EAEE4000_001_2022_3 Final Project
Machine Learning for Environmental Engineering and Science

Improving Solar Radiance Data Resolution with Convolutional Neural Nets

Pyranometer Onsite in Staten Island // Rachel Margolese, Plankton Energy, June 24, 2021

Name: Zaki Alattar
UNI: zaa2127
Github: github.com/zakialattar/superERA5

Introduction

The renewable energy transition has dramatically accelerated installations of solar photovoltaic (PV) power systems globally. As PV systems advance, the accurate monitoring and reporting of the performance of these sites is becoming increasingly important for developers and investors. For developers, maintenance operations depend on the timely identification of sites that underperform expectations. For investors, transparency on the performance of sites relative to the simulated performance underlying power purchase agreements builds is critical to secure financing for future projects. With billions of dollars staked on the performance of PV systems, accurately reporting the factors influencing solar power production at a given site is a vital component of the transition to renewable energy.

While temperature and wind can influence energy production, the performance of a photovoltaic system overwhelmingly depends on radiance. Radiance is a measure of the intensity of light on an object. For applications in solar power production, radiance refers to the degree to which the Sun's rays make contact with the solar panel over the course of a day. There are three common ways that PV system developers measure radiance on their sites: (1) the use of an on-site sensor known as a pyranometer, (2) pulling radiance measurements from public satellite data, and (3) purchasing additionally processed satellite data from a solar data provider. While regarded as the closest measure of ground truth, on-site sensors extend installation times and require continued maintenance. While inexpensive and easy to access, public satellite data typically has a maximum resolution of 15x15 or 30x30km which is far too imprecise for solar sites.[1] Developers and investors are therefore increasingly turning to solar data providers who aggregate a multitude of public and private satellite sources with their own proprietary numerical weather models to improve resolution and replicate the values generated by a physical pyranometer.

Advances in machine learning and artificial intelligence have shown tremendous success in climate and weather prediction, including in applications for solar modeling and radiance prediction (Yandav, Kumar, Chandel 2014).[2] Building on these findings, this project investigates the capacity for machine learning techniques to accurately model on-site pyranometers with low resolution public weather data. In particular, this project evaluates the performance of conventional and long short-term memory neural network models across four different solar PV sites in New Jersey and New York.

---

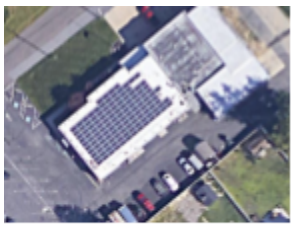
[1] "Solar radiation modeling," SolarGIS. https://solargis.com/docs/methodology/solar-radiation-modeling. Date Accessed Dec 23, 2022.

[2] Yadav, Amit Kumar, and S. S. Chandel. "Solar radiation prediction using Artificial Neural Network techniques: A review." Renewable and sustainable energy reviews 33 (2014): 772-781.

Data

Training and Training Sites

Plankton Energy operates nearly a dozen solar facilities across the Northeast United States with sites in New York, Connecticut, New Jersey, and Massachusetts. Each site contains a set of solar panel modules, optimizers, inverters and a data logging system but only a handful of sites have an operational on site weather station. The criteria for selection was based on the length of their operation and a clear maintenance log to ensure calibration. Figure 1 showcases the four solar production sites selected for this project:



Mantua Fire Department
City: Sewell, NJ
Latitude: 39.77
Longitude: -75.14

Staten Island UGE
City: Staten Island, NY
Latitude: 40.56
Longitude: -74.19

Waterford EMS
City: Atco, NJ
Latitude: 39.75
Longitude: -74.88

West Elementary School
City: New Canaan, CT
Latitude: 41.14
Longitude: -73.53

Figure 1: Site Locations

Solar installations at Mantua Fire Department (“mantua”), Staten Island UGE (“staten”), and West Elementary School (“west”) were selected as the training datasets and the Waterford EMS was selected as the test set. Both the testing and training sets begin on the same date April 21st, 2021 with the exception of Waterford EMS records which begins on September 16th, 2021. All sets’ records end at the same time on July 31st, 2022. As a result, there is an approximate ¾ split between training and testing data with 30,022 data points for training and 11,191 data points for testing.

