Blue Skies: A Methodology for Data-Driven Clear Sky Modelling

Kartik Palani[*], **Ramachandra Kota**[+], **Amar Prakash Azad**[+], **Vijay Arya**[+]
[*]University of Illinois at Urbana-Champaign, USA [+]IBM Research, India
palani2@illinois.edu, {rama.chandra, amarazad, vijay.arya}@in.ibm.com

Abstract

One of the major challenges confronting the widespread adoption of solar energy is the uncertainty of production. The energy generated by photovoltaic systems is a function of the received solar irradiance which varies due to atmospheric and weather conditions. A key component required for forecasting irradiance accurately is the *clear sky model* which estimates the average irradiance at a location at a given time in the absence of clouds. Current methods for modelling clear sky irradiance are either inaccurate or require extensive atmospheric data, which tends to vary with location and is often unavailable. In this paper, we present a data-driven methodology, Blue Skies, for modelling clear sky irradiance solely based on historical irradiance measurements. Using machine learning, Blue Skies is able to generate clear sky models that are more accurate spatio-temporally compared to the state of the art, reducing errors by almost 50%.

1 Introduction

Solar energy is gaining prominence across the world in response to climate change, depleting reserves, lowering costs, and favorable legislation. However, a key challenge in integrating solar power into the grid is the ability to forecast solar energy production accurately. The variability and uncertainty of solar power can negatively impact grid stability and increase the cycling costs of conventional power plants.

The energy generated by photovoltaic or solar thermal systems is a function of the received irradiance, which is the power density incident on a surface due to illumination from the sun, and measured in Watt/m^2. While the extraterrestrial irradiance above earth's atmosphere varies deterministically as a sinusoidal function, the radiation incident on solar panels varies stochastically as it is affected by both atmospheric and weather conditions such as aerosols, gases, and cloud cover. The expected solar irradiance on a given day can be decomposed into the clear sky irradiance minus any losses due to the presence of clouds. Clear sky irradiance is defined as the radiation received at a site under cloud-free conditions and is given by a *clear sky model* [Reno *et al.*, 2012]. Therefore, to predict the solar irradiance at a location, in addition to weather forecasts, one requires an accurate clear sky model appropriate for that location. Any errors or bias in the clear sky models translate into errors in the total irradiance forecast.

Additionally, clear sky models play an important role in siting solar installations and estimating peak loads induced on the power grid due to HVACs. Similarly, clear sky models are also used to estimate the impact of solar radiation on the quality of surface water bodies and in agriculture planning to estimate the amount of evapotranspiration expected at a given location [Chameides *et al.*, 1999].

The clear sky model is a physical model that computes the irradiance at a site at a given time and day of the year under a cloudless sky. Even on a clear day, the solar radiation percolates through the various layers of the atmosphere and is absorbed, scattered, and reflected by the composition of gases, aerosol content, water vapor, and other particulate matter in the atmosphere. A good model should therefore capture the attenuation due to these factors and local nuances in order to provide accurate clear sky irradiance estimates for a given location. However, the problem remains challenging due to the complexities involved in estimating clear sky radiation. Against this background, this work proposes a data-driven methodology for automatically building accurate and localised clear sky models using machine learning. Such a framework is especially relevant given the widespread adoption of solar photovoltaic systems. Our methodology overcomes several drawbacks of the existing approaches and clear sky models. Next, we provide a brief background on clear sky modelling and highlight the aspects that are not captured by current approaches.

1.1 Background

The sun is commonly modelled as an ideal black body. Therefore its radiation intensity H_0 at a distance D is given by $H_0 = \frac{R^2}{D^2}\sigma T^4$, where σ is the Stefan Boltzmann constant and R and T are the radius and surface temperature of the sun, respectively [Southworth, 1945]. Thus the *extraterrestrial irradiance* or *top of atmosphere irradiance* at a location, on a given day varies due to the axial tilt and elliptical orbit of the earth. This is the maximum irradiation that is available to the location on that day.

While the extraterrestrial irradiance is easily described using physics based models, the radiation falling on the earth's surface is highly variant due to atmospheric effects such as absorption and scattering, in addition to local variations such as water vapor and pollution. Different entities in the atmosphere affect the incoming irradiation differently: while atmospheric gases cause Rayleigh scattering [Bucholtz, 1995], atmospheric aerosols and dust particles cause Mie scattering [McCartney, 1976]. As a result, the irradiance reaching the earth's surface has a 'direct' (Direct Normal Irradiance, *DNI*) and a 'diffused' (Diffused Horizontal Irradiance, *DHI*) component. The total irradiation incident on the ground (Global Horizontal Irradiance, *GHI*) is a function of these two terms. GHI is easily measured using a pyranometer (we use such data in our approach). For a given location, GHI varies during the day as the solar position varies. The solar position is parametrized by the solar zenith and azimuth angles. GHI is also a function of the air mass, which is the length of atmosphere that the solar rays have to pass through. While there are complex models to evaluate air mass at a location [Bdescu, 1987; Gueymard, 1993; Kasten and Young, 1989], it is often approximated to be the secant of the zenith angle.

$$x = 1/cos\theta_z \tag{1}$$

This dependency of the surface irradiance on the atmospheric conditions even on a clear day, implies that any clear sky model will have a strong dependency on these atmospheric parameters, which in turn are affected by locational and seasonal variations. For example, a desert region will have a different local atmospheric composition compared to a tropical rain forest. Furthermore, the same location will have varying atmospheric compositions such as aerosol or dust content from one season to another.

