Deterministic and Probabilistic Forecasting of the Global Horizontal Irradiance (GHI)

Yajat Pandya
MECH 6342 Report – Project#2

***Abstract*—A historical dataset for two years (2014-15) of solar energy parameters like the hourly GHI, Wind speed, Clear-sky GHI, Temperature, Pressure etc. of a particular location is provided. Using Python's *scikitlearn* package, a time series forecasting of the hourly GHI value for the year 2015 is conducted. This has been done for Deterministic and Probabilistic based forecasting methods. A Random Forest based algorithm has been used for the deterministic forecasting which has been compared with the Persistence of Cloudiness (POC) method of forecast. Support Vector Regression (SVR) model is used to predict the standard deviation values in case of probabilistic forecasts. Corresponding error metrics including MAE, RMSE and mean pinball error are calculated.**

***Index Terms*—GHI, Deterministic forecasting, SVR, MAE.**

I. Introduction

THE solar energy domain has been increasing its share of the annual renewable energy production across the globe. This is because of the advancements in semiconductors which have enabled higher solar efficiency in these systems. With the ever-increasing trends of it, the accurate forecast for solar energy becomes more and more important. However, multiple factors affect the Global Horizontal Irradiance (GHI), which is the defining factor of the amount of solar energy available at a location. It represents the total amount of shortwave radiation received from above by a surface which is horizontal (parallel) to the ground. GHI is the most important parameter for calculation of PV electricity yield. The incoming solar radiation from space is first exposed to the cloud cover at a certain location on Earth. Based on the optical properties at this location, the radiation is either absorbed, reflected, or transmitted in the lower atmosphere. About 51% of the transmitted radiation is then absorbed at the surface of the Earth and the rest is reflected back in the atmosphere. The GHI index represents the total radiation available at the location at any instant in time. The uncertain and variable nature of clouds imposes challenges on the grid integration of solar power, particularly at high penetration levels. Solar forecasting plays an important role in reducing the uncertainty of solar power output in operations. This can be useful at different time horizons, from day-ahead for unit commitment to minutes- and hours-ahead for economic dispatch. Probabilistic solar power forecasts provide even more information about the possible solar output, thus their inclusion directly in system operations is an active research area.

The remainder of the report is organized as follows. Section II describes the details of the historical dataset provided and the Persistence of Cloudiness (POC) method for a benchmark forecasting. Section III explains the modeling of deterministic, probabilistic forecasting method used for the 1-hour ahead GHI forecast, and the corresponding results. Lastly, the findings are concluded in Section IV.

II. Dataset and the Benchmark Forecasting Method

A. Dataset information

The given dataset contains hourly values of the variables mentioned in Table 1 below.

TABLE I
Dataset Variables

Quantity	Units
Timestamp	Hourly – mmddyyyy, hh:mm:ss
Diffuse Horizontal Irradiance (DHI) & Clearsky DHI	W/m^2
Direct Normal Irradiance (DNI) & Clearsky DNI	W/m^2
Global Horizontal Irradiance (GHI) & Clearsky GHI	W/m^2
Temperature	Degree Celsius
Pressure	kPa
Relative humidity	%
Solar zenith angle	Degrees above the horizon
Wind direction	Degrees w.r.t. local north
Wind speed	m/s

This hourly dataset is provided in the .csv format for the complete calendar year of 2014 and 2015. In general, the relationship between DHI, DNI and GHI is given as follows in equation (1).

$$GHI = DNI\cos(\theta) + DHI$$

These three quantities also have a version of values for a '*clear-sky'* condition when there is no cloud cover. Since the effective GHI depends on the cloud cover present at the location, the ratio of actual GHI and the *clear-sky* GHI is an important parameter to quantify the presence of clouds. Although the performance of typical solar panels do not depends on the ambient temperature, the wind velocity and the temperature may play a crucial role in the large scale convective processes happening in the local region. These processes alongwith humidity can create a meaningful feature to forecast cloud presence. The solar zenith angle determines the contribution of the DNI in the effecting GHI value. For different day times in extreme summer and winter, this value is expected to vary significantly. Thus, the timestamp alone may not be sufficient to quantify the contribution of the DNI.