Input Data

The input data for this project is taken from the fifth generation European Centre for Medium-Range Weather Forecasts atmospheric reanalysis dataset (ERA5).[3] Produced by the Copernicus Climate Change Service (C3S), ERA5 was selected for its ease of access through the Climate data store (CDS) and its wide acceptance in research by the scientific community.

[3] “ERA5”, European Centre for Medium-Range Weather Forecasts, https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5.

The dataset is compiled historical observations from a network of satellites, ground sites, and weather balloons with advanced modeling and data assimilation systems. This reanalysis provides hourly estimates of a large number of atmospheric, land, and oceanic climate variables covering the Earth on a 30 kilometer grid. Quality-assured updates of ERA5 from 1959 to the present are available within three months of real time and preliminary results are available within five days of real time. Figure 1 depicts this process and the corresponding data output in the context of baseline temperature rise from 1980 to 2020.

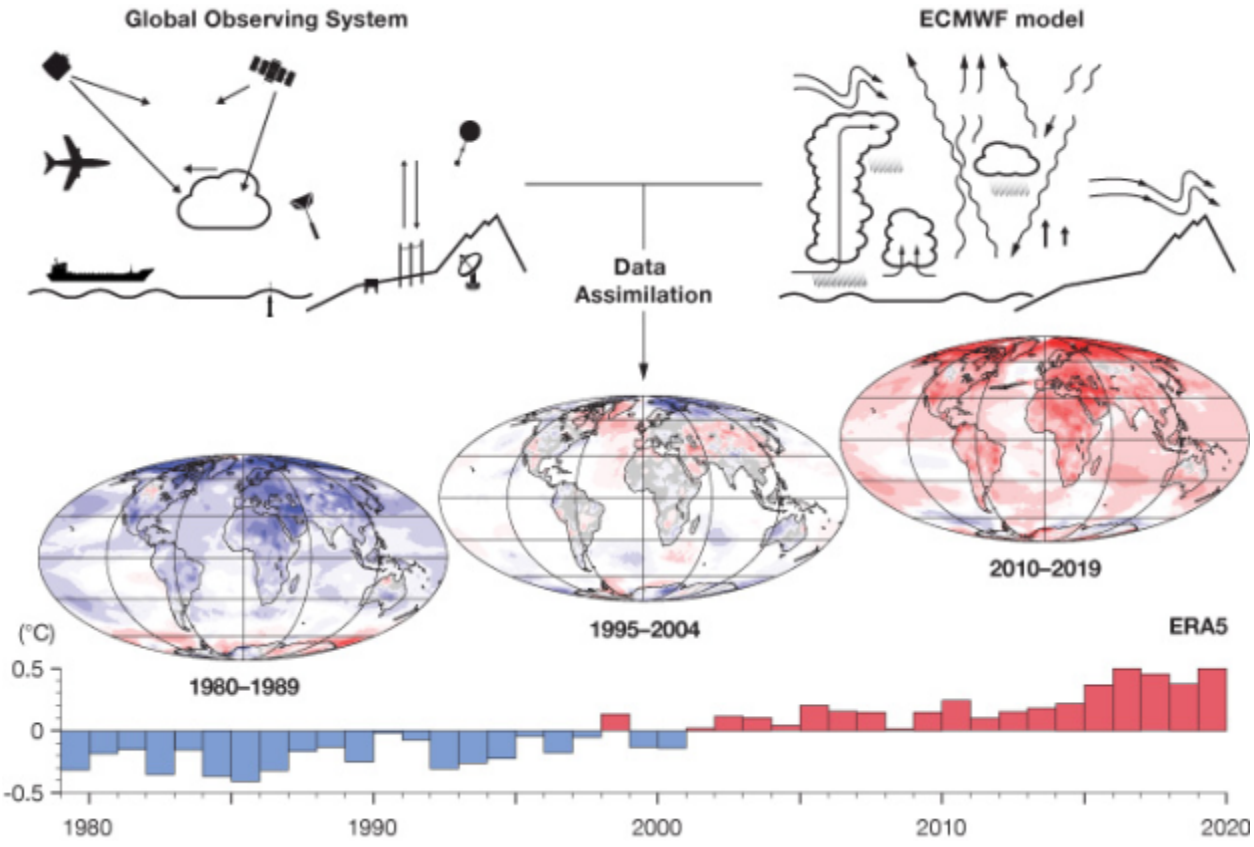


Figure 2: ERA5 Data Assimilation Schematic for Baseline Temperature Rise from 1980 - 2020[4]

ERA5 has known limitations. The dataset can contain non-physical trends and variability in the record due to changes in the observation methodology. For solar development, its low spatial resolution is unusable for the planning and monitoring of facilities with a 1 kilometer squared footprint. Yet, the wide temporal and spatial range of the data offers the opportunity to replicate this study for potentially any site and any given time period.

