# Machine Learning Modeling of Solar Output in the Northern Hemisphere

**Shubham Sharma**
190010065
Aerospace Engineering
shubh.am@iitb.ac.in

**Rupansh Parth Kaushik**
200260043
Engineering Physics
200260043@iitb.ac.in

**Shivam Ambokar**
200100145
Mechanical Engineering
200100145@iitb.ac.in

## Abstract

One of the most popular renewable energy sources in the world, solar energy is on the rise. However, it is reliant on sunshine, a transient natural resource. Because of this, the ability to estimate power generation is essential for integrating solar photovoltaics into our existing power systems. We dive into the data collected at 12 Northern hemisphere sites over 14 months and use only the location and weather data to make predictions about the power output, neglecting irradiance data which is difficult to collect and prone to errors.

## 1 Introduction

The amount of electricity produced by solar photovoltaics (PV) is predicted to increase by $30\%$ over the next five years, with distributed solar PV systems accounting for the majority of this rise. In urban areas as well as more dispersed locations (such as isolated military stations), where it may be difficult to construct substantial, centralised PV arrays, distributed PV can be advantageous to residential consumers and commercial/government institutions. The problem of intermittent solar energy is widely known and emphasises the need of projecting solar PV power generation, particularly in a dispersed setting.
Our contributions in this project are the following:

- Attempted to predict power output without irradiance data in order to save time, effort, and cost with no significant loss of accuracy.
- Explored variables for solar output prediction in the dataset by visualising and pre-processing before being passed to the machine learning algorithms.
- Experimented with different machine learning algorithms to see what strategies work best on the given problem.

## 2 Prior Work

Solar PV power forecasting has been studied extensively. Lorenz et al. (2014) provided an overview [1], and Raza et al. (2016) discussed recent advances [2]. Often, solar power forecasting studies are based on predicting irradiance or using historical power output. Yang et al. (2015) used exponential smoothing to improve predictions of horizontal irradiance [3]. Lorenz et al. (2010) used regional weather data to forecast irradiance, which was then converted to power [4]. Various studies have considered predicting irradiance or power using weather and prior power output data. Additionally, previous studies forecasting solar irradiance or power output are often based on data from a limited number of locations.
Our work is inspired by "Machine Learning Modeling of Horizontal Photovoltaics Using Weather and Location Data" submitted to the Journal of Renewable Energy in 2020. The dataset used accompanies the paper and is the largest available public datasets on the topic.

# 3 Dataset and Methodology

The dataset we used is made publicly available by the authors on Kaggle. It contains power output from horizontal photovoltaic panels located at 12 Northern hemisphere sites over 14 months. Several factors identified as important by prior research make up the data classes for the datase:

- **Cloud Ceiling:** the presence of clouds above a panel will scatter solar irradiance and decrease the amount of irradiation a panel receives; the cloud ceiling is measured at the altitude where at least 5/8ths of the sky above the weather station is covered by clouds.
- **Latitude:** the latitude of each location will dictate the sun deflection angle; this will affect the amount of sunlight the panel receives.
- **Month:** when the sun rises and sets and how high it will appear in the sky at any location on the earth is determined (in part) by the time of year at that location.
- **Hour:** the time of day determines how high the sun is in the sky—or whether or not it is present at all. Hour controls for the sun's position in relation to the time of day.
- **Humidity:** water affects incoming sunlight through refraction, diffraction, and reflection. Indirectly, humidity also affects dust build-up on panels due to the formation of dew increasing coagulation of dust.
- **Temperature:** the efficiency of a solar panel will generally decrease with an increase in panel temperature. Including temperature as an explanatory variable for power output has led to increased predictability
- **Wind speed:** the temperature of the panel may be affected by the speed of the wind surrounding the panel. Increased wind speed can also clean the dust off of the panel surface or stir up dust, thereby affecting the irradiance that reaches the panel.
- **Visibility:** this variable is a measurement of the distance at which a light can be seen and identified. Visibility will primarily affect how much irradiation reaches the panel and can have a negative effect on power output if visibility is low during daylight hours.
- **Pressure:** Pressure may have an effect on power output predictability by indicating a weather occurrence—such as a storm. this variable has not been extensively explored in solar panel power output literature.
- **Altitude:** there is less atmosphere for the sun to travel through at locations with higher altitudes; this results in a higher level of irradiation at locations farther above sea level.