Current approaches in clear sky modelling attempt to capture these atmospheric conditions via parameters in the model. Unfortunately, these methods either take a simple 'one-fits-all' approach or are incomplete and expect the user to fill in the requisite parameter values. In the simple methods, default parameter values are provided for use with the model. However, these do not capture local variations and result in inaccurate irradiance estimates. In contrast, the complex models, requiring extensive parameter values, are impractical because the atmospheric parameters are not readily available for most locations, especially in the developing world. In addition to generalizing across vastly different locations, existing models also fail to capture the seasonal variations in clear sky irradiance. Thus, most existing models are oblivious to spatio-temoparal variations.

To address these problems, we propose a methodology for building location and season-aware clear sky models called 'Blue Skies'. Based on machine learning techniques, Blue Skies takes in historical measurements of irradiance (GHI) for a location to build an accurate clear sky model for that location. Since GHI values recorded by pyranometers consist of both clear and cloudy periods, this process involves automatically retrieving clear sky periods from such a dataset. This clear sky dataset is then used to *learn* location specific and seasonal model parameters, thereby making the models spatio-temporally aware.

Our contribution, thus, is a comprehensive data-driven clear sky modelling methodology. The need for Blue Skies methodology and its performance is validated by evaluating on real datasets captured from three vastly different locations. The results show that clear sky models produced by Blue Skies perform considerably better than any of the current state of the art simple or complex clear sky models. At the same time, as opposed to complex models which require measurements of exogenous variables, Blue Skies requires only past irradiance measurements from the vicinity of the location, which are easy to obtain from nearby weather stations or an installed pyranometer. In the next section, we study the existing clear sky models. Section 3 describes our method. Following that, Section 4 details our experimental setup and results. Finally, Section 5 concludes the paper.

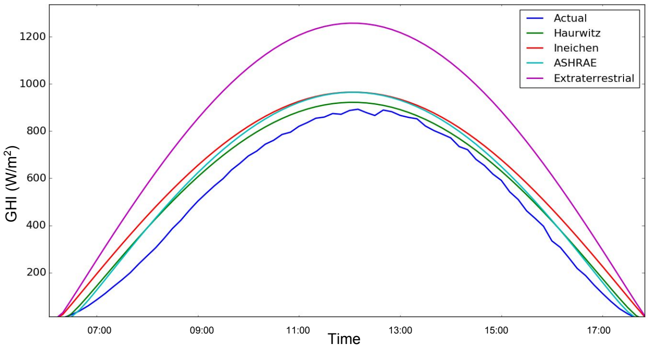

Figure 1: Performance of clear sky models in Bangalore.

2 Related Work

A clear sky model essentially involves geometric calculations that relate the irradiance incident at the location and estimates of atmospheric parameters that attenuate the incident power. Simple clear sky models are a function of extraterrestrial irradiance, the zenith angle and a few atmospheric parameters that generalize the atmospheric state for the location [Daneshyar, 1978; Kasten and Czeplak, 1980; Haurwitz, 1945; Badescu, 2008]. Most of them differ only in the ways they model and estimate the atmospheric parameters.

The Haurwitz model [Haurwitz, 1945] being one of the simplest models, is widely used due to its dependence only on the zenith angle. Models proposed by Kasten [1980] and Ineichen and Perez [2002] account for atmospheric effects along with the zenith angle. While Kasten's model uses air mass and the Linke turbidity factor (measure of turbidity of atmosphere), it is further improved by Ineichen and Perez by adding correction factors to the atmospheric parameter.

At the cost of adding complexity, improvements to simple models considered the effects of individual atmospheric components such as ozone, aerosol and perceptible water [Davies and McKay, 1989; 1982; Gueymard, 2008]. Another complex but well studied model is proposed by Bird [1984], which includes transmittance due to various atmospheric factors such as Rayleigh scattering, aerosol attenuation, mixed gas absorption, etc. Though these models provide accurate estimates of clear sky irradiance, they require a lot of inputs that are often not readily available. Moreover, methods for suitably estimating parameters for these models using external sensor and meteorological data is a tough process fraught with inherent errors.