[4] "Fact sheet: Reanalysis", ECMWF, 9 November 2020, Accessed December 22nd, 2022. https://www.ecmwf.int/en/about/media-centre/focus/2020/fact-sheet-reanalysis

The input data was downloaded via the Climate Data Store API. API calls were run for each month on an hourly resolution with the coordinates for each site and the following three variables:

1) Surface solar radiation downwards ("ssrd"), measured in Joule per square meter
2) Two-meter temperature ("t2m"), measured in Kelvin
3) Total cloud cover ("tcc"), measured as a proportion of grid box covered

The full documentation for each variable can be found in the ECMWF GRIB Parameter database.

Output Data

The output data for this project is the local radiance acquired with permission from Plankton Energy's on-site weather stations. Each weather station contains a pyranometer sensor that measures radiance and relays the data to a logger located adjacent to inverts as seen in Figure 3. The logger records the data and posts the data to the SolarEdge Monitoring Platform through a wireless gateway. That data was accessed and downloaded via the SolarEdge API.

Figure 3: Staten Island UGE (Left) Pyranometer | (Right) Data Logging System

The model's output variable is the Global Horizontal Irradiance (GHI). At West Elementary School, the direct radiance is captured which is equivalent to GHI for purposes of this study. At Staten Island UGE, Mantua Fire Department, and Waterford EMS, however, the plane of array radiance is captured by its pyranometer and must be converted to GHI as described in the following section.

Processing and Adjustments

A number of processing steps are required in order to weigh the corresponding input variables with the output variables. First, the dates and times between the datasets must be reconciled. The SolarEdge Monitoring API which distributes its data in the user profile's local timezone while ERA5 distributes its data in Coordinated Universal Time (UTC). In order to reflect daylight savings and hour of day on which peak radiance typically occurs, ERA5 data must be localized in Eastern Time (ET). Second, the plane of array radiance must be converted to GHI by dividing by the cosine of the angle of incidence (the angle between the direction of the sun and the angle of the panel). Third, the downward surface solar radiation must also be converted to GHI by dividing by the time of accumulation (3600 seconds per hour). Finally, the values are standardized.