B. Persistence of Cloudiness (POC) forecasting

As a benchmark forecast comparison, a Persistence of Cloudiness (POC) approach is implemented on the dataset, which assumes a constant clear-sky index within the forecasting time horizon. This is given as follows in equation (2).

$$GHI(t+1) = GHI(t) + SPI(t)[GHI_c(t+1) - GHI_c(t)]$$

Where, $GHI(t+1)$ is the GHI prediction at the next timestep and GHI_c is the clear-sky GHI value. The Solar Power Index (SPI) is the ratio of the actual and the clear-sky GHI value at any instant:

$$SPI(t) = \frac{GHI(t)}{GHI_c(t)}$$

As a result, the GHI forecast for every 1-hour was conducted using the dataset.

III. Forecasting Methodology

A. Deterministic Forecasting

In past, several solar forecasting methodologies have been developed. Majority of these methods particularly focused only on long-term forecasts, or day-ahead forecasts. A few methods like [1] used probabilistic methods to forecast for short-term periods. For the case of deterministic forecasting, a simple Machine Learning based algorithm has been used to predict 1-hr ahead forecasts for the year 2015.

A Random Forest regressor is implemented in Python's *scikitlearn* module using the *Pandas* library. For each time-step in 2015 (each hour), the data prior to that time-step was used to fit the regressor model. In training Machine Learning models, choosing the *right* features from the available dataset is critical and often requires a knowledge of the primary parameters that affect the output. For the scope of current study, the following features mentioned in Table 2 were chosen to train the model.

TABLE 2
Random Forest Regressor Input Features

Feature (previous timestamp)
Diffuse Horizontal Irradiance (DHI)
Day of the year
Hour of the day
Solar Zenith Angle

Here, other parameters like the Wind speed, direction, Relative Humidity etc. are omitted for the sake of simplicity and due to computational limitations. Consequently, two approaches were carried to use the training data:

1. RF1: All of 2014 data was used to train the model.
2. RF2: All of data before each target timestamp in 2015 was used to train the model.

The RF models were configured with default hyperparameters and with 100 number of decision trees. Thus, along with the POC method, a total of three approaches were carried for the comparison of deterministic forecasts. Figure 1 shows the time signal of the actual GHI value over the course of two years.

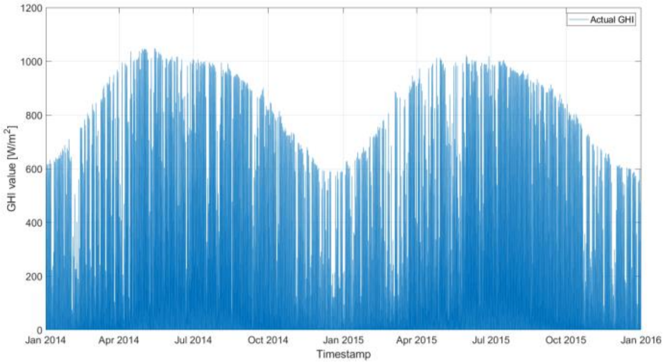

Fig. 1. Actual GHI values over the duration of two years.

Similarly, the forecasted GHI values from all three methods are also shown in figure 2 below. It is important to note that the peak of the GHI values happen between the months of May and July which is typical for any location in the northern hemisphere. There are significant number of days with very low GHI values. This implies the frequent happening of precipitation in the surrounding area.

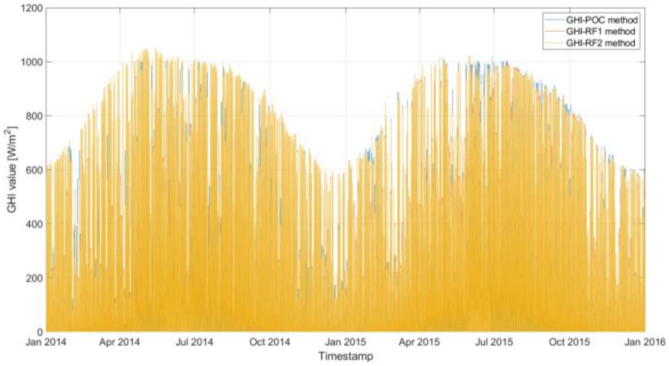

Fig. 2. 1-hr ahead forecasted GHI values over the duration of two years using all three methods – POC, RF1 and RF2.