The ASHRAE clear sky model [ASHRAE, 1979] reduces the dependency on many atmospheric measurements by providing a parameter look-up table for monthly values at a few locations. However, the model is not robust for locations where this table is unavailable. Fig. 1 shows the performance of the commonly used clear sky models in Bangalore, India compared against the actual measured irradiance on a particular clear sky day and the extraterrestrial irradiance. We see that none of the models closely match that actual clear sky trend of Bangalore. It is also interesting to note, from the figure, that the complex models (ASHRAE and Ineichen) perform worse than the simple Haurwitz model. Since both the complex models use globally averaged measurements as input parameters, for a given location they may perform better or worse depending on how closely the global averages reflect local atmospheric conditions.

In contrast to the above, a data-driven model is proposed by Grigiante *et al.* [2011] where measured irradiance of clear days are fitted with Bird's model. The detection of clear sky, however, is mainly done by visual inspection. NREL's Sunny Days [Long and Ackermani, 2000] is another similar approach where the coefficients of a base model are found for each clear day and parameters for other days are obtained by interpolating the coefficients between clear days. Both these methods depend on high resolution (spatially & temporally) satellite image data which is not available for most locations. Compared to all the above approaches, our methodology is not dependent on any exogenous parameters or satellite data and yet is capable of capturing variations due to location and seasons.

3 Blue Skies Methodology

In this section, we describe the three components of our methodology. First, we derive a generalised physical model that describes the irradiance received at a location, which we call the *base model*. The base model contains certain parameters which are dependent on the location. In the second stage, we present an algorithm to classify historical irradiance measurements received from a location into clear sky periods (Section 3.2). Finally, in the third stage, based on the classified clear sky dataset, learning approaches are used to determine the model parameters, thereby producing a customized model for a given location (Section 3.3).

3.1 Stage 1: Base Model

As seen in Section 2, a detailed analysis of existing clear sky models reveals that most models depend on the following variables: extraterrestrial irradiance, zenith angle, air mass, and a measure of atmospheric turbidity at the specified location. More elaborate models require multiple exogenous variables such as aerosol optical depth, water vapor content in lower and upper atmosphere, pollution levels, and so on.

In order to arrive at a generalized physical model of surface irradiance, we look at the Beer-Lambert law which describes the attenuation of a direct beam of light as it travels through fluids. The attenuation of light in a fluid is not only due to absorption but also due to scattering. The reduction of intensity due to both these effects is termed as *extinction*. The Beer-Lambert law states that an intensity I_0 attenuates exponentially with distance x based on the extinction coefficient β as:

$$I(x) = I_0 e^{-\beta x} \tag{2}$$

The earth's atmosphere is known to have many impurities. Most complex models attempt to quantify the effect of each of these impurities on the solar irradiation by having a different coefficient β_i for each impurity. However, capturing all the impurities in this way is a highly arduous and imprecise task. Instead, our model uses the average extinction coefficient β which we learn separately for each location based on data. Additionally, some processes, such as absorption by water vapor, do not follow the Beer-Lambert law [Elder and Strong, 1953]. For these, *transmittance* is defined as the ability of a medium to allow radiant energy to pass through and given by a clearness number:

$$C_n = \frac{I_{water}}{I_{standard}} \tag{3}$$

where I_{water} and $I_{standard}$ are the irradiance received in the presence of water vapor in the atmosphere and under standard atmospheric conditions respectively [Brutsaert, 1975].

The above equations, along with the knowledge of the extraterrestrial irradiance E and Eq. 1, allows us to calculate the direct normal irradiance falling on a surface perpendicular to the sun's rays as:

$$I_{ND} = EC_n e^{\frac{-\beta}{cos\theta_z}} \tag{4}$$

where θ_z is the zenith angle. The direct irradiance on a horizontal surface is then just the projection.

$$I_{HD} = EC_n cos\theta_z e^{\frac{-\beta}{cos\theta_z}} \tag{5}$$

In addition to direct irradiance, there exists a diffused component contributing to the overall irradiance. Previous work [Liu and Jordan, 1960] has shown that the effective method to obtain the diffused horizontal irradiation I_{diff}, is simply scaling I_{ND} by a factor C, which is the ratio of intensity of solar irradiation incident on a horizontal versus normal surface outside the atmosphere.

$$I_{diff} = CI_{ND} \tag{6}$$

The global horizontal irradiance is then a sum of these values

$$\text{GHI} = I_{HD} + I_{diff} \tag{7}$$

In summary, the global horizontal irradiance received at the earth's surface at a certain solar zenith angle θ_z, when the extraterrestrial irradiance E gets attenuated by a certain turbid atmosphere (with clearness number C_n and mean extinction coefficient β) and air mass (Eq. 1), is given as:

$$\text{GHI} = EC_n(cos\theta_z + C)e^{\frac{-\beta}{cos\theta_z}} \tag{8}$$

This base model is syntactically similar to some of the models studied in Section 2. We learn the coefficients (C, C_n, β) specific to a location (and/or time) using a clear sky dataset for that location. Next, we describe how we obtain such a dataset from historical irradiance measurements.