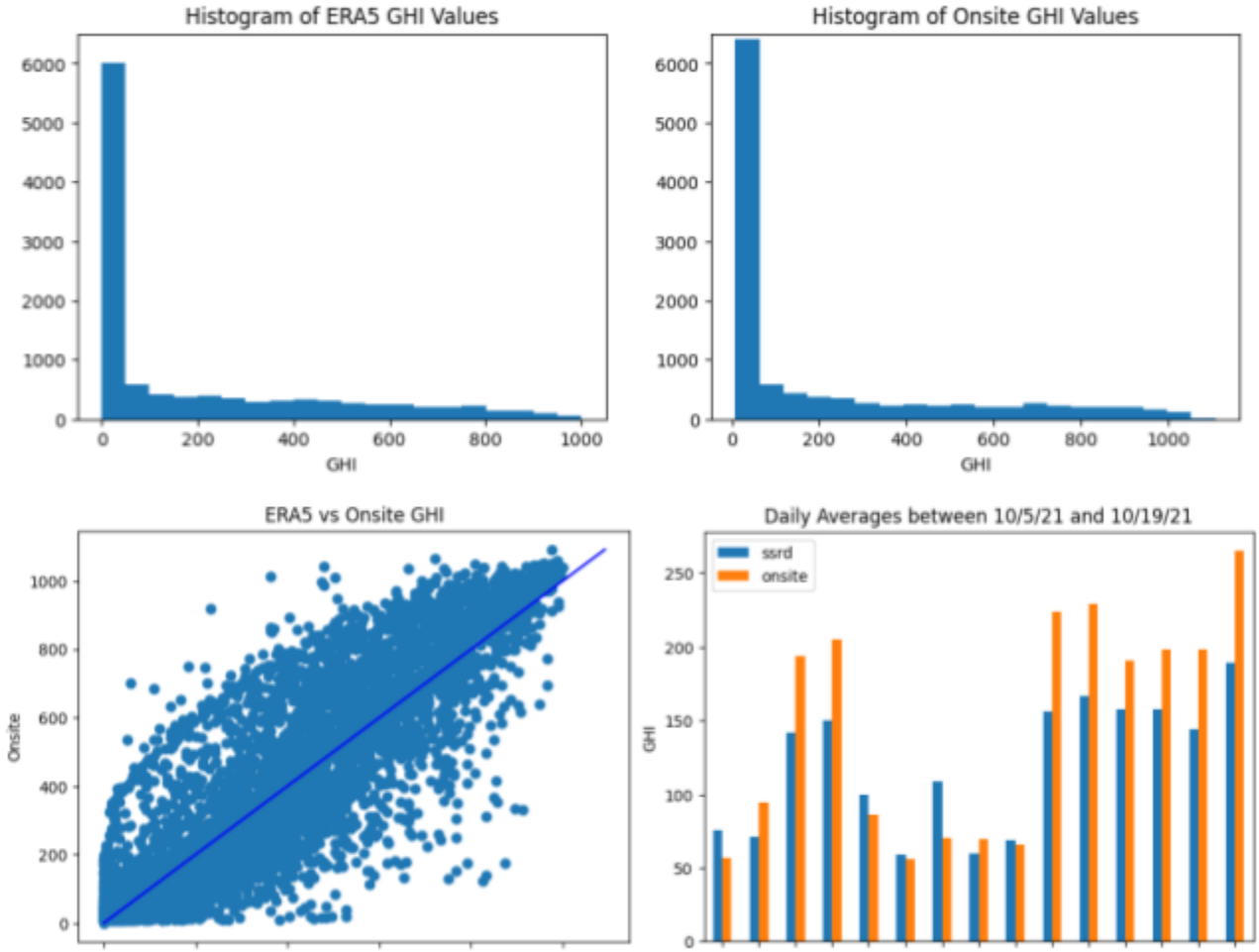


Figure 4: ERA5 vs Onsite GHI for Test Site

As shown in histograms in Figure 3, the input and output GHI distributions skew heavily toward the right due to the preponderance of low values during the night time in the Northern Hemisphere. As a result, ERA5 input values provide a strong baseline for predicting onsite

values after processing but fail to capture the maximums. On the left, the scatter plot depicts the relationship between the ERA5 input and onsite values after processing for the test site. The input and output values hold a correlation coefficient of 0.930417. This strong correlation belies the deviance for a given date or hour. To account for this skewness, the final adjustment to the dataset will be to strip the datasets of radiance values below 10 joules per square meter and calculate the residual. The resulting distributions can be seen in Figure 5 below. With these lower values removed, the correlation coefficient falls to 0.876271.

