AN EXPLORATION OF MACHINE LEARNING BASED DAY-AHEAD SOLAR IRRADIANCE FORECASTING METHODOLOGIES

by

AASHISH YADAVALLY

(Under the Direction of Frederick Maier)

ABSTRACT

Predicting solar irradiance is an important topic in renewable energy generation. In this work, the North American Mesoscale (NAM) Forecast System data is augmented with irradiance observations from the solar farm at the University of Georgia, towards forecasting 24 hours into the future. For the machine learning models used for this purpose, an input-selection scheme is presented and evaluated. This scheme significantly improved the performance, and resulted in a mean absolute error (MAE) of $72.63 W/m^2$, $44.94 W/m^2$ and $63.60 W/m^2$ for the dual-axis tracking, fixed-axis and single-axis tracking solar arrays respectively. The effect of geographic expansion, by including additional weather forecasts is evaluated. Furthermore, to correct the reported bias in global horizontal irradiance (GHI) in NAM Forecast System, theory-driven bias-correction approaches are explored, where NAM Forecast System is selectively combined with *Clear-Sky Scaling* and *Liu-Jordan* techniques. In addition, the ability of predictive models involving clear-sky index to capture seasonal patterns is evaluated.

INDEX WORDS: solar forecasting, machine learning, numerical weather prediction, bias correction, clear-sky index, clearness index

AN EXPLORATION OF MACHINE LEARNING BASED DAY-AHEAD SOLAR IRRADIANCE FORECASTING METHODOLOGIES

by

AASHISH YADAVALLY

B.Tech., Indian Institute of Information Technology Vadodara, INDIA, 2018

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2020

AN EXPLORATION OF MACHINE LEARNING BASED DAY-AHEAD SOLAR IRRADIANCE FORECASTING METHODOLOGIES

by

AASHISH YADAVALLY

Major Professor: Frederick Maier

Committee: Khaled Rasheed
Sheng Li

Electronic Version Approved:
Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
August 2020

ACKNOWLEDGEMENTS

I would like to thank all my committee members for their time and the numerous suggestions they provided throughout my research. In particular, I would like to thank Dr. Maier for his guidance, constant encouragement and motivation to keep me going. I owe a great debt to Zachary Jones and Chris Barrick who facilitated my introduction into the project, and whose contributions laid the foundation for my work. I am thankful to my friends and family who kept me sane during these unprecedented times, where the whole world has come to a standstill due to COVID-19. I would like to thank everyone at the *Institute for Artificial Intelligence* who presented me with an opportunity to be a part of this project. I would also like to thank Ms. Tino, whose assistance with the administrative tasks at the university made my life incredibly easy.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Fossil fuels are the dominant cause of climate change, and transitioning from energy sources depending on them to renewable forms is one of the most powerful ways in which we can reduce our ecological footprint as a society. However, owing to the unpredictability associated with carbon-free sources such as solar energy and wind energy, incorporating them into an electrical energy system is challenging. Thus, to ensure a balance between the consumption and production of solar energy, accurate prediction of solar irradiance is of utmost importance.

Solar irradiance forecasting can be performed by several methods depending on the temporal variability of the forecast horizon, ranging from a minute to several days [44]. Generally, the predictions worsen as the forecast horizon increases. Numerical weather prediction (NWP) models utilize mathematical models of atmosphere and ocean systems to predict weather variables from hours to months in advance. Thus, using them for day-ahead solar forecasting is a common strategy. The National Oceanic and Atmospheric Administration (NOAA) maintains various global and mesoscale weather prediction models for weather forecasting. Mesoscale models are three-dimensional regional models based on thermodynamic equations describing physical processes, which incorporate the inherent unpredictability of many small-scale phenomena. The North American Mesoscale (NAM) Forecast System [12] is one of the major mesoscale-based weather forecast models maintained by NOAA.

