# Predicting Solar Radiation Flux With Simple Input Data

*By Dmytro Kuskov*

## Introduction and problem statement

An accurate assessment and forecast of solar radiation on the Earth's surface plays a key role for the transition to sustainable energy.

This indicator directly determines the efficiency of photovoltaic installations, solar heating systems and a number of related processes from planning power grids to calculating the potential of agro-industrial projects.

This project considers the problem of forecasting solar radiation flux based on simple and widely available data. The direct value of solar radiation is not a typical forecast for the public, although it determines the efficiency of solar panels and many climate calculations.

It can be obtained from professional weather forecasts, which are not as easy to obtain as data from public forecasts. The goal of the project is to build a model that allows obtaining radiation flux values from a minimal set of input data: cloud cover, location coordinates and day of the year. All other derived features like astronomical length of the day are reproduced by the code independently, and the model learns from historical data. The project aims to make solar radiation forecasting as accessible and reproducible as possible: using simple input data, it is possible to reconstruct solar flux values and use them in sustainable development tasks.

The data source used in the project for modeling is AgERA5, a global climate dataset providing values of radiation flux and cloudiness. Additionally, elevation added from the GMTED2010 dataset and derived features were calculated that reflect geographical and astronomical factors: daylight hours, sunlight hours, extraterrestrial solar radiation.

Project goal: develop a reproducible and interpretable regression pipeline for solar radiation forecasting. Compare linear and non-linear models (Linear Regression, Ridge, Random Forest, etc.) by key metrics (MAE, RMSE, $R^2$). Assess how the obtained forecasts can be used by stakeholders to make decisions in the field of energy and sustainable development.

The practical value of the results is that better local solar flux forecasts allow:

- optimizing the placement and operation of photovoltaic systems,

- reducing uncertainty in generation planning,
- increasing the reliability of energy systems based on renewable energy sources,
- using climate data for agricultural decisions related to solar radiation (e.g. irrigation planning or crop yield assessments).

## Data and preprocessing

### Data sources

- AgERA5: historical climate data for 2024 from which the following were taken:
  - Solar radiation flux (J/m²) - target variable [1]. The values were converted to megajoules (MJ/m²).
  - Cloud cover - originally in percent, normalized to the range 0-1.
- GMTED2010 - global elevation, used as an additional geophysical predictor [2].
- Additional characteristics were calculated for modeling:
  - Daylength using FAO-56 equations [4].
  - Sunlight hours using standard astronomical formulas used by the National Oceanic and Atmospheric Administration (NOAA) and the United States Navy [3]. Both options (daylength and sunlight hours) were kept to empirically compare their impact on the model.
  - $H_0$ (extraterrestrial radiation) calculated using FAO-56 equations [4].
- Calendar features: Day of the year (1-366). Sine/cosine transformations to reflect cyclicity.
- Geographical features: Coordinates (latitude, longitude).

Feature approach for different models:

For linear models (Linear Regression, Ridge, Lasso), additional feature transformations were applied: cos (latitude), cos (longitude), log (elevation).
For nonlinear models (Random Forest, Gradient Boosting, etc.), these features were used untransformed to preserve the flexibility of trees when working with multi-scale data.

Quality control and reproducibility:
The dataset checked for gaps and outliers. Missing values were removed.

## Exploratory data analysis (EDA)

### General patterns

Seasonality: solar radiation flux shows a pronounced annual cycle, with maximums in the summer months and minimums in winter. This is expectable since 75% of the points in the dataset are in the Northern hemisphere.

Coordinates (latitude) determine the basic level of seasonality: the closer to the equator, the smaller the amplitude of radiation fluctuations.

There is a strong negative correlation between cloud cover and solar radiation flux. At high cloudiness values (close to 1), the radiation flux decreases significantly, while the spread of values due to other factors (geography, seasonality) remains.

Preliminary analysis showed a weak positive relationship of higher elevation, and radiation values which increase slightly with higher altitude. The relationship is not as strong as cloudiness or seasonality, but can be useful in combination with other features.

### Comparison of astronomical features

Daylength, sunlight hours, and extraterrestrial radiation show similar seasonal patterns, but differ in amplitude and shape of the curve.

