**SOLAR RADIATION ESTIMATION USING CLOUD PROPERTIES AND   
MACHINE LEARNING**

**SANSKAR R. SHETE** (TS90733)

**HEMANTH B.S.** (TS90719)

M.SC. AGRICULTURE ANALYTICS (2023-25)

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY, GANDHINAGAR

AND

INDIAN INSTITUTE OF REMOTE SENSING

(INDIAN SPACE RESEARCH ORGANIZATION), DEHRADUN

GUIDED BY: DR. N.R. PATEL AND MS. SHWETA POKHRIYAL

AGRICULTURE AND SOILS DEPARTMENT (ASD)

Abstract

Global solar radiation is the total amount of solar energy received in the horizontal plane and is defined as the sum of direct solar radiation, diffuse solar radiation, and reflected solar radiation. Global solar radiation is an important variable in agricultural, meteorological, hydrological, and climatological research. The objective of this study was to analyze various atmospheric properties obtained from satellite data, including cloud fraction, cloud optical depth, aerosol optical depth, aerosol index, and precipitable water vapor from (MODIS). is used to develop an effective method for estimating daily solar radiation. A model was developed using standard statistical methods and artificial neural networks based on seven combinations of atmospheric properties. To evaluate solar radiation, we compared the efficiency of the models based on the input data combinations.

# Introduction

## Global solar radiation is defined as the total amount of solar energy received by the Earth's surface and is the sum of direct, diffuse, and reflected solar radiation. Global solar radiation values ​​are important variables needed for a variety of applications, including ecological and hydrological modeling, plant growth and productivity estimation, and solar energy applications. Ground measurements are a reliable method. Many statistical modeling methods are designed for practical applications in estimating global solar radiation indirectly from other atmospheric variables. These models are less complex and require fewer computing resources. Statistical methods can be classified into parametric and nonparametric methods. However, parameters often vary from model to model. Instead of traditional statistical models, artificial neural networks are also used to estimate global solar radiation. Yadav and Chandel (2014) comprehensively investigated the applicability of artificial neural networks for predicting solar radiation. Satellite observations provide extensive continuous signals in space and time, capturing atmospheric parameters that influence daily global solar radiation. Deo and Şahin (2017) used satellite surface temperature data as an effective predictor in an artificial neural network to estimate global solar radiation. The results showed that the use of satellite data can outweigh the benefits of measured datasets in solar energy assessments, especially in some regions where observatory data are not available. Bisht and Bras (2010) estimated net radiation using satellite products such as cloud top temperature, cloud fraction, cloud emissivity, and cloud optical depth. The reliability of the proposed method relies on remote sensing data due to the lack of ground-based observations, which also has the potential to estimate the global surface energy budget. In order to develop an effective method to estimate daily global solar radiation, this study investigates cloud fraction, cloud optical depth, aerosol optical depth, aerosol index, aerosol index, and precipitation. Consider various atmospheric properties from satellite data, including possible water vapor. This study compares a specific class of neural networks with a standard statistical method, namely regression analysis, to evaluate daily global solar radiation based on recommended combinations of input data.

## Study Area

We have collected meteorological data from 16 stations in India namely: Bhubaneshwar (OR), Calcutta (WB), Dehradun (UK), Delhi (ND), Goa (GA), Jaipur (RJ), Jaisalmer (RJ), Machilipatnam (AP), Mumbai (MH), Nagpur (MH), Patiala (PB), Patna (BH), Port Blair (AN), Ranchi (JH), Shillong (ML) and Vishakhapatnam (AP).

### Datasets

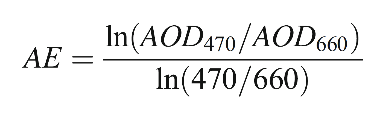
Two types of datasets are used in this study. The first is daily global solar radiation data from 2000 to 2012. Data quality control continued to check for errors and discrepancies, and data that did not meet important criteria was removed. All data were normalized and outliers were also normalized. The second dataset consists of daily satellite data including Moderate Resolution Imaging Spectroradiometer (MODIS) and Ozone Monitoring Instrument (OMI), which are used as predictor variables to estimate global solar radiation. will be done. MODIS, aboard the polar-orbiting satellite NASA-EOS Terra, has been providing aerosol-related parameters worldwide since 2000. Additionally, daily global coverage provides detailed information on aerosols and clouds over land and oceans. The spatial resolution of the data is 1° × 1°, and the uncertainty has been recorded to be approximately 25–30% (Hsu et al. 2006). In this study, we will consider his 5 products of MODIS equipment, including-