The specific purpose of this work is to develop machine learning models to effectively predict surface-level solar irradiance 24 hours into the future at multiple fixed and tracking solar arrays located at a solar farm near the University of Georgia. The developed models rely heavily on the use of NAM weather forecasts, which are released four times a day at six-hour intervals (00h, 06h, 12h, 18h UTC) for a grid of $12km$ x $12km$ cells covering the continental United States. NAM forecasts from the years 2017 and 2018 were used to develop and evaluate these models.

The current work extends prior work undertaken at the University of Georgia related to solar irradiance forecasting (which have been discussed in Chapter 2). Of special importance is the work of Jones [58], who developed a machine learning pipeline to forecast day-ahead solar irradiance using the weather forecast data from the NAM Forecast System. Chapter 3 begins by replicating these results. The models developed by Jones [58] utilized nine weather variables from the NAM data. Importantly, however, *total cloud cover*, which is defined as the fraction of the sky covered by visible clouds was not considered. Cloudiness is considered to be an important meteorological factor in determining the amount of solar radiation reaching earth's surface. Thus, all the ten weather variables were further analyzed with respect to the target irradiance observations recorded at the solar farm, intending to quantify their usefulness in the day-ahead irradiance prediction. It was found that surface temperature, global horizontal irradiance, total cloud cover and atmospheric height were more significant than others.

The NAM Forecast System predicts values for weather variables 84 hours into the future. These are referred to as feature projections. The first 37 feature projections are at a one-hour temporal resolution (starting at the zero-hour or reference time). Feature projections corresponding to the subsequent 48 hours are reported at a three-hour temporal resolution. In his work, Jones [58] used all 37 feature projections at a one-hour temporal resolution, for the nine weather variables as predictors to the machine learning models.

In the current work, it was hypothesized that this was unnecessary and needlessly increases model training time. To investigate this, the relationship between the first 25 feature projections of GHI in NAM forecasts (starting at the zero-hour or reference time), was studied with respect to the irradiance observations from the fixed-axis solar array in the forecast horizon. Using a mutual information matrix, we concluded that the target irradiance observations for a particular target hour did not depend on all the feature projections of the NAM weather variable. Because of this, only data from 6 hours prior to the target hour to 6 hours ahead of the target hour were chosen as predictors in the models developed. This input-selection scheme helped in reducing the computational cost of training the machine learning models significantly.

This input-selection scheme was used in a series of experiments to develop models analogous to those of Jones [58]. With respect to the performance of the latter, an average improvement in performance (across different machine learning models) by 19.05%, 19.68% and 10.65% was recorded for the dual-axis tracking, fixed-axis and single-axis tracking solar arrays. We attribute this improve-

ment in performance to the weeding out of less relevant features with the help of the input-selection scheme. The best performance was recorded using the *random forest* machine learning technique, with a mean absolute error (MAE) of 72.63 W/m^2, 44.94 W/m^2 and 63.60 W/m^2 respectively for each of the solar arrays.

Another series of experiments was performed to test the effect of geographic expansion, i.e. including a larger geographic area from the NAM forecast as input. The NAM Forecast System generates multiple grids of weather forecasts, where each cell corresponds to a $12km$ x $12km$ geographical area. In [58], Jones included NAM forecasts corresponding to a grid of cells surrounding the cell representing Athens. He noted that considering a 3 x 3 grid of cells was optimal, as the improvement in performance diminished for even greater grid sizes. We also investigated the significance of similar geographic expansion, but on the NAM data obtained by incorporating the input-selection scheme. It is observed that such an expansion does not significantly improve the solar irradiance predictions. In fact, in a few cases, the geographic expansion had a detrimental effect on the performance of the machine learning models.

Beyond their overall performance, the performance of the machine learning models was investigated in a stratified manner. The predictions of each of the 00h, 06h, 12h and 18h NAM weather forecasts were analyzed independently. For each, a *diurnal* analysis was performed, wherein the performance of the models for each target hour in the forecast horizon was compared. Such an analysis helped in understanding the ability of the machine learning models to gauge the diurnal characteristics.

