Physics-Guided Machine Learning for Prediction of Cloud Properties in Satellite-Derived Solar Data

Preprint

Grant Buster, Mike Bannister, Aron Habte,
Dylan Hettinger, Galen Maclaurin, Michael Rossol,
Manajit Sengupta, and Yu Xie

National Renewable Energy Laboratory

Presented at the 48th IEEE Photovoltaic Specialists Conference (PVSC 48)
June 20-25, 2020

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-6A20-79705
May 2021

Physics-Guided Machine Learning for Prediction of Cloud Properties in Satellite-Derived Solar Data

Preprint

Grant Buster, Mike Bannister, Aron Habte, Dylan Hettinger, Galen Maclaurin, Michael Rossol, Manajit Sengupta, and Yu Xie

National Renewable Energy Laboratory

Suggested Citation
Buster, Grant, Mike Bannister, Aron Habte, Dylan Hettinger, Galen Maclaurin, Michael Rossol, Manajit Sengupta, and Yu Xie. 2021. *Physics-Guided Machine Learning for Prediction of Cloud Properties in Satellite-Derived Solar Data: Preprint*. Golden, CO: National Renewable Energy Laboratory. NREL/CP-6A20-79705
https://www.nrel.gov/docs/fy21osti/79705.pdf

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-6A20-79705
May 2021

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Solar Energy Technologies Office. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government.

Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.

NREL prints on paper that contains recycled content.

Physics-Guided Machine Learning for Prediction of Cloud Properties in Satellite-Derived Solar Data

Grant Buster, Mike Bannister, Aron Habte, Dylan Hettinger, Galen Maclaurin, Michael Rossol, Manajit Sengupta, and Yu Xie

National Renewable Energy Laboratory, Golden, CO, 80401, United States

***Abstract*—With over 20 years of high-resolution surface irradiance data covering most of the western hemisphere, the National Solar Radiation Database (NSRDB) is a vital public data asset. The NSRDB uses a two-step Physical Solar Model (PSM) that explicitly considers the effects of clouds and other atmospheric variables on radiative transfer. High-quality physical and optical cloud properties derived from satellite imagery are perhaps the most important data inputs to the PSM, representing the greatest source of radiation attenuation and scattering. However, traditional methods for cloud property retrieval have their own limitations and are unable to accurately predict cloud properties outside of nominal conditions. We introduce a physics-guided neural network that can accurately predict cloud properties when traditional methods fail or are inaccurate. Using this framework, we show reductions in relative Root Mean Square Error (RMSE) for Global Horizontal Irradiance (GHI) up to 13 percentage points for timesteps that previously had missing or low-quality cloud property data. We expect that this methodology will be effective in improving the quality of cloud property and solar irradiance data in the NSRDB.**

Keywords—solar resource data, machine learning, physics-guided neural networks, cloud properties, remote sensing, satellite-derived irradiance

I. INTRODUCTION

Solar resource data is a fundamental input for virtually all solar related analyses including the analysis of solar energy conversion systems, power systems integration, market operations, and even financial investments in solar power systems. Satellite imagery has recently proven to be an effective resource for developing large quantities of solar resource data across large spatiotemporal extents [1]-[3]. Specifically, two-step physical models such as by Pinker et al. [4] and Xie et al. [5] which explicitly consider the effects of clouds and other atmospheric variables on radiative transfer have benefited from the recent improvements in satellite technology and reanalysis datasets [6]-[7].

A prominent example of solar resource data using the Physical Solar Model (PSM) by Xie et al. [5] is The National Solar Radiation Database (NSRDB), which is produced by the National Renewable Energy Laboratory (NREL) [8]. The NSRDB includes more than 20 years of surface irradiance and atmospheric data for most of the western hemisphere. The NSRDB can be freely accessed at https://nsrdb.nrel.gov/ and has been used widely by an ever-growing group of researchers and industry [8].

The cloud physical and optical properties used by the NSRDB are retrieved from satellite measurements in visible, near-infrared, and infrared channels from The Advanced Very High-Resolution Radiometer (AVHRR) Pathfinder Atmospheres–Extended (PATMOS-x) project [9]. While this cloud property data is accurate and of great utility to the NSRDB, the underlying methods such as the Daytime Cloud Optical and Microphysical Properties Algorithm (DCOMP) [10] can fail to converge under suboptimal conditions with certain surface types or extreme solar zenith angles, resulting in inaccurate or missing cloud property data. To compensate, the NSRDB executes a heuristic gap-fill procedure to fill cloud property data that is missing from the DCOMP output. The NSRDB version 3.0.0 gap fill procedure, described in Section 3.2 of Sengupta et al.'s 2018 paper [8], fills the irradiance at a timestep with missing cloud properties using a simple cloudy-to-clear Global Horizontal Irradiance (GHI) ratio from the nearest timestep with valid cloud properties. In the NSRDB version 3.1.0, a slightly modified gap-fill procedure was introduced that would fill missing cloud properties using the temporally nearest valid cloud properties of the same cloud phase (water or ice).