1. Aerosol optical depth (AOD) indicates the amount of aerosol loaded into the atmospheric column. It is the vertical integral of the fraction of incident light that is scattered or absorbed by the aerosol. The AOD value indicates the role of light extinction by atmospheric particles of different sizes along the optical path between the sun and the Earth's surface. AOD is written by the following vertical integral equation:

AOD(λ) = ∫σext(λ, z)dz

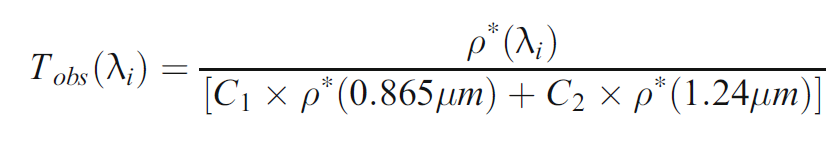
where σ extis is the extinction coefficient at wavelength λ at height z (Kokhanovsky 2008). Levy et al. (2010) showed that almost 80% of MODIS AOD searches had errors less than 0.01. Therefore, we introduce MODIS AOD as a quantitative product useful for model assimilation. Aerosol optical depth values ​​determined with MODIS were evaluated worldwide using ground-based measurements. Results show that more than 66% of MODIS AODs are within the expected error range of ± 0.05 + 15%, with a high correlation of approximately 0.9 (Levy et al. 2010).

1. Angstrom exponent (AE) measures aerosol size distribution. This shows the spectral dependence of aerosol extinction based on measurements of AOD at two different wavelengths (Chu et al. 2003). Higher AE values are indicate with smaller particles, lower values indicate the presence of larger particles. MODIS AE is defined as:

****

Where AOD470 and AOD660 are the aerosol optical depths at the wavelengths specified, 470 and 660nm, respectively (Remer et al. 2008). Previous studies show that the MODIS-derived AE is well correlated with ground-based measurements; however, there is a tendency to overestimate the presence of fine-mode aerosols (Kleidman et al. 2005).

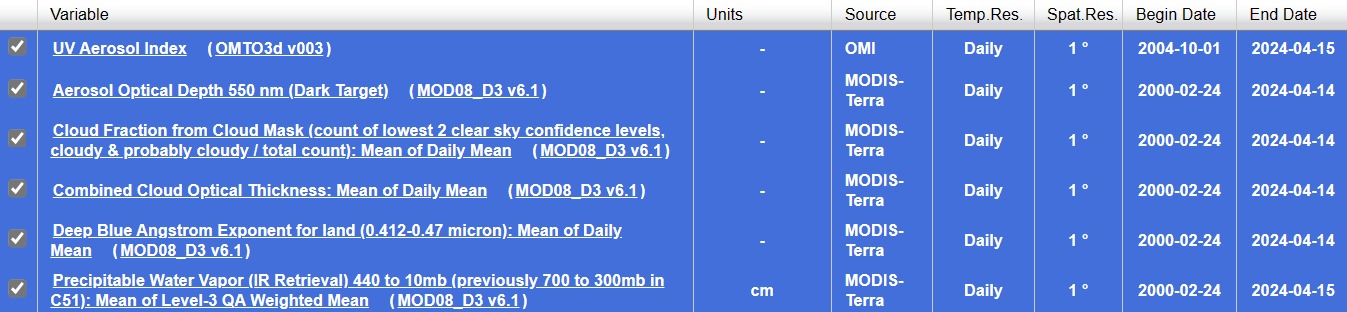
1. Cloud fraction (CF) is the fraction of hemispherical sky covered by clouds (Hahn et al. 2001). It is a critical cloud macrophysical parameter and a key factor in controlling Earth’s radiation budget both at the surface and at the top of the atmosphere (Stephens et al. 2002). The MODIS cloud fraction product is defined as the ratio of the number of cloudy pixels to the total number of pixels within a specified area centred on a selected site. It is worth mentioning that satellite cloud cover is affected by many uncertainties due to cloud inhomogeneities as well as surface brightness and changes. Therefore, the cloud detection algorithm developed for MODIS uses different tests to detect the presence or absence of clouds and then calculates a confidence score for each test. Confidence scores for each test are then evaluated and additional backtesting is used to reduce debugging due to brightness. Finally, a cloud detection confidence level is generated.
2. Cloud optical depth (COD) is the physical basis to retrieve the COD is the bi-spectral solar reflectance method which was first demonstrated with airborne data by Nakajima and King (1990). To retrieve the cloud optical depth, a radiative transfer model is applied to compute the reflected intensity field. The MODIS-derived COD is theoretically based on the reflection function of the cloud in the non-absorbent visible band (Li et al. 2019).
3. Precipitable water vapour amount (PWV) is the total atmospheric water vapour contained in a vertical column of a cross-section unit (Ichoku et al. 2002). To retrieve PWV from MODIS data, near-infrared bands are used. PWV is retrieved from atmospheric transmittance for water vapour absorption bands as follows:



where Tobs(λi) is the transmittance at band λi where i represents bands at the wavelength of 0.905, 0.936, and0.94 μm. ρ∗(λi) is the apparent reflectance at each band. C1 and C2 values are 0.8 and 0.2 for bands at 0.865 and1.24 μm, respectively (Wang and Liu 2020). MODIS- derived PWV has been widely evaluated against ground-based data in many previous studies (i.e. Gao et al. 2004; Prasad and Singh 2009; Wang et al. 2017).

Another product that is used in the current study is the UV aerosol index (AI), which is derived from OMI onboard NASA’s EOS-Aura satellite that was launched in July 2004. We measure the absorption of aerosols using the aerosol index. OMI is a nadir-viewing spectrometer, which measures solar reflected and backscattered radiation in the ultraviolet-visible spectrum. It has both the spatial and temporal resolution of 1° and once per day (Levelt et al. 2006). The Aerosol index is a quantity that shows the departure of the spectral dependence of backscattered UV radiation in an atmosphere containing aerosols from a pure Rayleigh atmosphere (Torres et al. 1998). OMI uses two wavelength intervals around 331nm and 360nm to define the aerosol index. AI is an indicator of light-absorbing aerosols that discriminate desert dust and sea salt aerosols that have UV-absorbing and non-absorbing properties, respectively.

It is also possible to separate the smoke as an absorbing aerosol from industrial pollution that is non-absorbing aerosol. Several regression methods and an artificial neural network algorithm are developed to estimate global solar radiation from satellite data. The data set was split into two parts: 80% of the records were used to develop the models and the remaining to evaluate the models

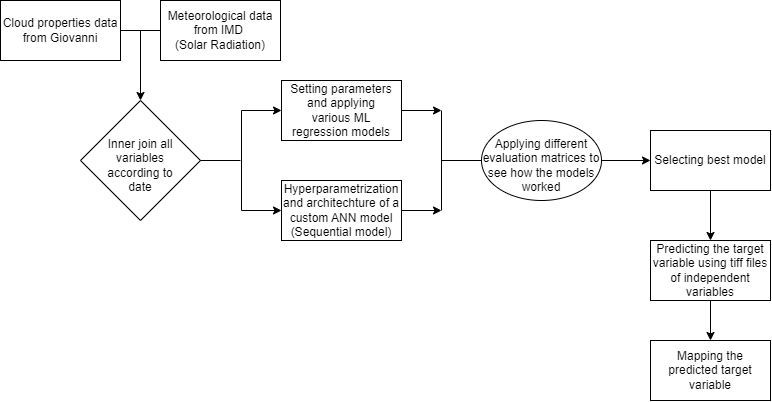
**Cloud properties used:**

#### Heatmap showing correlation between the different variables:

#### 

#### Methods

Multiple Linear Regression, Ridge, Random Forest Regressor, Decision Tree Regressor and Xgboost are some of the regression models and an artificial neural network (ANN) algorithm is developed to estimate global solar radiation using the aforementioned atmospheric parameters. Studies show that in most cases the cloud cover weakens solar radiation intensity (e.g., Matuszko 2012). As the atmospheric transmittance of radiation changes with the content of the atmosphere, an increase in cloud cover leads to an increasing scattering of radiation by water droplets and results in a decrease in the atmospheric light transmittance. Regressions were carried out between the daily mean measured global solar radiation and the cloud parameters. Briefly, the neural network used is a Sequential Model. The performances of the models are evaluated using the coefficient of determination (R2), mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE) and index of error (IoE)