Among the variables modelled by the NAM Forecast System is *downward short-wave radiation flux* (also called global horizontal irradiance, GHI [46]). It is an estimate of the total amount of short-wave radiation that reaches the Earth's surface, and is essential for short-term solar forecasting. It has been reported in literature that the NAM Forecast System tends to overpredict GHI when visible clouds are not present in the sky [46]. In order to correct such a bias, identifying the amount of clouds in the sky becomes essential. This can be estimated using empirical solar radiation models, which formulate relations between different meteorological variables through experimental observations. In Chapter 4, theory-driven bias correction methodologies were explored, which involved blending the physical NAM Forecast System with such solar radiation models, so as to selectively correct the bias in GHI.

From among the different empirical formulations proposed to estimate GHI from environmental conditions, *Clear-sky Scaling* [52] and *Liu-Jordan* [55] techniques were studied. For the purpose of correcting the bias in the GHI estimates from the NAM Forecast System, the GHI retrieved through each of these empirical solar radiation models was combined with the GHI from the NAM Forecast System depending on metrics such as *clear-sky index* and *clearness index*, which are different measures for estimating the amount of cloudiness in the sky.

A series of experiments was conducted using the *random forest* algorithm to test the effectiveness of such a model-blending approach. Three variants of NAM data was input to these machine learning models: GHI from the NAM Forecast System (GHI_{NAM}); adjusted GHI, obtained from blending NAM Forecast System with *Clear-Sky Scaling* technique ($\mathrm{GHI}_{\mathrm{NAM+CS}}$); adjusted GHI, obtained by blending NAM Forecast System with *Liu-Jordan* model ($\mathrm{GHI}_{\mathrm{NAM+LJ}}$). The performance of the *random forests* utilizing each of the adjusted GHI variants was compared with *random forests* utilizing GHI_{NAM}. It was observed that the blending methodology involving NAM Forecast System and *Clear-Sky Scaling* resulted in an improvement in performance (decrease in MAE) by 4.95%, 4.53% and 4.12% for the dual-axis tracking, fixed-axis and single-axis tracking solar arrays respectively. In comparison, the blending methodology involving NAM Forecast System and *Liu-Jordan* model recorded an improvement in performance by 4.17%, 4.14% and 3.62% for each of the solar arrays.

The above experiments were only based on GHI, however. That is, GHI (corrected or uncorrected) was the only predictor used to develop these models. A new series of experiments was conducted by including the other weather variables from NAM Forecast System such as *air temperature*, *total cloud cover* and *atmospheric height* along with the three variants of GHI used in the previous set of experiments. The input-selection scheme described earlier was incorporated into this weather forecast data, and select feature projections depending on the target hour in forecast horizon was selected for each of these weather variables. In this case however, the model-blending methodology involving NAM Forecast System performed slightly better than both of the blending methodologies. For the model-blending methodology involving *Clear-Sky Scaling* technique, an MAE of 72.57 W/m^2, 44.91 W/m^2 and 63.56 W/m^2 was recorded for the dual-axis tracking, fixed-axis and single-axis tracking solar arrays respectively. For the blending methodology involving *Liu-Jordan* model, an MAE of 72.74 W/m^2, 45.25 W/m^2 and 63.97 W/m^2 was recorded for each of the solar arrays.

Clear-sky index (which is computed using various meteorological variables such as GHI, solar position and solar zenith angle) is known to capture the diurnal and seasonal trends in weather data effectively. Because of this, it was suspected that the clear-sky index might improve model performance in comparison to using GHI. Thus, another series of experiments was conducted by developing predictive models utilizing *clear-sky index* (in place of GHI), along with air temperature, total cloud cover and atmospheric height. Select feature projections of each of these weather variables were used as predictors to the machine learning models.