While the overall accuracy of the NSRDB is quite high with relative GHI mean bias error typically below 5 percent [11], the missing cloud properties nevertheless represent a significant fraction (between 20 and 30 percent) of daylight cloudy timesteps. These timesteps typically have relative GHI Root Mean Square Error (RMSE) 2 to 10 percentage points higher than timesteps with valid cloud properties produced directly by the DCOMP algorithm. To address this issue, we have developed machine learning methods for cloud property retrieval that can be used to complement the traditional methods from PATMOS-x [9] and DCOMP [10].

Machine learning methods have been used in a variety of remote sensing applications such as the characterization of airborne particulates, cloud detection, and even the direct prediction of solar radiation [12]-[14]. For this work, we propose a method to leverage machine learning methods to predict missing cloud properties while preserving the key strengths of the NSRDB methodology. Namely, we preserve the cloud identification methods from Heidinger et al [9], the valid cloud properties produced by the DCOMP algorithm from Walther et al [10], and the PSM by Xie et al. [5] on which the NSRDB is based. In this fashion, we are able to make significant improvements to the NSRDB while maintaining the overall data product that is already widely used by the public.

Funding provided by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE).

II. Methodology

Predicting cloud properties can be described as a regression problem where input features X are transformed into the target output variable Y. The input features X can be any data resources available to the NSRDB including Geostationary Operational Environmental Satellite (GOES) imagery, and Y is the relevant physical and optical cloud properties. This problem can be handled by training a simple feed-forward neural network $f : X \rightarrow Y$ that produces cloud properties predictions $f(X) = \hat{Y}$ given some known input X. Such a model would be trained to minimize the empirical loss $\mathcal{L}_{NN}$ of the model predictions $\hat{Y}$ versus known outputs Y:

$$\text{argmin}_f \quad \mathcal{L}_{NN}[f(X), Y] \tag{1}$$

However, for such a formulation to be successful we would require data for X and Y over the entire expected observational range. Because this problem is specifically attempting to produce cloud properties where the traditional models do not, much of the desired prediction space includes out-of-sample data X with no known data Y. Indeed, a simple feed-forward neural network trained only on the cloud properties successfully produced by the DCOMP algorithm is observed to not predict accurate cloud properties when extended to the out-of-sample prediction space, as shown in Section III. Instead, we develop a physics-guided neural network (PHYGNN) architecture that is trained using the full NSRDB radiative transfer model along with additional training data sources to accurately predict cloud properties for all daylight timesteps, including data that is out-of-sample for the simple formulation in (1). This model architecture augments (1) by adding a physics-based loss term, $\mathcal{L}_{PHY}$:

$$\text{argmin}_f \quad \alpha_{NN}\, \mathcal{L}_{NN}[f(X), Y] + \alpha_{PHY}\, \mathcal{L}_{PHY}[f(X), P] \tag{2}$$

Where $\alpha \in \mathbb{R}$ are weighting factors for the two loss terms and P can be any supplemental input data used to calculate the physics-based loss term $\mathcal{L}_{PHY}$. In this case, $\mathcal{L}_{PHY}$ takes the cloud property predictions $f(X) = \hat{Y}$ along with supplemental inputs P, runs the full PSM by Xie et al. [5], and compares the predicted irradiance values against ground-measured irradiance. This method for training a PHYGNN model to predict cloud properties has several advantages for predicting cloud properties in the NSRDB. Primarily, the observation space of the training data X and Y can be extended using additional data P. An additional holistic benefit is that the PHYGNN model is trained on how cloud properties are used in the PSM and learns how to predict properties that result in more accurate irradiance values. The general PHYGNN architecture described in (2) has been used previously for a variety of applications in the physical sciences [15]-[16], and is shown in Section III to greatly outperform the simple feed-forward neural network described by (1).