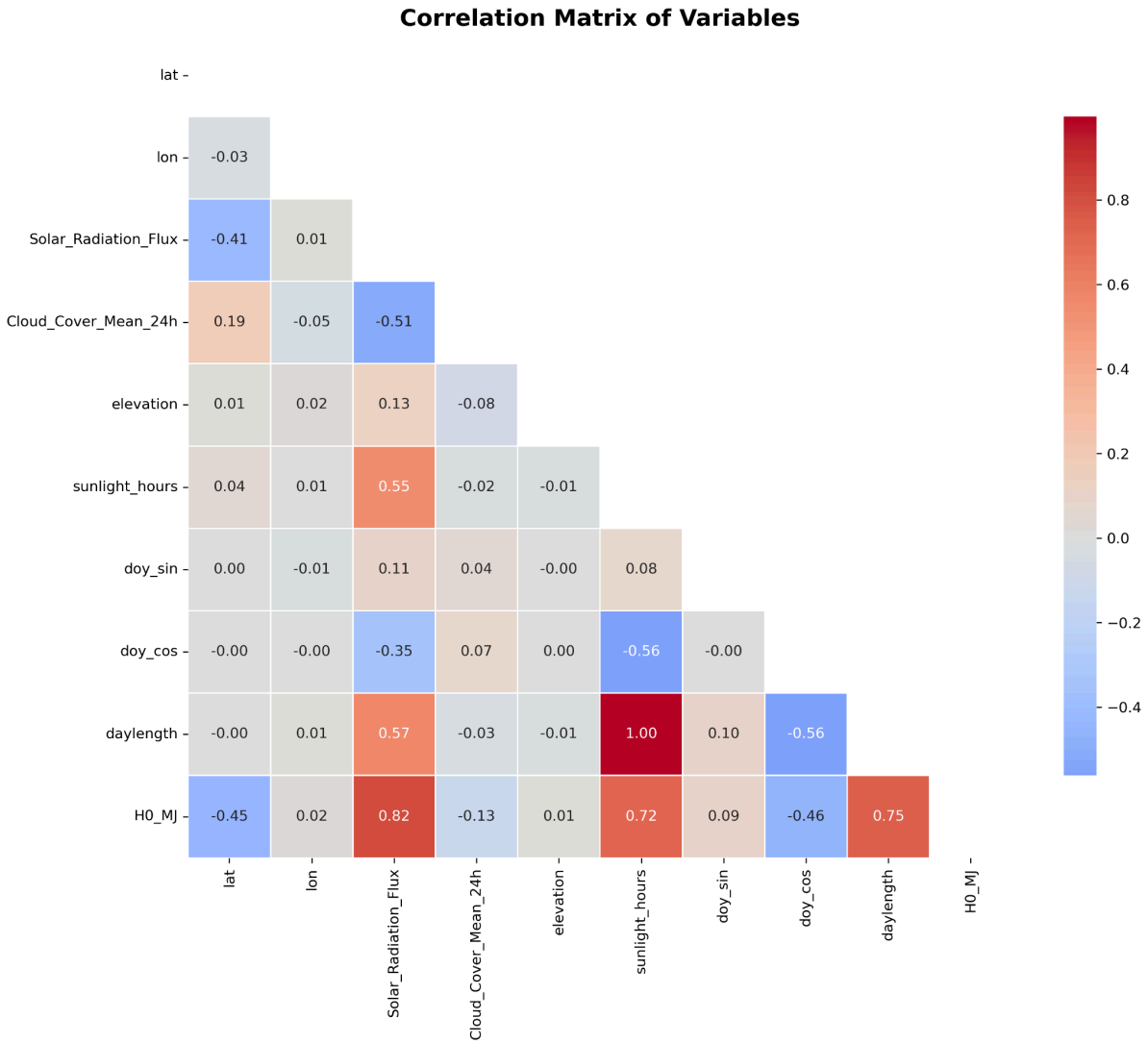


Figure 1: Correlation analysis

The target variable (solar radiation flux) is most strongly associated with cloudiness and astronomical features ($H_0$, daylength/sunlight hours). Geographical coordinates, and altitude provide additional, but less important contributions. Both daylength and sunlight hours have a strong correlation with the target variable, but daylength was chosen for training the model because it showed slightly higher correlation with the target.

## Modeling results

To assess the quality of predictions, both linear models (with different transformations) and nonlinear methods without additional transformations) were tested.

