A Data-driven Clear Sky Model for Direct Normal Irradiance

Yu Shen, Haikun Wei, Tingting Zhu, Xin Zhao and Kanjian Zhang

Key Laboratory of Measurement and Control of CSE, Ministry of Education,
School of Automation, Southeast University, Nanjing 210096, China

Email: hkwei@seu.edu.cn

Abstract. Solar energy is an ideal new energy source for power systems. In order to integrate solar energy into the power grid, an evaluation of the irradiance input to solar power systems is required in many applications. Clear-sky models describe the maximum input of solar power systems, which are particularly important for solar irradiance forecasting and numerical weather prediction. The existing clear sky models are empirical or semi-empirical models. In order to make use of the information in historical big data, a data-driven clear sky model for direct normal irradiance is proposed in this paper. Firstly, a clear sky detection algorithm is designed to collect clear-sky data from historical big data automatically. After that, the weighted k-Nearest Neighbors regression algorithm is used to calculate clear-sky direct normal irradiance by historical clear-sky data. The database of the National Renewable Energy Laboratory (NREL) is used for experiments to evaluate the performance of the proposed model, compared with other three clear sky models. The results of the experiments show that the proposed model is more accurate than other three clear sky models.

1. Introduction

Solar energy is an abundant, renewable, widely distributed, safe, and clean, making it an ideal new energy source for power systems. In order to integrate solar energy into the power grid, an evaluation of the irradiance input to solar power systems is required in many applications. Solar radiance may be attenuated by clouds and aerosols, thus the input of solar power system has some uncertainty. Clear-sky models describe the maximum input of solar power systems, which are particularly important for the design and sizing of solar power systems [1]. Besides, clear sky models can be used for solar irradiance forecasting and numerical weather prediction [2].

In the literature, simple clear sky models only use geometric variables to model the annual cycle of solar irradiance [3]. Daneshyar-Paltridge-Proctor model [4][5] uses zenith angle for calculation. Meinel model [6] includes zenith angle and extra-terrestrial solar irradiance. Laue model [7] adds elevation to modify Meinel model. Complex clear sky models take atmospheric variables into consideration to describe the solar irradiance attenuated by aerosols. Ineichen model [8] uses zenith angle, elevation, extra-terrestrial solar irradiance and Linke turbidity factor to calculate direct normal irradiance. Reference Evaluation of Solar Transmittance (REST) model [9] takes precipitable water, ozone amount, nitrogen dioxide amount, Angstrom turbidity coefficients and site pressure as inputs, and respectively calculates water vapor absorption, ozone absorption, nitrogen dioxide absorption, Rayleigh scattering, uniformly mixed gases absorption and aerosol extinction. REST2 model [1] separates the light wave into 2 bands, which gains more accuracy. There are many other complex clear sky models [10][11] such as Bird model [12][13], Simplified Solis model [14], Multilayer-weighted transmittance (MLWT) model [15][16].
The clear sky models mentioned above are empirical models or semi-empirical models. This article proposes a novel data-driven clear sky model for direct normal irradiance. In order to make use of the information in historical big data, a clear sky detection algorithm based on direct normal irradiance and air mass is designed in this paper. Firstly, clear-sky data is detected from historical big data automatically. After that, the weighted k-Nearest Neighbors regression algorithm is used to calculate clear-sky direct normal irradiance by historical clear-sky data. The rest parts of this paper are organized as follows: in section 2, a clear sky detection algorithm is described in detail; a data-driven clear sky model is proposed based on weighted k-Nearest Neighbors regression algorithm using historical clear-sky data in section 3; the experiments and results using NERL data set are shown in section 4 and conclusions are drawn in section 5.

2. Clear sky detection
In order to take advantage of the useful information about clear sky in the historical big data, a clear sky detection algorithm is designed in this section. As shown in figure 1, the blue line represents direct normal irradiance (DNI) curve about time. Before 12:00 is clear sky, and after 13:00 is cloudy.