However, it was observed that the performance of these predictive models paled in comparison to those utilizing the input-selected weather forecast data including GHI_{NAM}. An MAE of 79.58 W/m^2, 49.18 W/m^2 and 69.98 W/m^2 was recorded for the dual-axis tracking, fixed-axis and single-axis tracking solar arrays respectively. An analysis of the performance of individual 12h and 18h NAM forecasts was conducted. The MAE corresponding to the *spring* and *summer* season increased for the both the forecasts, with respect to the predictive models utilizing GHI_{NAM}. Consequently, it can be concluded that the presumed ability of *clear-sky index* to capture the diurnal and seasonal trends did not translate into improving the performance of the predictive models.

The theory-driven bias correction methodology undertaken in this work solely corrects the bias in GHI, and doesn't address the bias correction in other weather variables. In addition, the lack of improvement in the performance of predictive models upon including additional weather variables possibly indicates the inability of the models to identify the clear-sky conditions effectively. This, in turn prevents accurate bias correction in GHI. Future work can explore superior approaches for detecting clear-sky conditions, which will improve upon the bias correction in GHI, as well as other weather variables. Furthermore, it was observed that the predictive models utilizing clear-sky index performed worse than those utilizing GHI across all seasons in the year. We note that there are other clear-sky models in literature which can conceivably improve the ability of clear-sky index to capture such seasonal trends. These models can be investigated in further work.

CHAPTER 2

LITERATURE REVIEW

High variability in solar radiation necessarily results in variability in the output of photovoltaic (PV) power plants. What this essentially means is that, in order to effectively integrate PV systems into a larger electrical grid (which must compensate for the variable output of PV systems to ensure overall stability), effective prediction of solar irradiance is needed. In the last few decades, solar forecasting researchers have developed a variety of data-driven approaches to improve solar irradiance forecasting. Most of these approaches can be categorized based on the resolution of the forecast horizon, ranging from a few minutes to a couple of days, weeks or months; and the spatial resolution of the input data pertaining to a particular location.

For solar irradiance forecasts up to < 30 minutes ahead, a variety of techniques based on the ground-to-sky imagers have been explored. The spatial resolution for the total sky imagery is in the range of 10m - 100m. Using the sky images taken every 30 seconds, Chow et al [9] presented a method for determining sky cover and solar irradiance nowcasting. Marquez et al. [10] used image-processing techniques to calculate velocity fields and classify clouds in individual grids so as to employ it to forecast Direct Normal Irradiance (DNI), an essential component of global irradiance, for time horizons ranging from 3 minutes to 15 minutes.

Statistical time-series models such as Auto Regressive Integrated Moving Average (ARIMA) and non-linear model approximators such as Artificial Neural Networks (ANNs) have been shown to be more effective for forecasting solar irradiance up to 2 hours ahead [36]. Marquez and Coimbra [37] successfully used meteorological variables from US National Weather Service (NWS) Forecasting database as inputs to an ANN model for forecasting global and direct solar irradiance. In [38], Reikard reviewed a variety of time-series modeling techniques for predicting solar irradiance, and observed that the ARIMA models, in general, had the best forecasting results. However, as Lopez et al. [39] note, these developed models are not transferable to different locations, especially ones with varying cloudiness properties (this certainly applies to the work at UGA as well).

Input data from satellite imagery tracking cloud motion has been shown to be useful for an *intra-day* forecast horizon between 1 hour and 24 hours. The geostationary satellites detect clouds with the help of visible and infra-red images, which generally have a spatial resolution of ∼1 km. Various methods such as Heliosat-I [32], Heliosat-II [33] and Heliosat-III [34] have been implemented which employ motion-vector fields to track the clouds using these images. By applying the calculated motion vector fields on the actual image, the cloud index images can be determined. Hammer et al. [35] employed Heliosat-I technique on such cloud index images for forecasting solar irradiance ∼30 minutes to 6 hours in future. Cloudiness has a significant impact on the surface solar irradiance, and the basis of this methodology relies upon the determination of cloud structures.

