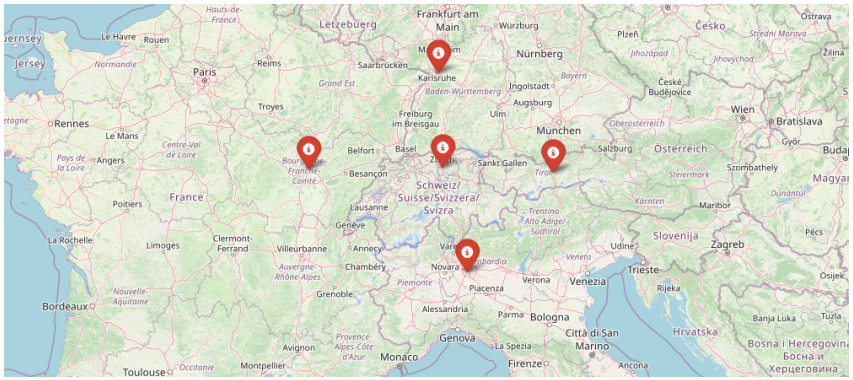


Figure: Plot with Folium library of the 5 locations

Averaging Measured Solar Radiation of the 4 Cities

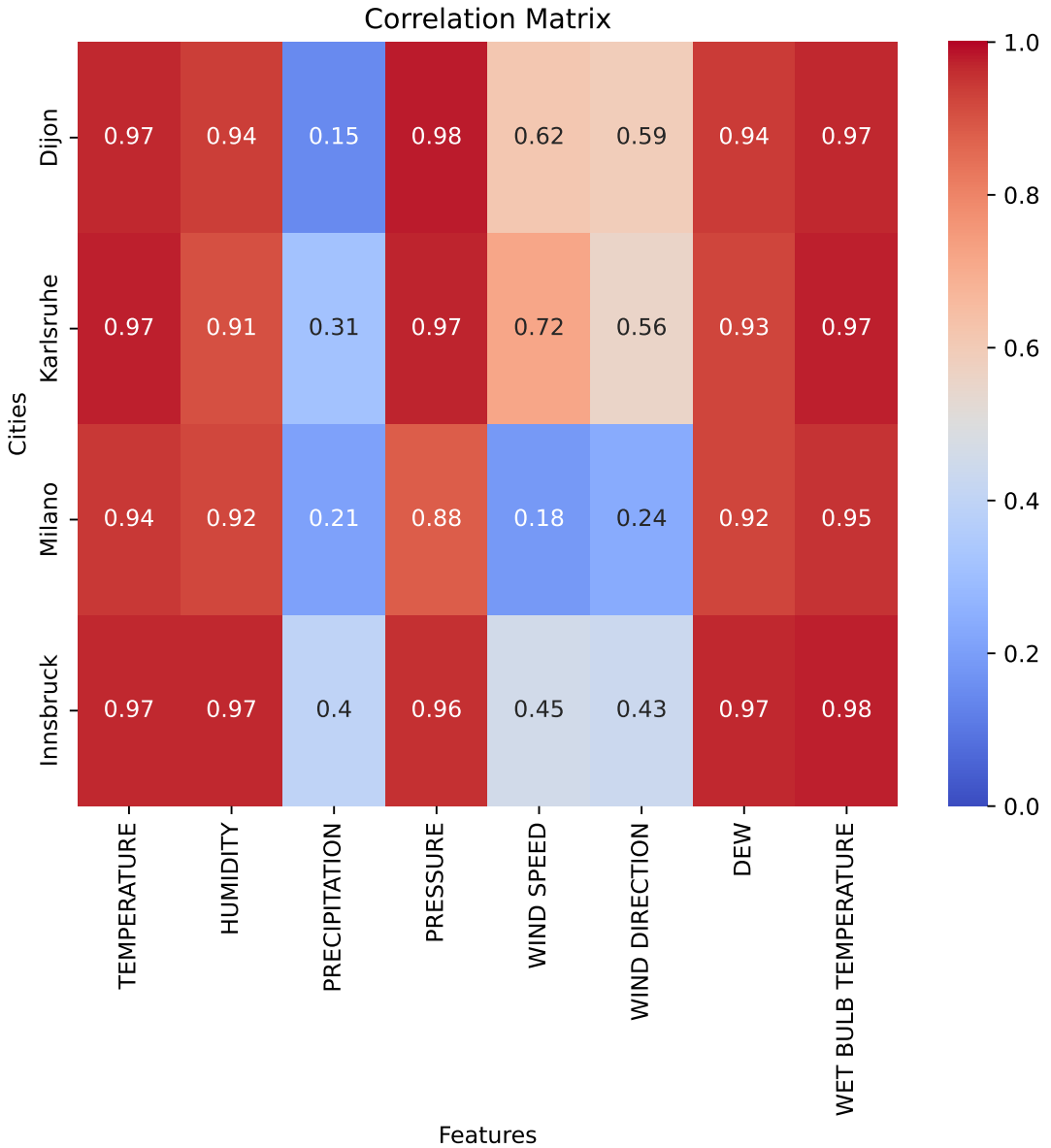
$$\texttt{MSE}\left(\frac{\text{Dijon} + \text{Milan} + \text{Karlsruhe} + \text{Innsbruck}}{4}\right) = 1979.2$$