In Figure 3, the 6-hr moving median of the absolute error in the GHI forecasts is shown for the year 2015. It is important to note that the peak of the error values occur during the summer months. This could be because the training data for one year may not be sufficient for the model to capture the annual trends in the atmospheric processes. It is also worth noting that the RF2 method does not have a decreasing error trend as time progresses in 2015, which would be expected of it as it trains on the most recent data available.

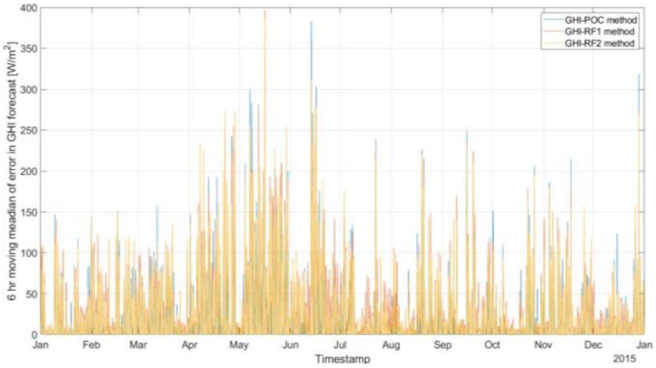

Fig. 3. 6-hr moving median of the absolute error in GHI forecasts for the year 2015 using all three methods – POC, RF1 and RF2.

Along with this, the % error PDF of the RF1 was also studied. In Figure 4., the % error PDF is plotted for the 2015 RF1 deterministic predictions for (a). All dataset and (b). Only day-time conditions. It is worth noting that the increment of variance in day-time condition proves that the overall dataset consists of significant proportion of night-time conditions when the observed and forecast GHI values are close to zero. These data may need to be filtered out for a true and rigorous analysis of the models used in this study.

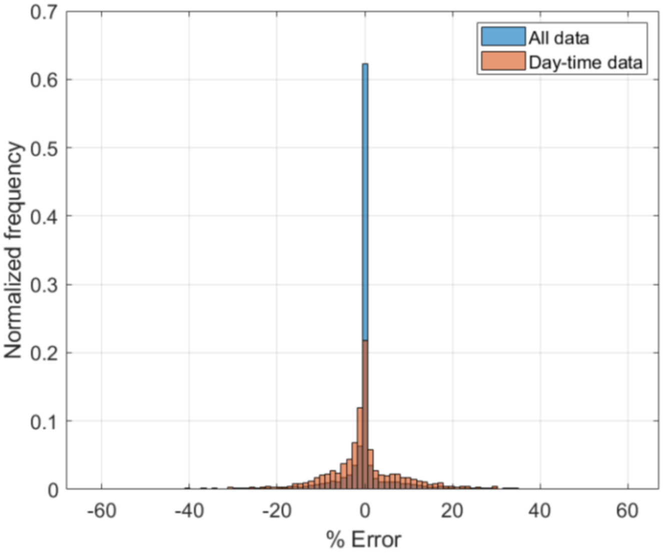

Fig. 4. % Error PDF for the RF1 deterministic forecasts for all dataset and for only day-time dataset of 2015 hourly forecasts.

A better way to quantify the error in the forecast is using Mean Absolute Error (*MAE*), normalized Mean Absolute Error (*nMAE*) and the normalized Root Mean Square Error (*nRMSE*). The normalization is done using the maximum value of the actual GHI value over the year 2015. The definitions of these quantities are described below.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|$$

$$nMAE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{f_i - y_i}{y_{max}}\right|$$

$$nRMSE = \frac{1}{y_{max}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2}$$

Where, f_i is the forecasted variable, y_i is the actual GHI value, y_{max} is the actual maximum of the GHI in 2015 and n is the number of samples (timestamps). This was calculated for all three methods and the results are shown in Table 3. Notably, the RF1 and RF2 methods do not have significant different in the forecasting accuracy. This implies that the models do not lack in the number of samples available, but rather in the features of training data. Also, the POC method has the least error values in all three domains. A better model would consist of an ensemble average of all three models based on their performance during certain range of features, e.g. Cloud cover.