We then perform exploratory data analysis on the dataset to find out the most relevant features in solar output prediction

# 4 Exploratory Data Analysis

## 4.1 Data pre-processing

The dataset consists of 21,045 rows and 17 columns shown below.

| | Location | Date | Time | Latitude | Longitude | Altitude | YRMODAHRMI | Month | Hour | Season | Humidity | AmbientTemp | PolyPwr | Wind.Speed | Visibility | Pressure | Cloud.Ceiling |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Camp Murray | 20171203 | 1145 | 47.11 | -122.57 | 84 | 2.017120e+11 | 12 | 11 | Winter | 81.71997 | 12.86919 | 2.42769 | 5 | 10.0 | 1010.6 | 722 |
| 1 | Camp Murray | 20171203 | 1315 | 47.11 | -122.57 | 84 | 2.017120e+11 | 12 | 13 | Winter | 96.64917 | 9.66415 | 2.46273 | 0 | 10.0 | 1011.3 | 23 |
| 2 | Camp Murray | 20171203 | 1330 | 47.11 | -122.57 | 84 | 2.017120e+11 | 12 | 13 | Winter | 93.61572 | 15.44983 | 4.46836 | 5 | 10.0 | 1011.6 | 32 |
| 3 | Camp Murray | 20171204 | 1230 | 47.11 | -122.57 | 84 | 2.017120e+11 | 12 | 12 | Winter | 77.21558 | 10.36659 | 1.65364 | 5 | 2.0 | 1024.4 | 6 |
| 4 | Camp Murray | 20171204 | 1415 | 47.11 | -122.57 | 84 | 2.017120e+11 | 12 | 14 | Winter | 54.80347 | 16.85471 | 6.57939 | 3 | 3.0 | 1023.7 | 9 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 21040 | USAFA | 20180928 | 1530 | 38.95 | -104.83 | 1947 | 2.018090e+11 | 9 | 15 | Fall | 11.66992 | 43.22510 | 9.79611 | 14 | 10.0 | 802.3 | 722 |
| 21041 | USAFA | 20180929 | 1300 | 38.95 | -104.83 | 1947 | 2.018090e+11 | 9 | 13 | Fall | 18.22510 | 28.98247 | 10.88992 | 13 | 10.0 | 799.2 | 722 |
| 21042 | USAFA | 20180929 | 1400 | 38.95 | -104.83 | 1947 | 2.018090e+11 | 9 | 14 | Fall | 15.52124 | 33.49167 | 8.24479 | 10 | 10.0 | 798.4 | 722 |
| 21043 | USAFA | 20180929 | 1500 | 38.95 | -104.83 | 1947 | 2.018090e+11 | 9 | 15 | Fall | 6.63452 | 51.62163 | 12.47328 | 10 | 10.0 | 797.8 | 722 |
| 21044 | USAFA | 20181001 | 1400 | 38.95 | -104.83 | 1947 | 2.018100e+11 | 10 | 14 | Fall | 22.58301 | 32.83958 | 6.39732 | 15 | 10.0 | 801.2 | 110 |

21045 rows x 17 columns

**Figure 1:** The dataset

To our relief we found that there are no missing values in the dataset. There is one feature **"YR-MODAHRMI"** that seems non-intuitive and is somehow related to date. We dropped this column as it was non-interpretable. Humidity, AmbientTemp, PolyPwr, Pressure, Cloud.Ceiling and Date are the continuous variables in our analysis. It is presumed that time precise to minutes is not something that should affect the solar irridiation to a very great extent. Besides, it may be noted that the time is not present for each and every minute, but rather for every few hours.