3.2 Stage 2: Generating Clear Sky Dataset

Typical low-cost pyranometers (located at weather stations, solar plants, and so on) provide time-indexed measurements of GHI at resolutions of 10 seconds, 1 minute, 10 minutes, etc. Given any such pyranometer data, the challenge is to label clear sky data points within it in order to generate a 'clear sky dataset'. Towards this, we built an automated process using a 2-step filtering approach. The goal of the filters is to weed out all instances of cloud cover and produce a refined clear sky dataset. The intuition behind the filters originates from two key observations:

1. All clear sky days have a smooth diurnal curve whose trend matches that of the extraterrestrial irradiation received that day. Fig. 1 illustrates this trend.
2. All smooth curves needn't correspond to clear sky periods. Uniformly cloudy days may also generate smooth irradiation curves which match the trend of the extraterrestrial curve but with significantly lower magnitude.

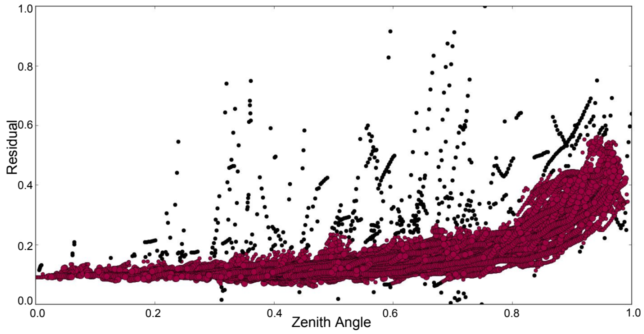

Figure 2: Scatter plot of normalized residuals against normalized zenith angles. The black points are characterized as outliers.

Moreover, pyranometer data is often sparse with missing values which reduces the possibility of finding completely clear sky days. Locations which are cloudy for most part especially face this problem. Therefore, we employ a sliding window based approach in order to identify and tag clear sky periods in the data rather than clear days. Our approach is not dependent on the size of this window. The size can be varied to best suit the availability of data.

Step 1 - The Correlation Filter:

The first filter uses a correlation metric to take advantage of the intuition that the trend of the GHI on clear sky periods matches that of the extraterrestrial irradiance. We pass a rolling window over the measured GHI data and correlate the data points with the corresponding timestamped values in the extraterrestrial data. A perfectly matching trend within the window presents a correlation of 1. In contrast, a cloudy period gives a negative correlation due to the drop in irradiance received on the earth (visually appearing as a negative spike in the pattern). For every window period, we use the Pearson correlation coefficient to calculate the correlation. Although rank based correlation may perform well for each half of the complete diurnal curve (due to the monotonic relationship), within small window sizes like in our case, the data points have a linear relationship thereby making Pearson correlation a better choice.

In this way, a period with negative correlation implies the existence of cloud cover or non-clear sky, thereby allowing us to remove such periods from the dataset. The periods with positive correlation, however, needn't necessarily be a clear sky (potentially uniformly cloudy). This is because the correlation filter does not take magnitude into account. For this, the positively correlated data points are passed through the second filter.

Step 2 - The Clustering Filter:

Although, measurements from fully cloudy periods also generate smooth curves having a high correlation with extraterrestrial irradiance, they have significantly lower magnitude. Such cloudy periods are distilled out by the second filter. A naive filtering approach would be to set a threshold for the magnitude of the measured GHI. However, this threshold will need to be time-variant as the clear period at 0700 hours will have much lower solar intensity than a clear period at 1200 hours. The magnitude also varies across seasons. These effects are seen even on the differences in magnitudes between GHI and extraterrestrial. To avoid these problems, we propose an automated unsupervised filtering method.

Instead of a time-series, we transform our instance space to solar angles. Since time is a construct of the position of the earth relative to the sun, it is possible to describe the time-series data in terms of the zenith and azimuthal angles. A tuple of zenith and azimuthal angle (θ_z, θ_a) represents the time and day in a year. This transformation is useful because solar intensity depends on the solar angles rather than the clock-time. Now, just like earlier, for each window, we compute the difference in magnitude between the measured GHI and corresponding extraterrestrial irradiation (hereby referred to as the residual). This difference can be computed between the mean values in the time window, or between the corresponding min or max values (discussed further in Section 4.1). Fig. 2 shows the scatter plot of the mean residuals against the zenith angles of a year for a particular location. The zenith angle is 0 degrees during the solar noon (directy overhead) while it is 90 degrees at sunrise and sunset. The region of high density in the plot captures the trend of the variation of the mean residual over day and seasons. The noise or outliers in the scatter plot represent instances of cloud cover or non-clear periods. Thus, the dense nature of the clear sky data points allows us to automate the process of grouping them into a single large cluster while removing the outliers (cloudy periods).