#### Software and Algorithms used (Theory)

1. **Libraries**
   1. **Pandas:**

Offers data structures and operations for manipulating numerical tables and time series. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data — load, prepare, manipulate, model, and analyse.

## Numpy:

NumPy serves as the foundational package for scientific computing in Python. In addition to its primary scientific applications, NumPy serves as a high-performance multi-dimensional data container supporting various data types. This flexibility enables seamless integration with diverse databases. NumPy also offers functions for operations in linear algebra, Fourier transforms, and matrix computations

* 1. **Scikit-learn**:

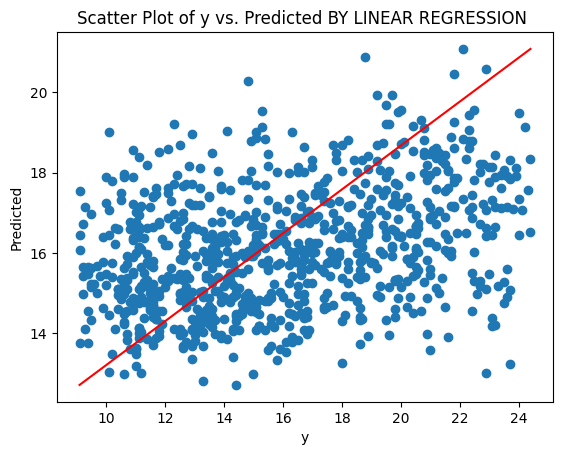
## Scikit-learn, known as sklearn, is a Python library for machine learning that provides supervised and unsupervised algorithms for tasks like classification, regression, clustering, and dimensionality reduction. Its API is consistent and easily integrates with popular Python libraries such as NumPy, Pandas, and Matplotlib. With comprehensive documentation and examples, scikit-learn caters to users of all levels, from beginners to experts. Being open-source and community-driven, it boasts a large developer and contributor community, making it a prevalent and robust library for machine learning applications in both industry and academia.

## Matplotlib:

Matplotlib, a Python library for data visualization, enables the creation of a diverse range of plots and charts. Leveraging NumPy as its foundation, Matplotlib seamlessly integrates with various scientific computing libraries. It offers a user-friendly interface for customizing plots and supports multiple output formats. Widely utilized in scientific computing, data analysis, and machine learning projects, Matplotlib is a go-to tool for visualizing data effectively.

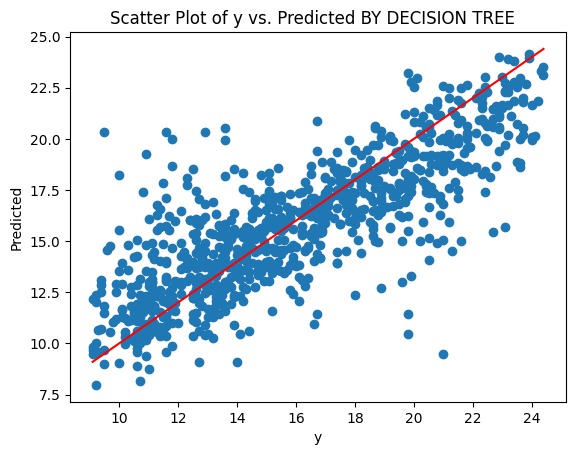
1. **Algorithms Used**
   1. **Multiple Linear Regression**

MLR is a simple yet effective algorithm used for predicting a continuous outcome variable (Y) based on predictor variables (X). It assumes a linear relationship between the independent variables (X) and the dependent variable (Y). It is calculated using the formula y=mx+c