![Figure 1. Direct normal irradiance curve.](image)

By observing DNI curve in figure 1, clear sky data is detected by the following steps:
- Estimate aerosol extinction coefficient
  The Beer-Lambert law reveals the relationship between DNI and aerosol extinction coefficient at some wave length $\lambda$, as equation (1).
  \[
  E(\lambda) = E_{\text{tot}}(\lambda) \cdot e^{-\tau_{AM}(\theta_z)}
  \] 
  (1)
  Ignoring the difference about aerosol extinction coefficient at any wave length, and take logarithm on both sides of the formula, the aerosol extinction coefficient can be approximately calculated in equation (2).
  \[
  \ln E = \ln I_0 - \tau AM(\theta_z)
  \] 
  (2)
  where $I_0$ is the total solar irradiance, varying from 1363 $W/m^2$ to 1368 $W/m^2$. The $I_0$ can be approximately represented by the average of total solar irradiance. Therefore, the aerosol extinction coefficient $\tau$ can be calculated if DNI and airmass are given. It is obvious that the larger the aerosol extinction coefficient is, the more attenuation is caused by aerosols, and the less DNI reaches at the ground. Hence, for clear sky, the aerosol extinction coefficient is not
larger than a certain value.

- **Estimate convexity**
  We can see obviously in figure 1 that before 12:00, clear sky DNI curve is approximately convex function about time, allowing a small range of noise; after 13:00, cloudy DNI curve fluctuates greatly and hardly has convexity. Therefore, we can estimate convexity to distinguish clear sky DNI curve from cloudy DNI curve.

- **Estimate duration**
  Clear sky duration is not less than a certain length of time. After two steps above, if the aerosol extinction coefficient is not larger than a certain value and the DNI curve satisfies convexity for some time, we also need to estimate whether this duration is less than a certain length of time, for example, 1 hour.
  After the steps above, the samples with label one can be considered as clear sky data.

3. **A data-driven clear sky model**

Considering the annual cycle of solar orbit and the earth’s climate [17], the clear-sky direct normal irradiance has obvious seasonal difference. In this paper, we assume that the direct normal irradiance in the same month and day may have the similar pattern [18].

In this section, a data driven clear sky model is proposed based on the weighted k-Nearest Neighbors regression algorithm [19]. Firstly, historical clear sky direct normal irradiance data is detected and collected by the method described in section 2. Then the weighted k-Nearest Neighbors (k-NN) regression algorithm is used to calculate the clear-sky direct normal irradiance (DNI). The clear sky model is shown in figure 2.

![Figure 2. The data-driven clear sky model.](image-url)

The selection of k, the distance measurement and the decision rule are three basic elements of the k-Nearest Neighbor algorithm.

- **The selection of k**
  In this paper, the value of k takes a relatively small number, and then gradually increases, using cross validation to select the optimal value of k.

- **The distance measurement**
  We define the distance from the sampling year to the current year as equation (3).

\[
d_i = currentyear - samplingyear_i
\]  

Where \( d_i \) is the distance of sample \( i \), and \( samplingyear_i \) is the year of sample \( i \).

- **The decision rule**
  We define the decision rule as equation (4) and equation (5).
Where $w_i$ is the weight of sample $i$, and $y_i$ is the DNI value of sample $i$.

After calculating the clear-sky direct normal irradiance by weighted k-Nearest Neighbors regression algorithm, we can get a DNI curve for each day. The historical clear-sky data may be not continuous in some cases, thus the DNI calculated by historical data may not be satisfying. We modify the DNI by handling outliers and missing values, then regain convexity of the curve.

4. Experiments and results

4.1. Data collection

Data analyzed in this paper is provided by Solar Radiation Research Laboratory, which belong to National Renewable Energy Laboratory. The local information is shown in table 1.

| Latitude     | Longitude | Elevation | Time Zone |
|--------------|-----------|-----------|-----------|
| 39.74°N      | 105.18°W  | 1829m     | GMT-7     |

In this paper, data from 2004 to 2017 at 1-min sampling rate is used, including direct normal irradiance (DNI) and zenith angle.