**Table 1:** Basic Statistics

| Variable Name | Units | Max | Min | Mean | Median | $1^{st}$ Quantile | $3^{rd}$ Quantile |
|---|---|---|---|---|---|---|---|
| Latitude | Degree | 47.52 | 20.89 | 38.21 | 38.95 | 38.16 | 41.15 |
| Longitude | Degree | -80.11 | -156.44 | -108.59 | -111.18 | -117.26 | -104.71 |
| Altitude | Metre | 1947 | 1 | 798.84 | 458.0 | 2.0 | 1370.0 |
| Humidity | % | 99.98 | 0.0 | 37.12 | 33.12378 | 17.52 | 52.59 |
| AmbientTemp | Celsius | 65.73 | -19.98 | 29.28 | 30.28 | 21.91 | 37.47 |
| PolyPwr | Watt | 34.28 | 0.25 | 12.97 | 13.79 | 6.40 | 18.86 |
| Wind.Speed | km/h | 49 | 0 | 10.32 | 9.0 | 6.0 | 14.0 |
| Visibility | km | 10.0 | 0.0 | 9.70 | 10.0 | 10.0 | 10.0 |
| Pressure | Millibar | 1029.5 | 781.7 | 925.94 | 961.1 | 845.5 | 1008.9 |
| Cloud.Ceiling | km | 722 | 0 | 515.96 | 722.0 | 140.0 | 722.0 |

### 4.2 Data Visualisation

To begin with, variables were plotted in the form of histograms

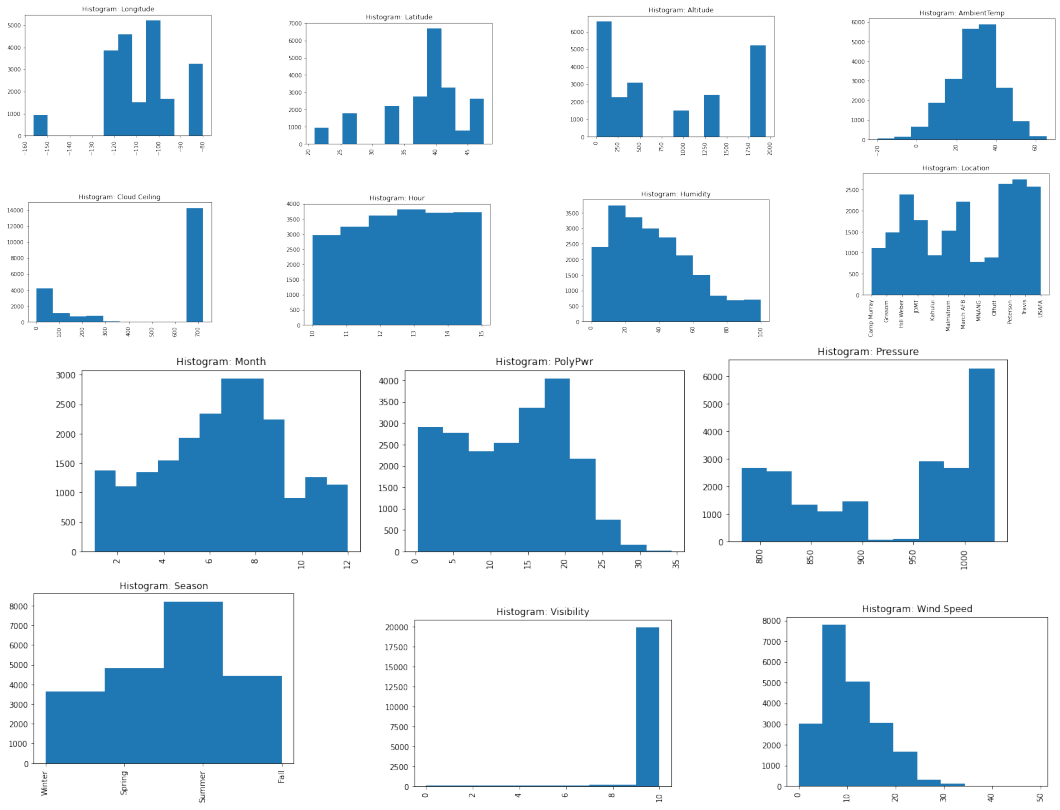
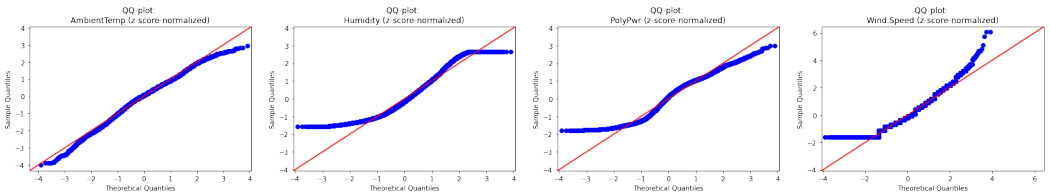