**Comparison of linear and nonlinear models**

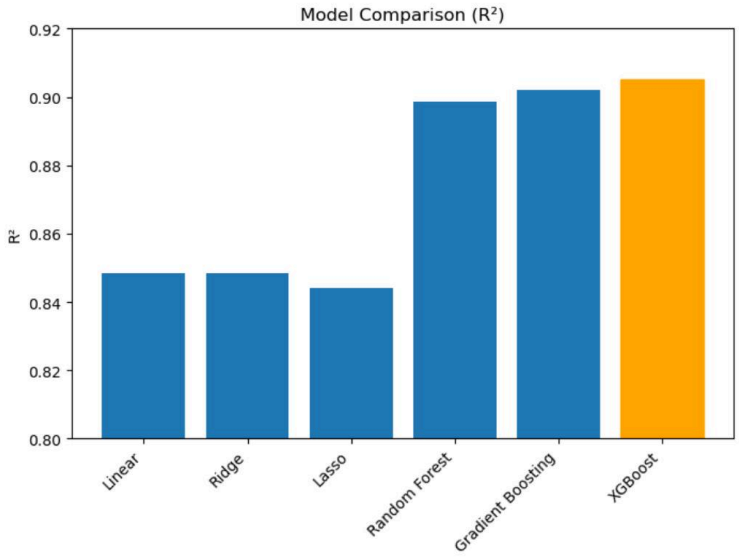


Figure 2:

Linear, Ridge, and lasso regressions perform similarly, with $R^2$ around 0.84, indicating that they capture general trends, but struggle with complex nonlinear relationships. Random Forest, Gradient Boosting, and XGBoost perform better, achieving $R^2$ slightly higher than 0.90. XGBoost has the highest $R^2$, indicating its ability to model complex feature interactions and improve prediction accuracy.

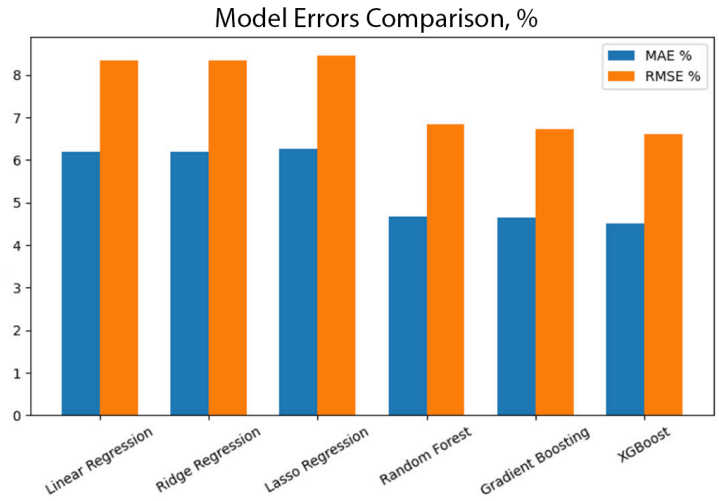


Figure 3:

Comparison of the model errors in terms of mean absolute error (MAE) and root mean square error (RMSE), expressed as a percentage. Linear, Ridge, and Lasso regressions show the highest errors (MAE ≈ 6-6.3% and RMSE ≈ 8.3-8.5%). Random Forest, Gradient Boosting, and XGBoost reduce errors (MAE ≈ 4.5-4.7%, RMSE ≈ 6.7%). XGBoost gives the lowest overall error, confirming the findings of $R^2$: it balances bias and variance better than other models. Nonlinear ensemble methods consistently reduce both absolute and squared errors, improving both accuracy and robustness.