4.2. Performance of clear sky detection

Define the threshold of aerosol extinction coefficient as 0.5, the threshold of noise as 10 $\text{W/m}^2$, and threshold of duration as 1 hour. Clear sky detection for 2017-01-01 is shown in figure 3, the dotted blue line denotes the original measured DNI. If it is clear-sky data, the solid red line coincide exactly with the blue line. Otherwise, the solid red line falls to zero.

![Figure 3. Clear sky detection.](image-url)
4.3. Performance of the data-driven clear sky model

Experiments of the data-driven clear sky model proposed in this paper are conducted to evaluate the performance. For comparison, we adopt other three existing clear sky models, including a simple model (Meinel) and two complex models (Ineichen and REST). The Meinel mode [6] is defined as equation (6).

\[ E = I_0 \times 0.7^{AM^{0.678}} \]  
(6)

The Ineichen model [8] is defined as equation (7).

\[ E = b \cdot I_0 \cdot \exp(-0.09 \cdot AM \cdot (T_{LK} - 1)) \]  
(7)

where \( T_{LK} \) is Linke turbidity factor, \( b = 0.664 + 0.163 / f_h \) and \( f_h = \exp(-\text{altitude} / 8000) \).

The REST model [9] is defined as equation (8).

\[ E = T_R \cdot T_g \cdot T_o \cdot T_n \cdot T_w \cdot T_a \cdot I_0 \]  
(8)

where \( T_R \), \( T_g \), \( T_o \), \( T_n \), \( T_w \) and \( T_a \) are the band transmittances for Rayleigh scattering, uniformly mixed gases absorption, ozone absorption, nitrogen dioxide absorption, water vapor absorption and aerosol extinction, respectively.

We use normalized root mean square error \( nRMSE \), normalized mean absolute error \( nMAE \) and normalized mean bias error \( nMBE \) to evaluate the performance of each model.

Normalized root mean square error is defined as equation (9).

\[ nRMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (E_{ci} - E_{mi})^2} \times 100\% \]  
(9)

Normalized mean absolute error is defined as equation (10).

\[ nMAE = \frac{1}{N} \sum_{i=1}^{N} \left| E_{ci} - E_{mi} \right| \times 100\% \]  
(10)

Normalized mean bias error is defined as equation (11).

\[ nMBE = \frac{1}{N} \sum_{i=1}^{N} (E_{ci} - E_{mi}) \times 100\% \]  
(11)

In the formulas above, \( N \) is the number of samples, \( E_{ci} \) represents the calculated DNI by sample \( i \), \( E_{mi} \) denotes the measured DNI and \( \bar{E}_{mi} \) represents the average of measured DNI.

| Year | Models     | nRMSE (%) | nMAE (%) | nMBE (%) |
|------|------------|-----------|----------|----------|
| 2015 | Meinel     | 13.45     | 10.91    | -5.37    |
|      | Ineichen   | 6.22      | 4.98     | 2.40     |
|      | REST       | 5.56      | 4.90     | 1.11     |
|      | Data-driven| **4.44**  | **3.47** | **-0.84**|
| 2016 | Meinel     | 14.42     | 12.00    | -2.44    |
|      | Ineichen   | 8.61      | 7.18     | 2.18     |
|      | REST       | 6.40      | 5.49     | 2.60     |
|      | Data-driven| **6.19**  | **5.05** | **-1.21**|
| 2017 | Meinel     | 15.73     | 13.38    | -3.61    |
|      | Ineichen   | 8.35      | 7.13     | 1.41     |
|      | REST       | 6.26      | 5.72     | **0.95** |
|      | Data-driven| **4.89**  | **3.67** | -1.87    |
Select $k = 10$, and Table 2 lists the results of calculating clear sky DNI from 2015 to 2017, with three existing models and the data-driven model proposed in this paper. The best $nRMSE$, $nMAE$ and $nMBE$ for each year are shown in bold. Plots of measured DNI, the data-driven model predicted DNI and other three existing models predicted DNI for 2017-01-17 are shown in figure 4.