**Figure 2:** Plotting histograms of different variables against the prediction target

We infer the following:

1. Visibility appears to be very great in most cases and is heavily skewed

2. Cloud ceiling also appears heavily skewed, with either very high or very low values
3. Pressure values have a "gap" in between, which could be related to the altitudinal placement of the power generation devices
4. Variation in latitudes is not high. We have noted that latitudes have significant effects on Power generation.
5. Most of the other variables have a somewhat Gaussian distribution (This is inferred from the QQ-plots presented hence)
    - Windspeed is skewed
    - Visibility and Cloud ceiling are heavily skewed
    - The rest have thin tails, except for AmbientTemp
    - Our target variable, **PolyPwr** appears to fit to a normal-qq plot when z-score normalized, with thin-tails

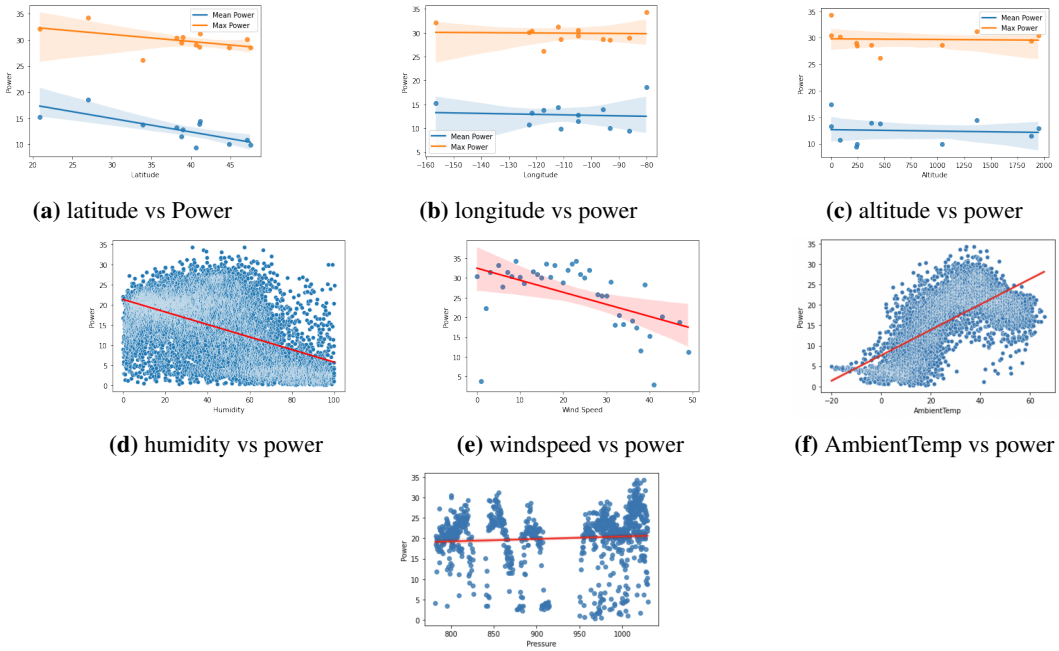



**(a)** AmbientTemp **(b)** Humidity **(c)** Power **(d)** Windspeed

**Figure 3:** QQ-plots of z-score normalized variables

### 4.3 Variable Trends

We now plot some relevant variables against the mean and max power values to get an idea of the most influential variables in the analysis.