TABLE 3
FORECASTING ACCURACY FOR GHI VALUES

	POC	RF1	RF2
MAE [W/m^2]	24.52	29.59	28.70
nMAE	2.33 %	2.82 %	2.73 %
nRMSE	6.67 %	6.79 %	6.73 %

B. *Probabilistic Forecasting*

To estimate the probabilistic forecasts of the GHI values, a Gaussian kernel is used to predict the probabilistic range at each hour in 2015. The Gaussian function has $f\left(\frac{x}{\mu},\sigma\right)$ two input parameters – the mean of the distribution (μ) and the standard deviation (σ). The Cumulative Distribution Function of the Gaussian function is shown below.

$$F(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sigma}\right)\right]$$

At each time stamp, the mean μ is the deterministic forecast value of the GHI around which the probabilistic forecasting was conducted. The σ was optimized at each time stamp using the Pin-ball loss minimization method in MATLAB. At any given time stamp, the pinball loss was calculated using the following method:

$$L_m(q_m, x_i) = \begin{cases} \left(1 - \frac{m}{100}\right)(q_m - x_i)\ for\ x_i < q_i \\ \frac{m}{100}(x_i - q_m)\ for\ x_i \geq q_i \end{cases}$$

Where q_m is the mth percentage predicted quantile using the Gaussian distribution and x_i is the actual observation of GHI value at that time stamp. Thus, at each time stamp in the training dataset (2014), the following optimization was performed to estimate the optimal σ:

$$\min_{\sigma} \sum_{m=1}^{99} L[q_m(\sigma), x_i]$$
$$sub.to: \sigma_l \leq \sigma \leq \sigma_u$$

Where $\sigma_l = 0$ and $\sigma_u = 200 \frac{W}{m^2}$ were set as the bounds. Here, MATLAB's *fmincon* function was used to get the optimal σ in the training data of 2014. A surrogate model of $\sigma = f(x_p)$ was created using Support Vector Regression (SVR) on the training data of 2014, where x_p is the input vector of the features at each time stamp. This surrogate model was then used to predict the σ values at 1-hr ahead timestamps of 2015, thereby resulting in the probabilistic forecasts. The features used here are the same as those for the Random Forest algorithm to predict the deterministic forecasts, as in Table 2. The histogram of the predicted σ values are shown below in Figure 5. As expected, the standard deviation of day-time forecasts is significantly higher than that of all datapoints, which comprise large number of night-time conditions.

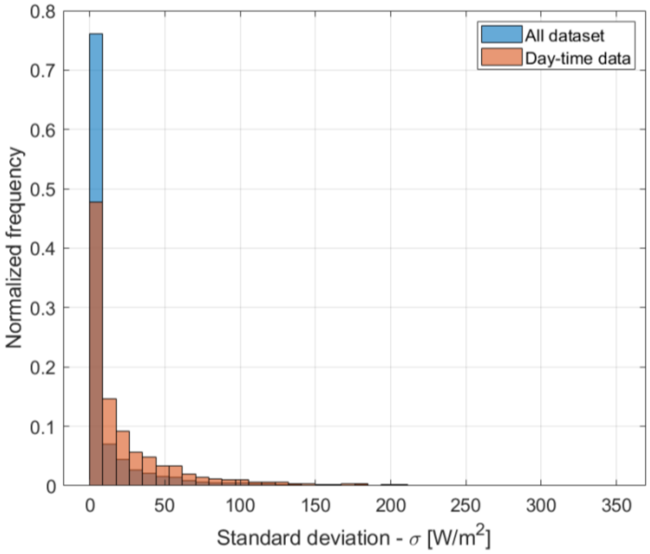

Fig. 5. Histogram of predicted σ values for 2015 hourly forecasts using SVR trained from 2014 dataset – All dataset and day-time conditions.

It is important to note that most of the occurrences of σ happen near the zero values, which implies (a). Large proportion of overfitting the dataset and (b). Large proportion of night-time conditions with close to zero GHI values of observations and forecasts. This could be because the annual variation is captured only once is the training data and other random variables could be possibly overpowered by it in the SVR training phase. A sample time-signal of the probabilistic forecast for 2015 is also shown below in Figure 6. Notably, the night-time GHI forecasts effectively are 100% accurate with no probability intervals.