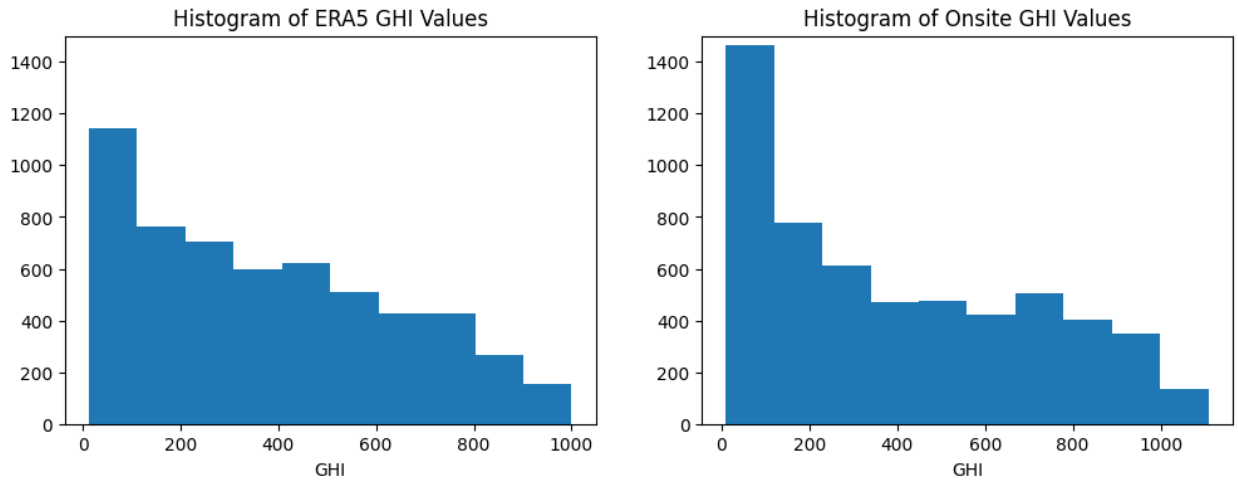


Figure 5: Histograms with GHI < 10 removed

Methodology

Applying two machine learning techniques and two loss functions, four approaches were evaluated for the purposes of this study:

1) Artificial Neural Network with Mean Squared Error Loss Function
2) Artificial Neural Network with Quantile Loss Function
3) Long Short-Term Memory Model with Mean Squared Error Loss Function
4) Long Short-Term Memory Model with Quantile Loss Function

While each model was trained on the same three datasets and tested on the fourth dataset for consistency, their parameters were tuned separately to maximize their potential.

Model Architecture

Artificial Neural Networks

An Artificial Neural Network (ANN) is a multi-level perceptron with at least one input layer, a hidden layer, and an output layer. The models trained for this study are constructed with two hidden layers and trained on a learning rate of 0.001, a minibatch size of 64, and a 50/50 validation split. The two ANN models differ on the number of neurons with the model used in conjunction with the Mean Squared Error loss function having layers with 16 neurons while the model used with the Quantile Loss Function is trained on layers of 64 neurons. The activation function for both models is the Rectified Linear Unit (ReLU) which passes along its value unless negative in which case it passes along a zero. Figure 6 visualizes the model architecture associated with the Mean Squared Error loss function.