**(a)** latitude vs Power **(b)** longitude vs power **(c)** altitude vs power

**(d)** humidity vs power **(e)** windspeed vs power **(f)** AmbientTemp vs power

**(g)** pressure vs power

**Figure 4:** Plotting features against the prediction target

### 4.4 Correlation Heatmap

To see the dependence amongst these variables, we plot a correlation heatmap of the data shown in **Figure 5**.
It would be great if we could get a continuous variable considering time, month and year altogether. A continuous time feature, that is. However, an even better feature would be one which is periodic in nature such that for a given date across a year, the value of the variable would be the same. We will need the data only in a specific interval of time. The time in which the systems are expected to generate power. Taking the cyclic date features using sine and cosine functions we again plot the heatmap with the new columns in **Figure 6**.

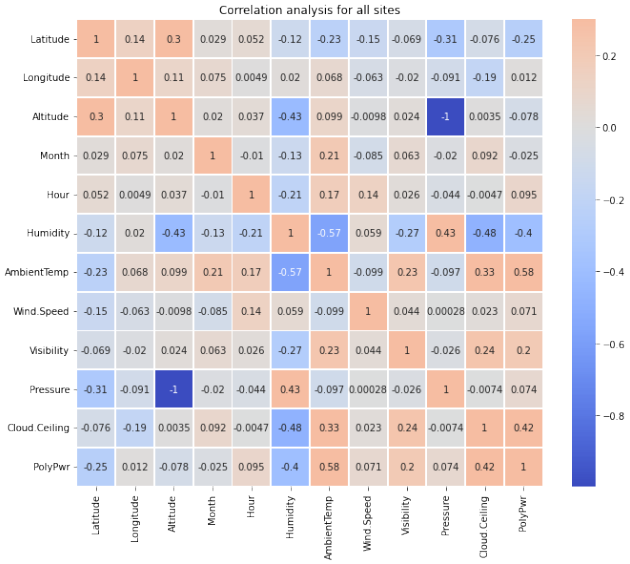



**Figure 5:** Correlation Heatmap

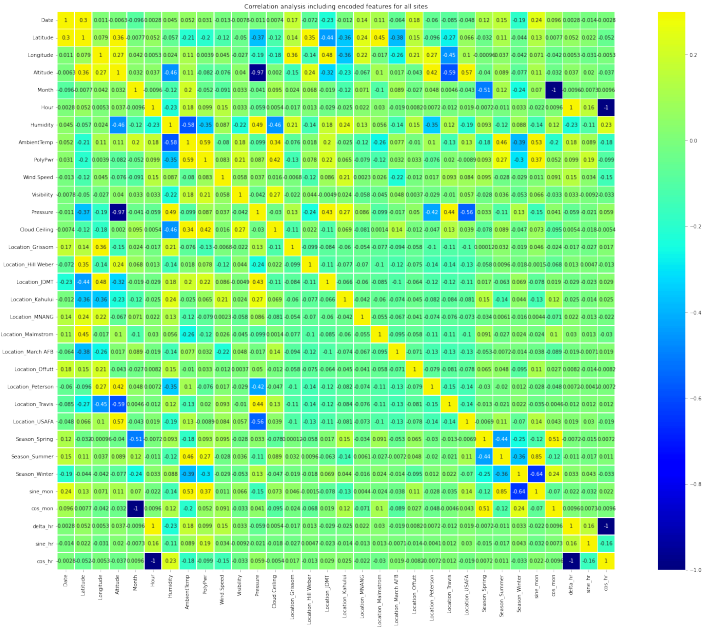



**Figure 6:** Correlation Heatmap with encoded features

### 4.5 Observations

At this stage, we make the following observations:

1. It is presumed that time precise to minutes is not something that should affect the solar irridiation to a very great extent. Besides, it may be noted that the time is not present for each and every minute, but rather for every few hours.
2. The Latitudinal position of the plant seems to be affecting the power output. Higher latitudes correspond to lower power extraction, which is what is expected (solar irradiation is the strongest near the equator, that is, 0-degree latitude)
3. It seems the longitudinal position doesn't affect the power to a great extent. This is also in line with our expectations that the mean power at a given latitude should be the same, irrespective of the longitude
4. Altitude seemingly does not affect the power output to a great extent.
5. The Solar Power appears to be decreasing with increasing wind speeds. However, the linear fit is not the best curve, as is apparent, though it clearly shows the relationship between the two quantities.
6. The overall trend seems to be a decreasing one, though it is not very crystal clear, in the case of humidity vs power.
7. The trend of Power with respect to AmbientTemp is overall increasing
8. The variation of Power with Pressure is too chaotic, but overall slightly upwards
9. Pressure and altitude are heavily correlated. One must therefore be removed. We will remove altitude since pressure is more readily measurable

## 5 Seasonal Analysis

To get more insights into the high productive times of the year for solar power generation, we do a seasonal analysis of the power output from these 12 sites individually as they are located at different geographical locations in the northern hemisphere (**Figure 7**).