Weighted Averaging Measured Solar Radiation of the 4 Cities

$$\texttt{MSE}\left(\frac{\alpha * \text{Dijon} + \beta * \text{Milan} + \delta * \text{Karlsruhe} + \gamma * \text{Innsbruck}}{4}\right) = 1888.75$$

Where the coefficients (i.e., $\alpha, \beta, \delta, \gamma$) are defined as the normalized mean correlations over the data sets features between each city and Zürich.

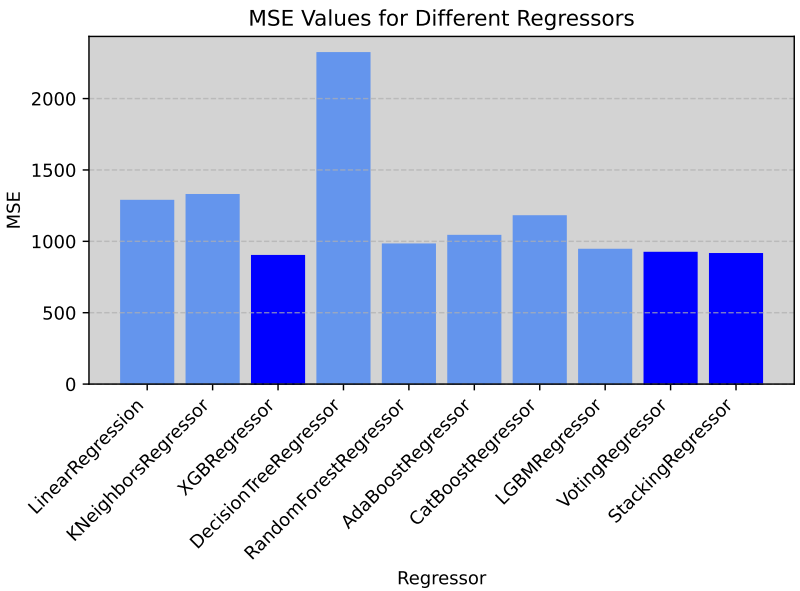
City	Corr
Dijon	0.768
Karlsruhe	0.792
Milano	0.656
Innsbruck	0.764



Custom Data Set

- We create a custom data set by concatenating the features from the 4 cities.
- Features: [TEMPERATURE], [HUMIDITY], [PRECIPITATION], [PRESSURE], [WIND SPEED], [WIND DIRECTION], [DEW], [WET BULB TEMPERATURE], [LIGHT], [IRRADIANCE].
- Data set shape: 78888 entries x 45 features.
- 3-Fold Cross Validation with the same models used before (no retuning).

Results Custom Data Set



Regressor	MSE
LinearRegression	1284.6
KNeighborsRegressor	1325.3
XGBRegressor	898.4
DecisionTreeRegressor	2319.4
RandomForestRegressor	978.9
AdaBoostRegressor	1039.2
CatBoostRegressor	1176.6
LGBMRegressor	941.7
VotingRegressor	920.5
StackingRegressor	911.8

Table: Results for all Regressors with a custom data set

Research questions about the dataset

- What are the relations in our data set?

$\Rightarrow$ Data exploration and machine learning showed various

- Are there non-trivial relations between the different variables?

$\Rightarrow$ Notable behaviour of variables

Research questions about the prediction pipeline

- Can we find an accurate machine learning model to predict solar radiation?

$\Rightarrow$ Yes e.g. XGB, LGBM, Random Forest and Ada Boost

  - Which variables have the biggest impact and which ones are negligible?

  $\Rightarrow$ That depends heavily on the regressor

  - Which model best explains the relationships among variables? Why?

  $\Rightarrow$ Random Forest and Ada Boost

  - Does the prediction reach a level of accuracy such that it could be exploited for application purposes?

  $\Rightarrow$ Yes with an root mean squared error of $\pm 40\ W/m^2$

Research questions about the "Quadrangulation" of Zurich

- Is it possible to estimate this measure by averaging over 4 other neighboring locations?
  - Averaging the measured solar radiation in the 4 cities?

  $\Rightarrow$ Does work surprisingly good and is fairly simple
  - Assembling the measured solar radiation in the 4 cities with different weights

  $\Rightarrow$ This does improve the prediction by a bit
  - Training an model on a custom data set featuring information of all of the 4 cities?

  $\Rightarrow$ Achieved a remarkable small root mean square error of only $\pm 30\ W/m^2$

Questions?