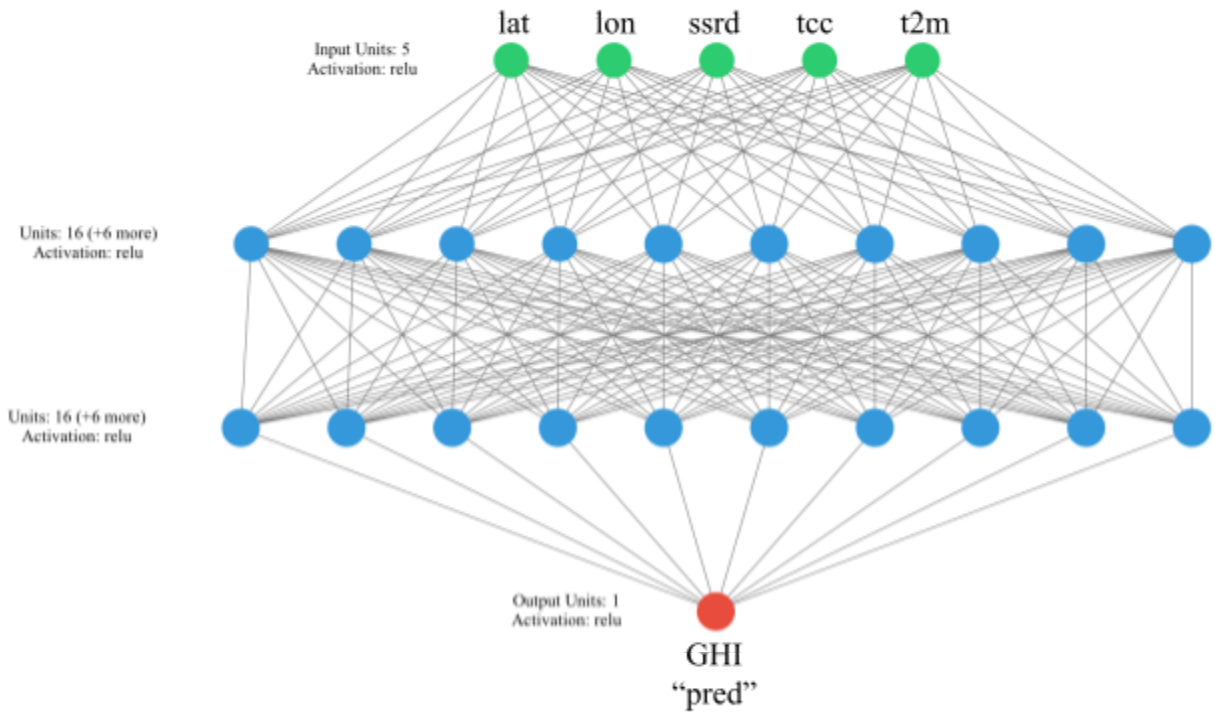


Figure 6: Artificial Neural Network Architecture with MSE Loss Function

Long Short-Term Memory Neural Networks

Long Short-Term Memory (LSTM) neural networks is a form of an artificial neural network that has feedback connections. As such, it operates as a recurrent neural network that can keep track of dependencies across time series. This function can support longer term trends that affect solar radiance accuracy such as the change in the length of sunlight over the course of a year and stormy weather events that extend cloud cover over multiple days. The LSTM neural network for

this study functions as a hybrid as it includes an additional dense layer that synthesizes weights from the prior memory layer. The Mean Squared Error loss function LSTM model is constructed with 64 neurons per layer while the Quantile loss function LSTM model has 32, primarily to reduce run times for the latter. Both models share a learning rate of 0.001, a minibatch size of 128, and an Rectified Linear Unit activation function. Figure 7 depicts the architecture for the LSTM model employed in this study.

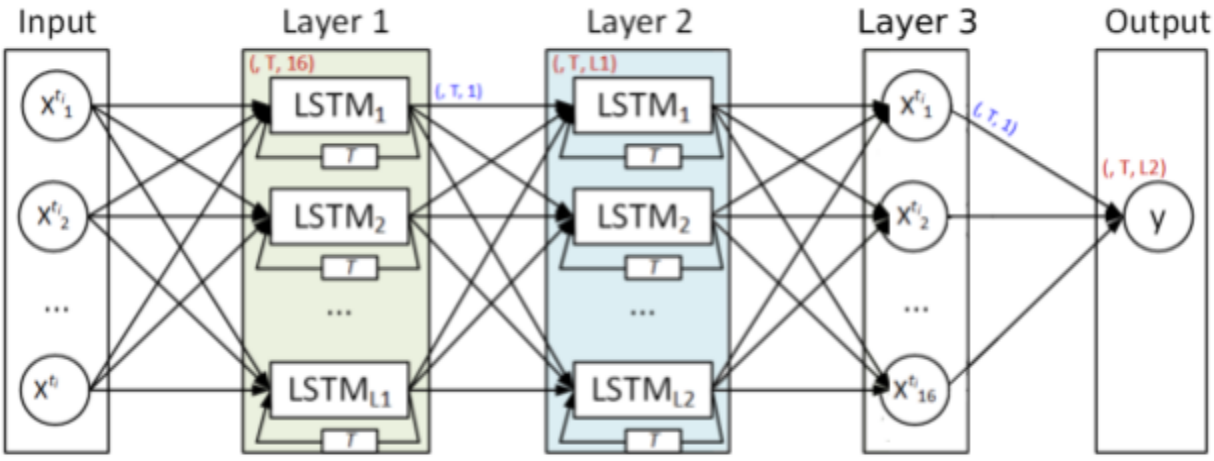


Figure 7: LSTM Neural Network Architecture

Loss Functions

Mean Squared Error

Mean Squared Error is the most common loss function for evaluating neural networks. The method takes the average of the errors squared, essentially capturing the weighted distance of each value from the mean. The formula for this function is:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

As the squaring function increases the penalty exponentially as values drift away from the mean, the mean squared error heavily penalizes outliers. Marginalizing outliers presents adverse consequences for the prediction of solar radiance in which most production occurs during peak seasonal conditions and most production is lost due to exceptional cloud events.

Quantile Loss

To address the penalization of outliers, this study considers Quantile Loss functions. A Quantile Loss function, also known as a Huber Loss function, minimizes the weight on outliers by weighting the distance from given quantiles as opposed to the mean. The formula for this function is as follows:

$$\sum_{k=0}^{n} \rho_q(y_k - p_k) \quad with \quad \rho_q(y) = y * (q - \mathbb{I}_{(y<0)})$$

Five quantiles (1%, 25%, 50%, 75%, and 99%) were selected for the ANN model while only one quantile (40%) was selected for the LTSM model loss function due to incongruencies between multiple quantiles and LTSM models that were beyond the scope of this study.