Towards this, we use a well-known density based clustering algorithm, DBSCAN [Ester *et al.*, 1996]. Given two inputs, ϵ and *minpts*, DBSCAN defines an ϵ-neighborhood for point x as: $N_\epsilon(x) = \{y \in X | d(x,y) \leq \epsilon\}$. Core points are points which have more than a minimum number of points, *minpts*, in their neighborhood. A point y is then considered *density reachable* from a core point x, if there exist a finite sequence of core points between x and y where each such point belongs to the ϵ-neighborhood of the previous point. Then, every point that is reachable from core points is factored into a maximally connected cluster. DBSCAN's flexibility in generating arbitrary shaped clusters allows us to capture the clear sky data points into a single cluster, with cloudy periods being characterized as outliers or noise.

3.3 Stage 3: Learning Clear Sky Model

The base model for clear sky derived in section 3.1 yields the global horizontal irradiance at time t as

$$GHI(t) = E(t)C_n(cos\theta_z(t) + C)e^{\frac{-\beta}{cos\theta_z(t)}} \qquad (9)$$

The value for extraterrestrial irradiance $E(t)$ is obtained from NREL's Solar Position Algorithm [Reda and Andreas, 2004], which also computes the zenith $\theta_z(t)$ and azimuth $\theta_a(t)$ with a precision of $\pm 0.0003°$, and is used as a standard in all weather calculations. This leaves the clearness number C_n, the diffusion coefficient C, and the atmospheric extinction coefficient β, as unknowns in the model. As these coefficients tend to depend on the location, we *learn* these unique values from the clear sky dataset generated for the corresponding location. In particular, let s denote the set of time steps that are classified as clear sky conditions by the algorithm in the previous section. Then, given irradiance measurements $\{GHI(t), t \in s\}$, the parameters C, C_n, and β are regressed by minimizing the L_2 norm of the error:

$$(\hat{C},\hat{C}_n,\hat{\beta}) = \operatorname*{argmin}_{(C,C_n,\beta)} \sum_{t\in s} \Big(GHI(t) - E(t)C_n(cos\theta_z(t)+C)e^{\frac{-\beta}{cos\theta_z(t)}}\Big)^2 \qquad (10)$$

The function is convex in each variable, which indicates that the minimization does not yield local minimas. This was also empirically verified. L2-minimization is obtained using Levenberg-Marquardt algorithm.

The parameters C, C_n, and β at a location may also be affected by diurnal and seasonal variations which should be captured by an accurate clear sky model. In order to determine the best model, we use the following *learners*:

1. **Basic**: The simplest clear sky model for a location can be obtained by learning a single tuple (C,C_n,β) corresponding to all clear sky irradiance measurements. Hence, this learner does not capture any seasonality.
2. **Seasonal**: More detailed model would be obtained by dividing clear sky dataset by seasons and learning n tuples, (C,C_n,β), for n seasons experienced at the location.
3. **Azimuthal**: Since (θ_z, θ_a) give a construct for time, it makes intuitive sense to learn (C, C_n, β) for different ranges of azimuthal angles to capture temporal variations.
4. **Hourly**: Instead of azimuthal transformation, the clear sky time series could be used to learn (C,C_n,β) for the different hours of the day. Considering a 12 hour daylight period, this would result in 12 tuples of (C,C_n,β).

Also, the Seasonal learner could be combined with either Azimuthal or Hourly to create **Seasonal-Azimuthal** and **Seasonal-Hourly** learners respectively. For example, given n seasons and 12 hours, Seasonal-Hourly will have $n\times 12$ tuples of (C,C_n,β). We empirically evaluate all these learners to determine their performance on given set of locations.

4 Experiments

In order to empirically validate our methodology, we conducted experiments and evaluated the performance of both the components of our approach: (i) generating clear sky dataset and (ii) learning clear sky model. For evaluation, our irradiance dataset consisted of GHI pyranometer measurements at 1-minute resolution from three different locations. The locations were picked in different latitudinal zones in order to understand the clear sky patterns in various regions:

1. **Tucson**, Arizona, US [Andreas and Wilcox, 2010]: This mid-latitude location is characterized by a desert climate and has two major seasons, summer and winter with some monsoon showers in Jul and Aug (NOAA). The dataset is made up of 7 years from Nov 2010 to Aug 2016.
2. **Bangalore**, India: This tropical region has a tropical savanna climate which is characterized by distinct dry and wet periods. It is also at a higher altitude (900m) and is affected by both the northeast and southwest monsoons with Aug and Sept having 13 rainy days on average (Jensen). Our dataset covers a 1 year period from Jan 2016 to Dec 2016.
3. **Seria**, Brunei: This equatorial location is characterized by significant rainfall across the year. Even the driest month (Feb) has, on average, 14 rainy days and Oct and Nov have 25 rainy days on average (WWO). Our dataset spans 3 years from Jan 2012 to Dec 2014.