TABLE I. PHYGNN Dataset Names and Uses

Dataset Name	Use
Solar zenith angle	Feature, supplemental P input
Air temperature	Feature
Dew point	Feature
Relative humidity	Feature
Total precipitable water	Feature, supplemental P input
Surface albedo	Feature, supplemental P input
Cloud type	Feature, supplemental P input
Cloud probability	Feature
Cloud fraction	Feature
0.65 µm reflectance	Feature
0.65 µm reflectance standard deviation (on a 3x3 grid)	Feature
3.75 µm reflectance	Feature
3.75 µm brightness temperature	Feature
11.0 µm brightness temperature	Feature
11.0 µm brightness temperature standard deviation (on a 3x3 grid)	Feature
Aerosol optical depth	Supplemental P input
Alpha (aerosol angstrom exponent)	Supplemental P input
Surface pressure	Supplemental P input
Aerosol single scattering albedo	Supplemental P input
Aerosol asymmetry parameter	Supplemental P input
Total ozone	Supplemental P input
Time index	Supplemental P input
Ground-measured GHI	Supplemental P input
Cloud optical depth	PHYGNN output
Cloud effective particle radius	PHYGNN output

Besides the custom loss function described in (2), the PHYGNN architecture used in this work is a standard feed-forward neural network with 3 layers, 64 nodes per layer, 18 input features (including one-hot encodings), and 2 output channels. The network also includes a 1 percent dropout rate on all hidden layer output connections during training. The model is trained using the Adam optimizer with a learning rate of 0.002. The training is split into 100 pre-training epochs with $\alpha_{NN} = 1$ and $\alpha_{PHY} = 0$, and 100 final training epochs with $\alpha_{NN} = 0.5$ and $\alpha_{PHY} = 0.5$. Model weights are updated 64 times per epoch (64 batches per epoch). Loss values are calculated using mean absolute error.

The PHYGNN model is trained using satellite data from GOES [6], reanalysis data from Modern Era Retrospective Analysis for Research and Applications Version 2 (MERRA2) [7], cloud identification from PATMOS-x [9], surface albedo data derived from MODIS [17], and ground measurement data from the NOAA Surface Radiation Budget (SURFRAD)

Network Observations sites [18]. The full set of training features (X), supplemental data inputs (P) to calculate the physical loss term $\mathcal{L}_{PHY}$, and PHYGNN outputs ($\hat{Y}$) are recorded in TABLE I. Four years of data is used from four GOES satellites (GOES West 15/17 and GOES East 13/16) over the seven SURFRAD sites. Including the data from the newer GOES satellites (GOES 16 and 17) with higher temporal resolution, a total of 1,226,400 daylight data observations were used. The available training data was randomly partitioned with 20 percent used for post-training validation, 20 percent of the remaining data used for in-situ training validation, and the remaining data (64 percent of the total dataset) used for training. For the results in this paper, the SURFRAD data, which is typically available at 1-minute intervals, is averaged using a 15-minute centered moving window average. This helps reduce the variability associated with a point-source measurement when compared to the large spatial extent that the NSRDB grid cells represent.

III. Results

The results presented in Fig. 1 and Fig. 2 show the Global Horizontal Irradiance (GHI) and Direct Normal Irradiance (DNI) RMSE for four years of the NSRDB irradiance data vs. ground-measured irradiance, respectively. Data presented in Fig. 1 and Fig. 2 is exclusively from the 20 percent of the data that the simple feed-forward neural network and PHYGNN models were not trained on. For these results, the four years of NSRDB data at each of the seven SURFRAD sites is produced twice: once using source data from the GOES East satellites, and once using source data from the GOES West satellites. Fig. 1 and Fig. 2 present results from all daylight cloudy timesteps that are missing cloud property inputs from the DCOMP algorithm [10]. The "DCOMP + Gap-Fill" data in Fig. 1 and Fig. 2 was produced using the NSRDB version 3.1.0 cloud property heuristic gap-fill procedure described in Section I. Relative error metrics are calculated with respect to the mean of the data. It should be noted that because these results are for all daylight cloudy timesteps, the absolute magnitude of the errors can be quite high because the mean data value which is used to normalize the metrics includes low irradiance timesteps when the sun is rising or setting. However, these timesteps are important to include because the DCOMP algorithm performs poorly when the sun is close to the horizon.

As shown in Fig. 1 and Fig. 2, NSRDB data produced from the heuristic gap-filled DCOMP cloud properties exhibits high relative RMSE. The simple feed-forward neural network with loss function defined by (1) is shown to predict accurate cloud properties for some locations but performs worse than the heuristic gap-fill method for others. In contrast to the simple feed-forward neural network model, the PHYGNN model with loss function defined by (2) significantly improves the validation statistics for all locations, reducing the relative GHI RMSE by 6 to 13 percentage points from the heuristic gap-filled DCOMP results.