Results and Discussion

To compare the accuracy of models with differing loss functions, the root mean square error and correlation coefficient are calculated for each approach. The baseline for the evaluation of these neural networks is comparison against the GHI values derived from publicly available ERA5 data. The correlation coefficient and root mean squared error between the ERA5 data and the onsite data is 0.876271 and 151.29, respectively. Figure 8 captures the results of each approach with comparison to ERA5 data in the first row and the onsite data in the second row.

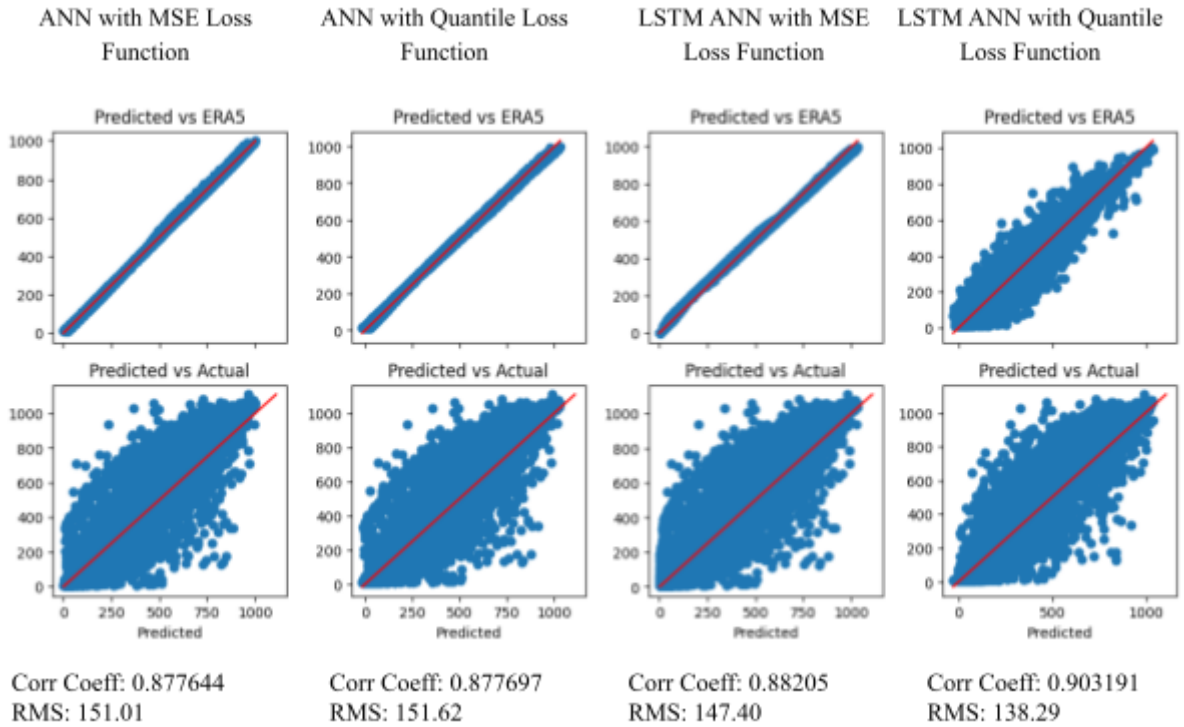


Figure 8: Comparison of Predicted Values, ERA5 Input Values and Onsite Output Values

While all approaches demonstrated an improvement over the base ERA5 data, the results show that the first three approaches made little change to inputs and therefore improved only slightly. The failure to deviate from the baseline is most likely driven by the choice to set the residual as the predicted output value. With the input data already heavily correlated to the output data, simpler models may require higher quantities of data to build confidence in deviating from the baseline.

The LSTM ANN with Quantile Loss model demonstrates the reduction in the root mean square error and highest correlation with the onsite pyranometer data. In aggregate, a 2.5% improvement in the accuracy of GHI measurements can equate to a difference of hundreds of thousands of kilowatt hours over the course of a year. This suggests that ANN LSTM models with Quantile Loss functions offer meaningful adjustments to publicly available solar data.

Yet when parsing data at an hourly level, a 2.5% increase in correlation coefficient is not statistically significant. Figure 9 plots the hourly data across the entire time period of the dataset for Waterford EMS. The model falls short in predicting GHI during the daylight limited seasons between October and March in the Northern Hemisphere. Additional input data on days from/to the summer solstice and the number of daylight hours may improve the model's ability to account for long term periodic trends.