Table 1: Performance of the dataset generation algorithm

Algorithm	Precision	Recall
Mean-based residuals	91	86
Min/Max-based residuals	96	84

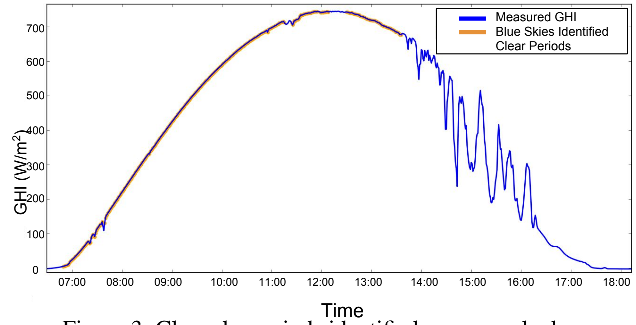

Figure 3: Clear sky periods identified on a sample day.

Finally, to evaluate our approach, we utilize NREL's National Solar Radiation Database (NSRDB) [Wilcox, 2007]. The dataset for Tucson, created by processing satellite images, classifies cloud cover with labels from 0-12, where each integer represents a type of cloud and 0 is clear sky. The dataset also contains the clear sky GHI for Tucson based on the REST2 model [Gueymard, 2008].

4.1 Evaluation: The Dataset Generation Process

A clear sky dataset contains timestamps and corresponding GHI measures of time instances when the sky was cloud free. A good clear sky dataset, thus, must contain very few false positives (cloudy periods being classified as clear periods). In the dataset generation process, we proposed two methods for computing the residuals (used to determine whether a time window should be classified as clear sky): one is based on mean values and other is based on min/max values of irradiance in the window. To evaluate the two methods, we compared our generated clear sky dataset with the NSRDB database. While our dataset has 1-minute resolution, the NSRDB database has a snapshot every 30 minutes. Therefore, we compared the common timestamps, that is the value at every 30^{th} minute.

Table 1 shows the results for Tucson. Here, *recall* gives the percentage of clear sky data points in NSRDB that were identified by our methods. *Precision* gives the percentage of actual clear sky data points (as given by NSRDB) in our generated clear sky datasets. We see that while both methods have similar recall, min/max based method has much higher precision. This is because min/max residual method makes the anomalies more extreme thereby allowing DBSCAN to easily classify them as outliers. For creating a clear sky dataset, the method with higher precision must be preferred even if it has slightly lower recall as long as the recall is sufficient to produce a large enough dataset. As that is the case with the min/max residual method, we propose that this method be used as part of the Blue Skies methodology. While clear sky data was available only for Tucson for evaluation, Fig. 3 gives a glimpse of how well our method identifies clear sky periods for a given day in the Bangalore location (similar results were seen for Seria).

4.2 Evaluation: The Clear Sky Learning Methods

In Section 3.3, we proposed several learning methods that can be used to determine the parameters (C,C_n,β) from the clear sky dataset. For our experiments, we split the dataset into train-test parts in the following manner. For Tucson and Seria, where the dataset spans several years, some years were used for training and others for test. For Tucson, 2011, 2012, 2013 & 2015 were in the training set and 2010, 2014 & 2016 in the test set. For Seria, 2012 & 2014 were used for training and 2013 for test. Since the Bangalore dataset only spanned 12 months, 80% of the days in each month (24 days) were used for training and rest 20% for test. All results are presented in the form of RMSE values with normalized RMSE (nRMSE) in paranthesis. We use the following formula for nRMSE:

$$\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}}{mean(Y_i)} \times 100 \qquad (11)$$

where, $\hat{Y}_i$ are the predicted values of GHI and Y_i are the actual values of GHI for n such GHI values.

Table 2 compares the performance of three current state-of-the-art clear sky models – ASHRAE, Haurwitz and Ineichen with our methodology, named Blue Skies-Basic (BS-Basic) as it uses the Basic learner described in Section 3.3. For comparison, we also included BS-Ineichen, which is a hybrid of Blue Skies with Ineichen (most popular among the current models). BS-Ineichen uses Ineichen as the base model. However, instead of employing the default parameter values provided by Ineichen model, it learns them in a data-driven manner like in our approach, by following stage-2 and stage-3 of our methodology.

While it is clear that the customised data-driven clear sky models, BS-Ineichen and BS-Basic, perform considerably better than the standard models for all the three locations we make a few interesting observations. As described before, most current models are dependant on detailed sensor measurements, and in locations where such data is available these models perform well. Ineichen, for e.g. does better than ASHRAE and Haurwitz in Tucson due to the rich sensor data available for the region. Another spatial trend we note is that unlike other models whose performance falls drastically when moving from mid-latitudinal Tucson desert to tropical regions which are harder to model, Blue Skies has a more gradual decline. The variation in spatial performance is attributed to the local climatic conditions. Any clear sky model benefits from large volumes of clear sky data in order to model the atmospheric conditions at a location. While Tucson's desert like climate allows for a large number of clear days, it is harder to get such data points from the rain forests of Seria.

Finally, we see that BS-Basic outperforms BS-Ineichen significantly even though they both only differ in the base model. This is due to the fact that the Ineichen model is a very rigid model with only one learnable parameter - Linke Turbidity. We believe that our base model abstracts out most of the physical rigidity imposed on the Ineichen model, thereby making our base model more robust and better suited for deriving the data-driven clear sky models.