The sum of pinball loss is averaged over all quantiles from 1% to 99% and normalized by the maximum observed GHI value in 2015. The pinball loss is thus calculated to be equal to **1.36 units**. The corresponding standard deviation however was calculated to be equal to **3.06 units**. This is dramatically low mean value with approximately three times higher standard deviation. This was realized to be primarily due to the presence of night-time conditions in the dataset when the observations and the forecast GHI values were close to zero. Thus, a filter of day-time-only conditions was applied in this calculation. As a result, the day-time dataset resulted in an average normalized pinball loss value of **3.01 units** with a standard deviation of **4.22 units**. This is the true measure of the accuracy of the probabilistic forecasts. These are also mentioned below in Table 4.

TABLE 4
NORMALIZED ERROR METRICS OF PROBABILISTIC FORECASTS

	All data	**Day-time data**
Mean of pinball error	1.36	3.01
Std. dev. (σ) of pinball error	3.06	4.22

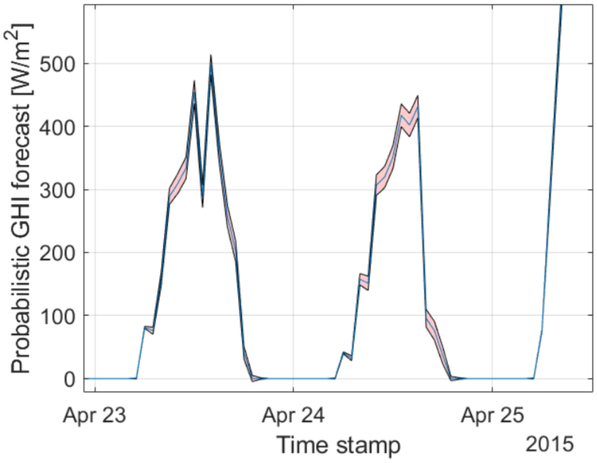

Fig. 6. 1-hr ahead probabilistic forecast of the GHI values in a sample 2015 duration. Blue line indicates the observed GHI values.

IV. CONCLUSION

Using the provided dataset with the actual and clear-sky GHI values for two years, a set of 1-hr ahead GHI forecasts for deterministic and probabilistic methods was modeled in Python's *scikitlearn* library. Firstly, the persistence-of-cloudiness method is used as a benchmark for the deterministic forecasts. Two methods (RF1 and RF2) of deterministic forecasts are utilized using Random Forest algorithm. These include training the RF model using only 2014 dataset and training the model using every data point available before the target prediction timestamp. There is not significant change in the accuracy of both the methods, which mean that the 2014 dataset has more than sufficient points to train from. Furthermore, a distinct difference in error was noted during daytime and night-time conditions which was realized to happen because of close to zero GHI values during the night for both the observations and the forecasts. Using the RF1 deterministic forecast, a pin-ball optimization approach was utilized to predict the probabilistic forecasts for 1-hr ahead in 2015 time stamps. The SVR model was used to create a surrogate formulation of the standard deviation of Gaussian (Normal) distribution at each time stamp GHI. Because of abovementioned differences in night-time and daytime conditions, the average normalized pin-ball error in daytime conditions was more than 2.5 times more than the total average error. Thus, various forecasting methods taught in the course were analyzed and implemented during this project.

Acknowledgment

Yajat acknowledges the MECH 6342 course instructor Dr. Jie Zhang for his continuous support and guidance during this project work.

References

[1]. P. Pinson, *Very-short-term probabilistic forecasting of wind power with generalized logit–normal distributions*, Journal of the Royal Statistical Society: Series C (Applied Statistics) 61 (4) (2012) 555–576.

[2]. M. Evans, N. Hastings, and B. Peacock, “*Statistical distributions*,” 2000.

[3]. I. Steinwart, A. Christmann et al., “*Estimating conditional quantiles with the help of the pinball loss,*” Bernoulli, vol. 17, no. 1, pp. 211–225, 2011.

[4]. J. Juban, N. Siebert, and G. N. Kariniotakis, “*Probabilistic shortterm wind power forecasting for the optimal management of wind generation*,” in Power Tech, 2007 IEEE Lausanne. IEEE, 2007, pp. 683–688.

[5]. Feng, C., Cui, M., Hodge, B.-M., Lu, S., Hamann, H., & Zhang, J. (2018). *Unsupervised Clustering-Based Short-Term Solar Forecasting. IEEE Transactions on Sustainable Energy*.