![Figure 4. DNI forecasting plot for 2017-01-17](image_url)

The results of the experiments show that the data-driven model proposed in this paper achieves the least $nRMSE$ and $nMAE$, performs better than other three clear sky models.

5. Conclusions
In this paper, a data-driven clear sky model is proposed for direct normal irradiance. Firstly, a clear sky detection algorithm is designed to collect clear sky data from historical big data automatically. Then, the weighted k-Nearest Neighbors regression algorithm is used to calculate the clear sky direct normal irradiance. Experiments are conducted using the database of NREL. The performance of the proposed model and three other clear sky models are compared, including Meinel, Ineichen and REST. The results of the experiments show that the data-driven model proposed in this paper achieves the least $nRMSE$ and $nMAE$, performs better than other three clear sky models.

6. References
[1] Gueymard C A 2008 J. Solar Energy. REST2: High-performance solar radiation model for cloudless-sky irradiance, illuminance, and photosynthetically active radiation – Validation with a benchmark dataset. 82(3) 272-85
[2] Law E W, Prasad A A and Kay M 2014 J. Solar Energy. Direct normal irradiance forecasting and its application to concentrated solar thermal output forecasting – A review. 108 287-307
[3] Stein J S, Hansen C W and Reno M J. 2012 J. Sand. Global horizontal irradiance clear sky models: implementation and analysis.
[4] Daneshyar M 1978 J. Solar Energy. Solar radiation statistics for Iran. 21 345-49
[5] Paltridge G W and Proctor D 1976 J. Solar Energy. Monthly mean solar radiation statistics for Australia. 18 235-43
[6] Meinel A B and Meinel M P 1976 *Applied solar energy* (MA: Addison-Wesley)

[7] Laue E G 1970 *J. Solar Energy. The measurement of solar spectral irradiance at different terrestrial elevations*. 13 43-50

[8] Ineichen P and Perez R 2002 *J. Solar Energy. A new airmass independent formulation for the Linke turbidity coefficient*. 73(3) 151-57

[9] Gueymard C A 2003 *J. Solar Energy. Direct solar transmittance and irradiance predictions with broadband models. Part I: detailed theoretical performance assessment*. 74(5) 355-79

[10] Inman R H, Edson J G and Coimbra C F M 2015 *J. Solar Energy. Impact of local broadband turbidity estimation on forecasting of clear sky direct normal irradiance*. 117 125-38

[11] Janjai S, Sriraroen K and Pattarapanitchai S. 2011 *J. Applied Energy. Semi-empirical models for the estimation of clear sky solar global and direct normal irradiances in the tropics*. 88(12) 4749-55

[12] Bird R E and Hulstrom R L 1980 *J. Solar Energy. Direct insolation models.*

[13] Bird R E and Hulstrom R L 1981 *J. Solar Energy. Simplified clear sky model for direct and diffuse insolation horizontal surfaces.*

[14] Ineichen P A 2008 *J. Solar Energy. broadband simplified version of the Solis clear sky model*. 82(8) 758-62

[15] Gueymard C 1996 *Conf.Solar. Multilayer-weighted transmittance functions for use in broadband irradiance and turbidity calculations*. 281-8

[16] Gueymard C A 2010 *J. Applied Meteorology. Turbidity Determination from Broadband Irradiance Measurements: A Detailed Multicoefficient Approach*. 37(4) 414-35

[17] Zhu T T, Wei H K, Zhao X, Zhang C and Zhang K J 2017 *J. Renewable Energy. Clear sky model for wavelet forecast of direct normal irradiance*. 104 1–8

[18] Zhang J X, Ehinger K A, Wei H K, Zhang K J and Yang J 2017 *J. Pattern Recognition. A novel graph-based optimization framework for salient object detection*. 64 39–50

[19] Harrington P 2012 *Machine Learning in Action* (Manning)

**Acknowledgments**

The authors gratefully acknowledge the National Renewable Energy Laboratory providing all data in this research. This work is supported by the National Natural Science Foundation of China (Grant No.61374006, Grant No.61773118 and Grant No.61703100), and by the Fundamental Research Funds for the Central Universities.