Next, we studied the benefits of learning temporal variations. We conducted experiments by replacing the Basic learner with the advanced temporal learners in our methodology. Since the Seasonal, Seasonal-Azimuthal & Seasonal-Hourly methods model each season distinctly, we further separated the training and test datasets into seasons for them. As Tucson has 2 major and 3 minor seasons, we divided the data into 5 parts, each part corresponding to a season. Similarly, the Seria and Bangalore dataset was divided into 4 parts for the 4 seasons there. The results presented for the seasonal learning methods are averages of the results across seasons. The results (see Table 3) show that Seasonal-Hourly (S-H) and Seasonal-Azimuthal (S-A) perform the best. In fact, their errors are 3 to 4 times less than the state of the art models. This confirms our reasoning that seasonal and diurnal environmental effects play a significant role on the clear sky irradiation and should be taken into account by a good clear sky model.

Table 2: RMSE (nRMSE) of various clear sky models

Method	Bangalore	Seria	Tucson	Avg nRMSE
ASHRAE	116 (24.7)	101 (20.9)	51 (8.5)	18
Haurwitz	108 (23.1)	75 (15.2)	48 (8.2)	15.5
Ineichen	153 (32.6)	118 (24.8)	40 (7.1)	21.5
BS-Ineichen	68 (14.5)	84 (17)	28 (5)	12.5
BS-Basic	40 (8.5)	54 (10.9)	24 (4.1)	7.83

Table 3: RMSE (nRMSE) of temporal Blue-Skies models

Learner	Bangalore	Seria	Tucson	Avg nRMSE
Seasonal (S)	35 (7.8)	44 (6.9)	20 (3.4)	6
Hourly (H)	34 (7.2)	52 (11.1)	26 (4.7)	7.7
Azimuthal (A)	32 (6.9)	51 (10.5)	26 (4.7)	7.4
S-H	29 (6.5)	42 (6.6)	19 (3.2)	5.4
S-A	27 (6)	42 (6.7)	18 (3)	5.2

We also compared with the more complex Numerical Weather Prediction models. Due to the complexity of these models, accurate parameter values are available for only certain locations. For Tucson, the REST2 [Gueymard, 2008] model has RMSE of $36.8W/m^2$ (nRMSE: 15.08) while for Bangalore, the StreamerRT model [Key and Schweiger, 1998], has RMSE of $126.24W/m^2$ (nRMSE: 25.32). Moreover, a recent model by Kim *et al.* [2016], which uses satellite images to determine visible reflectance and brightness corresponding to various GHI values, also reports RMSE value of 41.4 W/m^2 on the Tucson dataset.

5 Conclusion and Future Work

A clear sky model is a critical input to forecasting solar irradiance at any location. However, current models are highly inaccurate due to a variety of factors. In this paper, we presented a novel data-driven methodology for creating customized clear sky models for any location using a combination of AI & ML techniques. When evaluated on real datasets, the models generated by our method have errors 3 to 4 times lower than current state of the art. For future work, we intend to build such a data-driven methodology for forecasting solar energy production. Another direction is to use our approach for environmental impact analysis such as studying the changes in the clear sky conditions over the years (e.g., comparing pollution levels of Beijing between winter 2006 and winter 2016).

References

[Andreas and Wilcox, 2010] A Andreas and S Wilcox. Observed atmospheric and solar information system (oasis); tucson, arizona (data). Technical report, NREL Report No. DA-5500-56494. doi: 10. 5439/ 1052226, 2010.

[ASHRAE, 1979] ASHRAE. *Handbook of Fundamentals 1979, American Society of Heating, Refrigeration, and Air- Conditioning Engineers*. New York, 1979.

[Badescu, 2008] Viorel Badescu. Verification of some very simple clear and cloudy sky models to evaluate global solar irradiance. *Solar Energy*, 61(4):251 – 264, 2008.

[Bird, 1984] Richard E. Bird. A simple, solar spectral model for direct-normal and diffuse horizontal irradiance. *Solar Energy*, 32(4):461 – 471, 1984.

[Brutsaert, 1975] Wilfried Brutsaert. On a derivable formula for long-wave radiation from clear skies. *Water Resources Research*, 11(5):742–744, 1975.

[Bucholtz, 1995] Anthony Bucholtz. Rayleigh-scattering calculations for the terrestrial atmosphere. *Applied Optics*, 34(15):2765–2773, 1995.

[Bdescu, 1987] V Bdescu. Can the model proposed by barbaro et al. be used to compute global solar radiation on the romanian territory? *Solar Energy*, 38(4):247–254, 1987.

[Chameides *et al.*, 1999] William L Chameides, H Yu, SC Liu, M Bergin, X Zhou, L Mearns, G Wang, CS Kiang, RD Saylor, C Luo, et al. Case study of the effects of atmospheric aerosols and regional haze on agriculture: An opportunity to enhance crop yields in china through emission controls? *Proceedings of the National Academy of Sciences*, 96(24):13626–13633, 1999.