**Figure 7:** Seasonal distribution of the 12 sites

As expected, the most power is extracted in the "warm" seasons of summer and spring with the following being the most productive months for these places:
**Summer:** Travis; Hill Weber; USAFA; JDMT; Peterson; March AFB; Malmstrom; Grissom; Offutt; Camp Murray. MNANG
**Spring:** Kahului
The largest share is mostly that of Summer, except for at Kahului where it is maximum in the spring season

# 6 Predictive Analysis

Relevant features were decided upon based on the Elementary Data Analysis. The data was then split between training and test datasets in 80:20 ratio. The data was standardized and was fed into "Random Search CV" for hyperparameter optimization on a **4-fold** cross validation set. The performance of the various models with best-parameters on unseen test data was measured using RMSE, R2-score and MAE. The four models used are KNN, Deep Neural Network, Random Forest and LGBM.
Finally, a stacked ensemble model was created using all the above models. These four distinct models (namely, KNN, DNN, RF, and LGBM) were combined using the stacking-regressor module in the library Scikit-Learn. The meta learner was a simple-linear-regression model which was trained on the 4-fold cross validated base model predictions, as well as the original input features.

## 6.1 Performance Summary

The performance of each model is evaluated using the hold-out set which is 20% of the entire dataset. The results are summarized below. The stacked model has the overall best performance with a 10% improvement compared with the KNN (baseline) model. In addition, the LGBM model is the best base model based on all metrics considered.

**Table 2:** R2-score for Validation Set

| Model | R2-score |
|---|---|
| KNN | 0.629 |
| DNN | 0.661 |
| LGBM | 0.670 |
| Random Forest | 0.671 |

**Table 3:** Performance of models over Test Dataset

| Model | R2-score | RMSE | MAE |
|---|---|---|---|
| KNN | 0.62 | 4.4 | 2.97 |
| DNN | 0.659 | 4.16 | 2.70 |
| LGBM | 0.674 | 4.06 | 2.73 |
| Random Forest | 0.670 | 4.09 | 2.78 |
| Stacked Model | 0.681 | 4.02 | 2.66 |

### 6.2 Feature Importance

We now, using our predictive algorithm, list the feature importance

**Table 4:** Feature Importance

| Model | Random Forest | LGBM |
|---|---|---|
| AmbientTemp | 100 | 92.22 |
| Humidity | 57.89 | 100 |
| CloudCeiling | 51.96 | 38.25 |
| Pressure | 28.67 | 99.04 |
| sine_mon | 27.77 | 29.80 |
| Latitude | 22.41 | – |
| Windspeed | – | 59.77 |

## References

[1] J.; Heinemann D. In Weather Matters for Energy; Troccoli A. Dubus L. Haupt S. Eds. Lorenz, E.; Kühnert. Overview of irradiance and photovoltaic power prediction. In *Springer: New York*, 2014. 1

[2] M.; Ekanayake C. Raza, M.Q.; Nadarajah. On recent advances in PV output power forecast. In *Sol. Energy*, pages 125–134, 2016. 1

[3] V.; Ye Z.; Lim L.I.; Zhao L.; Aryaputera Yang, D.; Sharma. A.W. Forecasting of global horizontal irradiance by exponential smoothing, using decompositions. . In *Sol. Energy*, page 111–119, 2015. 1

[4] T.; Hurka J.; Heinemann D.; Kurz Lorenz, E.; Scheidsteger. On recent advances in PV output power forecast. In *Regional PV power prediction for improved grid integration.*, page 757–771, 2010. 1