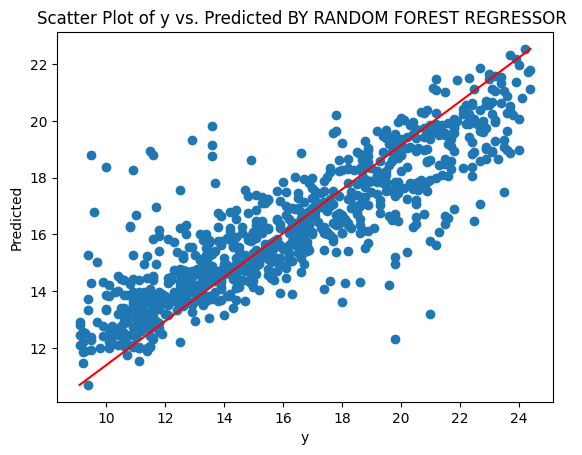
* 1. **Decision Tree**

A Decision Tree is a tree-like structure in which each internal node represents values, each branch represents a decision rule, and each leaf node represents an output. The toppest node is known as the root node. It learns to partition based on the attribute value.



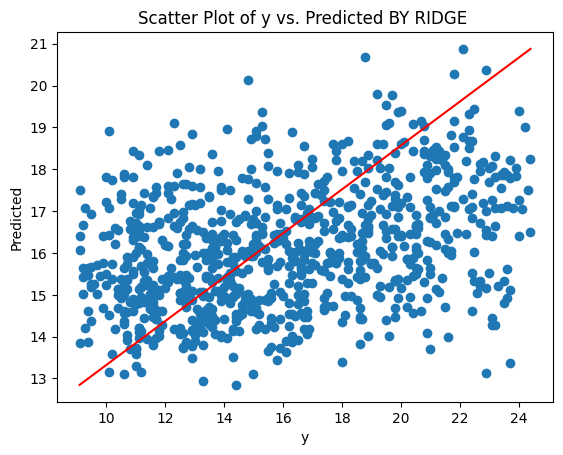
* 1. **Random Forest**

Random Forest is a supervised learning technique. It is based on the concept of ensemble machine learning, which is a process of combining multiple tree like structures to solve a particular problem.



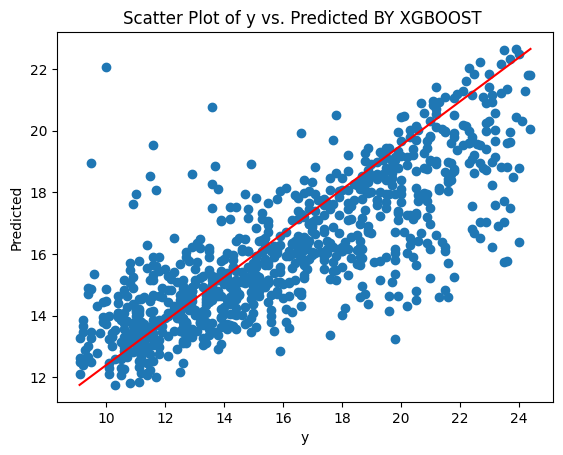
* 1. **Ridge**

When the data suffers from multicollinearity, we use ridge. It is like linear regression but to avoid over-fitting, we add slope square to the cost function. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.



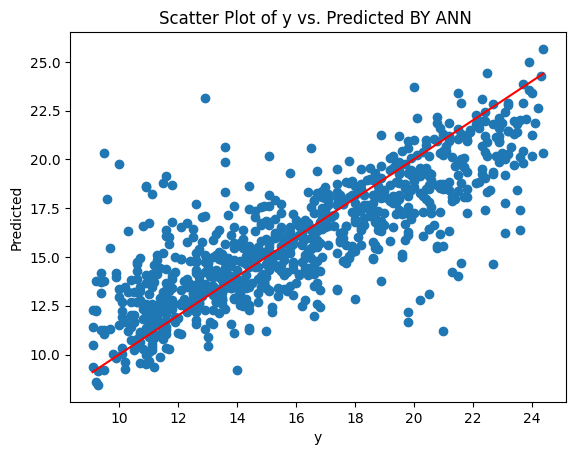
* 1. **XGBoost**

XGBoost stands for eXtreme Gradient Boosting. It is just like random forest but the boosting function is introduced. It gives a much better result than decision tree and random forest.