**Feature importance for XGBoost to show which factors really matter.**

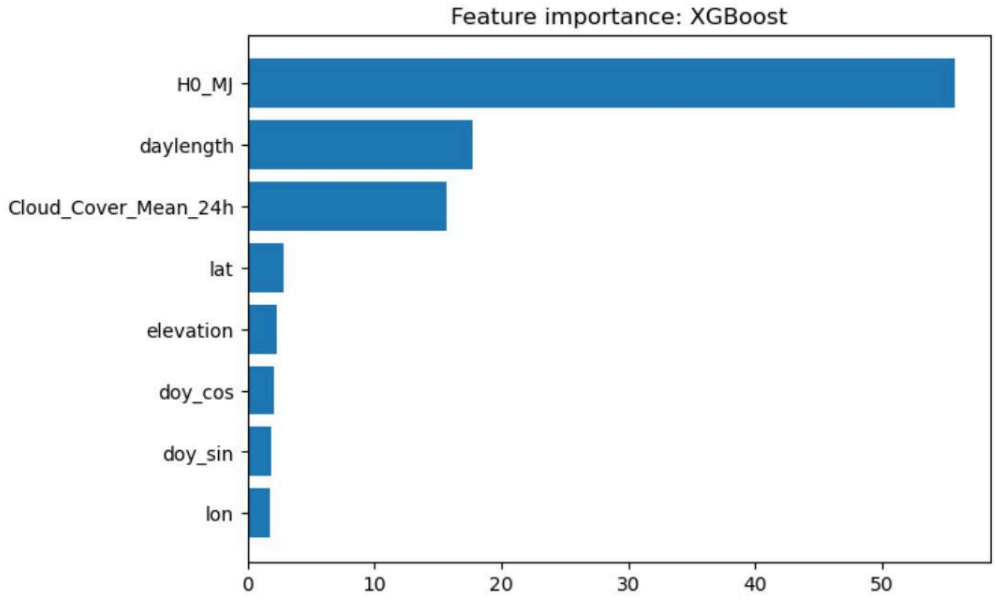


Figure 4:

The feature importance analysis obtained by the best performing model (XGBoost). The most influential predictor is H0_MJ (daily extraterrestrial solar radiation in MJ/m²), which dominates the model's prediction. This is physically justified as it reflects the maximum theoretical energy input before atmospheric effects. Daylength and cloudiness variables are the next most important factors, highlighting the role of seasonal photoperiod and weather conditions. The geographic variables (latitude, altitude, longitude) and cyclic representations of the day of the year (doy_cos, doy_sin) make minor contributions, but still help the model fine-tune the forecasts. The model relies primarily on astronomical-geometric factors (solar irradiance potential and day length), while weather (cloudiness) and local geography provide secondary refinements.

**Conclusion**

The problem of predicting solar radiation based on simple meteorological and geographical data is well solved by machine learning. Even linear models explain a significant part of the variation, but for practical applications, Gradient Boost, XGBoost and Random Forest are preferable.

**Stakeholder-facing summary**

The goal of the project is to make solar radiation forecasting simple and accessible to people and communities that want quick forecasts using simple input data. We have shown that it is enough to have simple input data - cloudiness from a weather forecast, geographic coordinates, and date.

Everything else is restored automatically using open geodata and code calculations.

Key results: Simple linear models already explain 84.8% of solar radiation variations. More advanced methods (XGBoost, Random Forest, Gradient Boosting) up to 90.5%. The forecast error was ~2.4 MJ/m² per day, which is enough for use in planning energy systems, agriculture, and sustainable projects.

**Importance:**

Applicable for solar energy planning, irrigation deficit in agriculture, and climate-resilient solutions. Simple and open data allows the method to be scaled without access restrictions.

**Ethical, social, and environmental aspects**

Environmental: accessible solar radiation forecasts help farmers optimize water use and energy systems, reducing their carbon footprint and resource intensity.

Social: simplified methods make climate data accessible to countries and regions that do not have access to professional weather forecasts.

Ethical: no personal data is used, high explainability of model's predictions.

**Limitations**

Cloud cover values from the AgERA5 reanalysis were used for training. However, in real-world applications, public weather forecasts are expected to be used. There may be differences in accuracy between these sources, which will affect the quality of the predictions.

The solar radiation flux from AgERA5 was used as the target variable. It is based on modeling and averaging, not on actual measurements. This means that the model is trained to reproduce not the "ideal" radiation, but the one calculated in the reanalysis. In some cases, this may introduce systematic deviations.

## Bibliography


1. Copernicus Climate Data Store (CDS). AgERA5 Solar Radiation Flux and Cloud Cover Datasets. https://cds.climate.copernicus.eu/
2. GMTED2010 elevation data at different resolutions https://www.temis.nl/data/gmted2010/
3. National Oceanic and Atmospheric Administration. General Solar Position Calculations https://gml.noaa.gov/grad/solcalc/solareqns.PDF
4. Chapter 3 - Meteorological data FAO - Food and Agriculture Organization of the United Nations Rome, 1998 https://www.fao.org/4/x0490e/x0490e00.htm#Contents