[Daneshyar, 1978] M. Daneshyar. Solar radiation statistics for iran. *Solar Energy*, 21(4):345 – 349, 1978.

[Davies and McKay, 1982] John A. Davies and Donald C. McKay. Estimating solar irradiance and components. *Solar Energy*, 29(1):55 – 64, 1982.

[Davies and McKay, 1989] J.A. Davies and D.C. McKay. Evaluation of selected models for estimating solar radiation on horizontal surfaces. *Solar Energy*, 43(3):153 – 168, 1989.

[Elder and Strong, 1953] Tait Elder and John Strong. The infrared transmission of atmospheric windows. *Journal of the Franklin Institute*, 255(3):189–208, 1953.

[Ester *et al.*, 1996] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[Grigiante *et al.*, 2011] M. Grigiante, F. Mottes, D. Zardi, and M. de Franceschi. Experimental solar radiation measurements and their effectiveness in setting up a real-sky irradiance model. *Renewable Energy*, 36(1):1 – 8, 2011.

[Gueymard, 1993] Christian Gueymard. Critical analysis and performance assessment of clear sky solar irradiance models using theoretical and measured data. *Solar Energy*, 51(2):121–138, 1993.

[Gueymard, 2008] Christian A Gueymard. Rest2: High-performance solar radiation model for cloudless-sky irradiance, illuminance, and photosynthetically active radiation–validation with a benchmark dataset. *Solar Energy*, 82(3):272–285, 2008.

[Haurwitz, 1945] Bernhard Haurwitz. Insolation in relation to cloudiness and cloud density. *Journal of Meteorology*, 2(3):154–166, 1945.

[Ineichen and Perez, 2002] P. Ineichen and R. Perez. A new airmass independent formulation for the linke turbidity coefficient. *Solar Energy*, 73(3):151 – 157, 2002.

[Jensen,] Ingrid Stver Jensen. Weather statistics for bangalore. `https://www.yr.no/place/india/karnataka/bangalore/statistics.html`.

[Kasten and Czeplak, 1980] F. Kasten and G. Czeplak. Solar and terrestrial radiation dependent on the amount and type of cloud. *Meteorologische Rundschau*, 24:177–189, 1980.

[Kasten and Young, 1989] Fritz Kasten and Andrew T Young. Revised optical air mass tables and approximation formula. *Applied optics*, 28(22):4735–4738, 1989.

[Kasten, 1980] F. Kasten. A simple parameterization of the pyrheliometric formula for determining the linke turbidity factor. *Meteorologische Rundschau*, 33:124–127, 1980.

[Key and Schweiger, 1998] Jeffrey R Key and Axel J Schweiger. Tools for atmospheric radiative transfer: Streamer and fluxnet. *Computers & Geosciences*, 24(5):443–451, 1998.

[Kim *et al.*, 2016] Chang Ki Kim, William F Holmgren, Michael Stovern, and Eric A Betterton. Toward improved solar irradiance forecasts: Derivation of downwelling surface shortwave radiation in arizona from satellite. *Pure and Applied Geophysics*, pages 1–19, 2016.

[Liu and Jordan, 1960] Benjamin YH Liu and Richard C Jordan. The interrelationship and characteristic distribution of direct, diffuse and total solar radiation. *Solar energy*, 4(3):1–19, 1960.

[Long and Ackermani, 2000] C. N. Long and T. P. Ackermani. Identification of clear skies from broadband pyranometer measurements and calculation of downwelling shortwave cloud effects. *Journal of Geophysical Research-Atmospheres*, 105:15609–15626, Jan 2000.

[McCartney, 1976] Earl J McCartney. Optics of the atmosphere: scattering by molecules and particles. *New York, John Wiley and Sons, Inc., 1976. 421 p.*, 1, 1976.

[NOAA,] NOAA. National Weather Service - NWS Tucson. `http://www.wrh.noaa.gov/twc/monsoon/monsoon.php`.

[Reda and Andreas, 2004] Ibrahim Reda and Afshin Andreas. Solar position algorithm for solar radiation applications. *Solar energy*, 76(5):577–589, 2004.

[Reno *et al.*, 2012] Matthew J Reno, Clifford W Hansen, and Joshua S Stein. Global horizontal irradiance clear sky models: implementation and analysis. *SANDIA report SAND2012-2389*, 2012.

[Southworth, 1945] GC Southworth. Microwave radiation from the sun. In *Classics in Radio Astronomy*, pages 168–181. Springer, 1945.

[Wilcox, 2007] Stephen Wilcox. National solar radiation database 1991-2005 update: User's manual. Technical report, National Renewable Energy Laboratory (NREL), Golden, CO., 2007.

[WWO,] WWO. Seria monthly climate average, brunei darussalam. `http://www.worldweatheronline.com/seria-weather-averages/bn.aspx`.