* 1. **Artificial Neural Network (ANN)**

Artificial Neural Networks (ANNs) are multi-layer fully connected neural nets that look like the human brain. They are composed of 3 layers; input, hidden and output layers. In each of the output from thee nodes, bias and activation function will be added.



1. **Evaluation Matrices**
   1. **R-squared (R2)**

R-squared, or the coefficient of determination, is a statistical measure that represents the goodness of fit of a regression model. As the r square increases, the model fits better. It provides a measure of how well the observed outcomes are replicated by the model, as it quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables.

* 1. **Root Mean Square Error (RMSE)**

It is a standard way to measure best fit line. It signifies the standard deviation of the residuals or prediction errors. Residuals are a measure of how far from the regression line data points are, and RMSE is a way to measure how spread out these residuals are.

* 1. **Mean Absolute Error (MAE)**

It is an alternative for rmse. It’s an average of the absolute differences between the predicted and actual values. It gives an idea of how predictions vary from the actual values.

* 1. **Mean Absolute Percentage Error (MAPE)**

Mean Absolute Percentage Error (MAPE) is a statistical measure to define the accuracy of an algorithm in statistics, specifically trending algorithms. It is used to describe the percent difference between the actual and predicted values.

* 1. **Index of Agreement (IoA)**

It is a standardized measure to analyse the performance of the model and ranges from 0 to 1, with 1 being a perfect match, 0 being no match at all. It is a good measure to determine the overall performance of a model’s predictions.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Alg\Eval | R2 | RMSE | MAE | MAPE | IoA |
| MLR | 0.12 | 3.79 | 3.23 | 0.22 | 0.33 |
| Decision Tree | 0.73 | 5.66 | 4.58 | 0.31 | 0.91 |
| Random Forest | 0.67 | 3.94 | 3.33 | 0.23 | 0.67 |
| Ridge | 0.12 | 3.78 | 3.23 | 0.22 | 0.32 |
| XGBoost | 0.58 | 3.97 | 3.25 | 0.22 | 0.62 |
| ANN | 0.61 | 4.75 | 3.95 | 0.27 | 0.69 |

## Results

In our study, we evaluated six different machine learning algorithms for their predictive performance: Multiple Linear Regression (MLR), Decision Tree, Random Forest, Ridge Regression, XGBoost, and Artificial Neural Network (ANN). The Decision Tree model demonstrated the highest R-squared value of 0.73, indicating that it was able to explain 73% of the variance in our target variable. However, it had a relatively high Root Mean Square Error (RMSE) of 5.66, suggesting that the model’s predictions deviated from the actual values.

The Ridge Regression and MLR models showed the lowest RMSE and Mean Absolute Error (MAE), both indicating smaller average prediction errors. However, these models had a lower R-squared value of 0.12, suggesting they were less effective at explaining the variance in the target variable.

The Random Forest and XGBoost models demonstrated a balance between predictive accuracy (as indicated by RMSE and MAE) and explanatory power (as indicated by R-squared), with the Random Forest model slightly outperforming XGBoost.

The ANN model showed a relatively high R-squared value of 0.61 but had higher error rates (RMSE and MAE) compared to other models.

In terms of the Index of Agreement (IoA), the Decision Tree model again performed the best with an IoA of 0.91, indicating a high degree of agreement between the model’s predictions and the actual values.

**Conclusions**

The Decision Tree Regressor performed well with R2 Score of 0.73 and RMSE Score of 5.66 in predicting solar radiation using cloud properties. However, its performance on satellite-based data needs further investigation. The Random Forest Regressor also showed promise with R2 Score of 0.67 and especially RMSE Score of 3.94. It could be considered as an alternative to the Decision Tree model. The ANN has potential but requires more data to explore further because it gave an R2 Score of 0.61 and RMSE Score of 4.75. If additional data becomes available, the ANN could be a valuable option.

**References:**

Sara Bamehr, Samaneh Sabetghadam (2020) Estimation of global solar radiation data based on satellite-derived atmospheric parameters over the urban area of Mashhad, Iran. Environmental Science and Pollution Research (2021) 28:7167–7179

Akpabio LE, Udo SO, Etuk SE (2004) Empirical correlations of global solar radiation with meteorological data for Onne, Nigeria. Turk J Phys 28(3):205–212