A noteworthy result is the highly inaccurate NSRDB GHI data at the Penn. State University (PSU) location predicted by both the heuristic gap-fill and simple feed-forward neural network models (relative GHI RMSE of 54.2 and 65.5 percent, respectively) and the significant improvement by the PHYGNN model (relative GHI RMSE of 40.4 percent). The inaccurate irradiance at PSU is primarily because the site is at a very extreme viewing angle from the GOES West satellites, which dramatically increases the RMSE even though the predicted irradiance from the GOES East satellites is accurate. In fact, in the actual NSRDB data, locations as far east as Pennsylvania would never be produced using data from the GOES West satellites. Nevertheless, this provides a challenging prediction scenario for these models and shows that the PHYGNN model is able to learn how to produce accurate cloud properties even in the worst out-of-sample conditions, reducing the relative GHI RMSE by 13.8 percentage points.

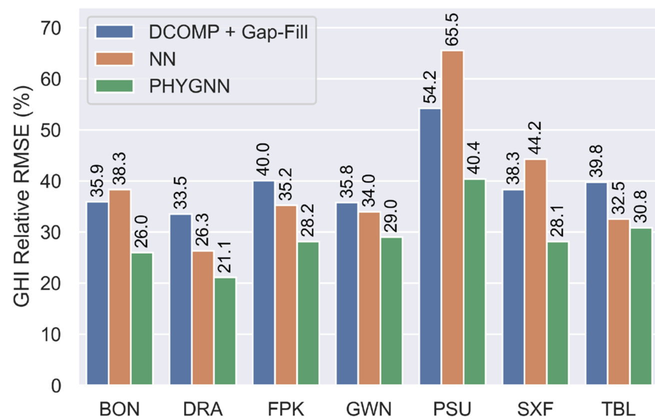

Fig. 1. Ground-truth validation of cloudy NSRDB GHI with missing cloud property inputs for seven SURFRAD sites. Includes validation of NSRDB data produced using only heuristic gap-filled cloud properties ("DCOMP + Gap-Fill"), using a simple feed-forward neural network ("NN"), and using a physics-guided neural network ("PHYGNN").

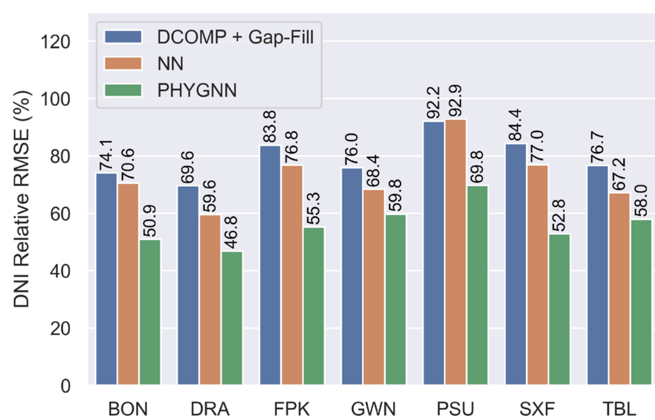

Fig. 2 Ground-truth validation of cloudy NSRDB DNI with missing cloud property inputs for seven SURFRAD sites. Includes validation of NSRDB data produced using only heuristic gap-filled cloud properties ("DCOMP + Gap-Fill"), using a simple feed-forward neural network ("NN"), and using a physics-guided neural network ("PHYGNN").

IV. Conclusions

In this work, we use machine learning techniques to predict physical and optical cloud properties for input to satellite-derived solar resource data. By training a neural network with an understanding of a full radiative transfer model, our physics-guided approach is able to significantly increase the accuracy of cloud property predictions as related to the surface irradiance experienced on the ground. We validate this PHYGNN model against four years of ground measurement data with inputs from four GOES satellites, along with a simple feed-forward neural network model and the heuristic gap-fill methodology that is currently used in the NSRDB. We show that the PHYGNN model greatly outperforms the simple feed-forward neural network and heuristic gap-fill methodology and is able to improve the accuracy of irradiance data in the NSRDB, especially for timesteps that were previously missing cloud property data from the traditional cloud property retrieval algorithms. Open-source software for creating PHYGNN models has been made available on GitHub [19], and NSRDB data including the improvements from the PHYGNN predictions will be available to the public in the NSRDB 2020 data (NSRDB version 3.2.0).

ACKNOWLEDGMENT

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Solar Energy Technologies Office (Systems Integration Subprogram) Contract Number 36598. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

A portion of this research was performed using computational resources sponsored by the Department of Energy's Office of Energy Efficiency and Renewable Energy and located at the National Renewable Energy Laboratory.