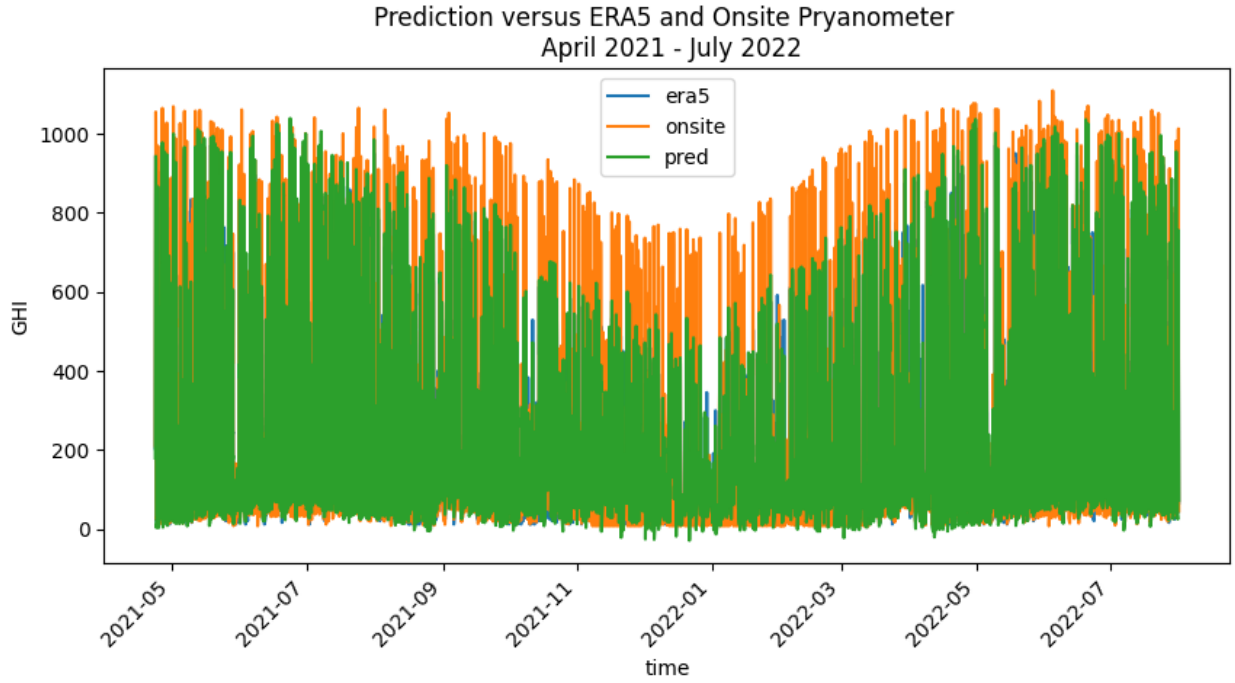


Figure 9: LTSM with Quantile Loss Model Prediction vs ERA5 and Onsite GHI

Conclusion

The transition to renewable energy has rapidly accelerated the need for accurate monitoring and reporting on solar radiance on photovoltaic power facilities in the United States. This paper evaluates the capabilities of ANNs with and without LTSM and Quantile Loss functions to improve the resolution of publicly available ERA5 weather datasets vis-a-vis pyranometers at four existing sites across the Northeast United States. The results of this study suggest that contemporary machine learning techniques, particularly ANNs with LTSM and a Quantile Loss function, can improve the accuracy of public satellite data. While hourly values are difficult for these models to estimate, the increasing availability of onsite solar data and improved parameterization offer the opportunity to improve on these models for future work.

References

"ERA5", European Centre for Medium-Range Weather Forecasts, https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5. Accessed December 22nd, 2022.

"Fact sheet: Reanalysis", ECMWF, 9 November 2020, Accessed December 22nd, 2022. https://www.ecmwf.int/en/about/media-centre/focus/2020/fact-sheet-reanalysis

"Solar radiation modeling," SolarGIS, https://solargis.com/docs/methodology/solar-radiation-modeling. Accessed December 22nd, 2022.

Yadav, Amit Kumar, and S. S. Chandel. "Solar radiation prediction using Artificial Neural Network techniques: A review." Renewable and sustainable energy reviews 33 (2014): 772-781.