REFERENCES

[1] Gurtuna O, Prevot A. An overview of solar resource assessment using meteorological satellite data. Recent Advances in Space Technologies (RAST), 2011 5th International Conference on. 10.1109/RAST.2011.5966825:209 − 12; 2011.

[2] Pinker R, Frouin R, Li Z. A review of satellite methods to derived surface shortwave irradiance. Remote Sens Environ 1995;51:108–24.

[3] Hammer A, Heinemann D, Hoyer C, Kuhlemann R, Lorenz E, Müller R, et al. Solar energy assessment using remote sensing technologies. Remote Sens Environ 2003;86:423–32.

[4] Pinker RT, Laszlo I. Modeling surface solar irradiance for satellite applications on a global scale. J Appl Meteorol 1992;31:194–211.

[5] Xie Y, Sengupta M, Dudhia J. A fast all-sky radiation model for solar applications (FARMS): algorithm and performance evaluation. Sol Energy 2016;135:435–45.

[6] Schmit T, Gunshor M, Menzel W, Gurka J, Li J, Bachmeier A. Introducing the nextgeneration advanced baseline imager on GOES-R. Bull Am Meteorol Soc 2005;86:1079–96.

[7] Gelaro R, McCarty W, Suárez M, Todling R, Molod A, Takacs L, et al. The modern era retrospective analysis for research and applications, version 2 (merra-2). J Clim 2017;30:5419–54.

[8] Sengupta M., Xie Y., Lopez A., Habte A., Maclaurin G., Shelby J., 2018. The National Solar Radiation Data Base (NSRDB). Renewable and Sustainable Energy Reviews, 89, 51-60, ISSN 1364-0321.

[9] Heidinger A, Foster M, Walther A, Zhao X. The pathfinder atmospheres-extended AVHRR climate dataset. Bull Am Meteorol Soc 2014;95:909–22.

[10] Walther A, and Heidinger A. "Implementation of the Daytime Cloud Optical and Microphysical Properties Algorithm (DCOMP) in PATMOS-x." Journal of Applied Meteorology and Climatology, vol. 51, no. 7, 2012, pp. 1371–1390. JSTOR, www.jstor.org/stable/26175210. Accessed 13 Jan. 2021.

[11] A. Habte, M. Sengupta, A. Lopez, Y. Xie and G. Maclaurin, "Assessment of the National Solar Radiation Database (NSRDB 1998-2016)," 2018 IEEE 7th World Conference on Photovoltaic Energy Conversion (WCPEC), Waikoloa Village, HI, 2018, pp. 2305-2308, doi: 10.1109/PVSC.2018.8547589.

[12] Lary D, Alavi A, Gandomi A, Walker A, Machine learning in geosciences and remote sensing, Geoscience Frontiers, Volume 7, Issue 1, 2016, Pages 3-10, ISSN 1674-9871, https://doi.org/10.1016/j.gsf.2015.07.003.

[13] Bai T, Li D, Sun K, Chen Y, Li W. 2016. "Cloud Detection for High-Resolution Satellite Imagery Using Machine Learning and Multi-Feature Fusion" Remote Sens. 8, no. 9: 715.

[14] Cornejo-Bueno L., Casanova-Mateo C., Sanz-Justo J., Salcedo-Sanz S., Machine learning regressors for solar radiation estimation from satellite data, Solar Energy, Volume 183, 2019, Pages 768-775, ISSN 0038-092X, https://doi.org/10.1016/j.solener.2019.03.079.

[15] Forssell U. and P. Lindskog. "Combining Semi-Physical and Neural Network Modeling: An Example of Its Usefulness." IFAC Proceedings Volumes 30 (1997): 767-770.

[16] Karpatne A, Watkins W, Read J, and Kumar V, "Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling". arXiv:1710.11431v2 (2018).

[17] Maclaurin G, Sengupta M, Xie Y, and Gilroy N. Development of a MODIS-Derived Surface Albedo Data Set: An Improved Model Input for Processing the NSRDB. United States: N. p., 2016. Web. doi:10.2172/1335471.

[18] NOAA Earth System Research Laboratory, 1995: Surface Radiation Budget (SURFRAD) Network Observations. Data used from all 7 SURFRAD sites for years 2016-2019. NOAA National Centers for Environmental Information. 1-14-2021.

[19] Grant Buster, Michael Rossol, Mike Bannister, and Dylan Hettinger. Physics-Guided Neural Networks (phygnn). https://github.com/NREL/phygnn (version v0.0.9), 2021. https://doi.org/10.5281/zenodo.4498541.