For a *days-ahead* solar forecasting range which is essential for utility applications, the knowledge of the meteorological weather parameters in that period is paramount. Numerical Weather Prediction (NWP) models are physical models which make use of the current meteorological conditions and predict weather conditions days into the future, on the basis of atmospheric equations. Notably, NWP models are able to forecast up to two days ahead or beyond, depending on the spatial domain of the model. Examples of the NWP models maintained by the National Oceanic and Atmospheric Administration (NOAA), which record data across different spatial resolutions, across varying geographical expanse are the Global Forecast System (GFS) [11], North American Mesoscale (NAM) [12], Rapid Refresh (RAP) and High Resolution Rapid Refresh (HRRR).

Several researchers have concentrated their efforts on comparing the effectiveness of each of the NWP models for solar irradiance forecasting purposes at various locations. Mathiesen and Kleissl [13] compared the irradiance parameter forecast in NWP models such as NAM, GFS and ECMWF within the continental United States, with respect to solar forecasting. In this work, they extensively studied the predictions using each of the NWP models in varying cloud conditions, establishing a database to validate numerical weather predictions. In [14], Ruiz-Arias et al. found that the NWP models based solar irradiance forecasting *significantly outperforms* satellite-based methodologies while forecasting 6 hours and beyond, and attributed it to the effective simulation of weather parameters of the entire atmospheric system in the NWP models.

Lorenz et al. [15] performed benchmarking studies to gauge the reliability of different solar irradiance forecasting approaches. They investigated the seasonal dependence of forecast errors using several techniques. They concluded that post-processing the weather parameters in the NWP models significantly captures the dependence between forecast accuracy and climatic conditions. Perez et al.

[16] validated the performance of the NWP models across seven stations in the SURFRAD network. They extracted the hourly GHI forecasts by time-interpolating the 3-hour and 6-hour cloud cover parameter forecasts in the NWP models, and further adjusting them using sky-cover-to-irradiance fits. In this work, the authors explored a diverse set of climatic environments and noted that the models' performance in winters tends to be poorer than in summers. They also concluded that forecasts from the one-hour time-interpolated data are on par or better than the forecasts from the satellite-imagery based data for a forecast horizon up to 5 hours.

In [13], Mathiesan and Kleissl also infer that the NWP models like GFS and NAM are biased towards forecasting clear conditions resulting in large biases in global horizontal irradiance (GHI) parameter in these conditions. They obtain the metric Mean Bias Error (MBE) for each NWP model based on the solar zenith angle and the clear sky index (k_c) metric, which is the ratio of the measured GHI in the model to the clear sky GHI. However, like Diagne et al. [17] note, the methodology used by Mathiesan and Kliessl was not adequate, as they did not present information about the bias source, which is important to selectively correct forecasts. They observed that these bias corrections did not help in reducing the Root Mean Squared Error (RMSE) metric, as even the accurate forecasts were unnecessarily corrected - indicating a need for a better approach for GHI bias corrections in the NWP models.

The effectiveness of physical NWP models such as the NAM Forecast System for the purpose of day-ahead solar irradiance forecasting, demonstrated in the aforementioned works, prompted us to study and analyze it further. Accurate identification of clear-sky conditions is essential towards selectively correcting the bias in GHI [18]. Such an identification requires additional measurements such as direct or diffuse components of global solar irradiance, which is not measured by the NAM Forecast System. However, these irradiance components can be estimated by means of empirical solar radiation models. The main difference between physical models such as the NAM Forecast System and empirical solar radiation models is that the former parameterizes cloud microphysics through mathematical equations, while the latter formulates the relation between meteorological variables derived through experimental observations.

Empirical Solar Radiation Models

Several empirical formulations have been proposed in literature which help predict the irradiance metrics such as global horizontal irradiance (GHI), diffuse horizontal irradiance (DHI) and

direct normal irradiance (DNI) from atmospheric properties. GHI is the total amount of shortwave radiation reaching a surface horizontal to the ground, and is the most useful solar radiation data parameter. In comparison, DHI is the part of global solar radiation which passes through the atmosphere, and is absorbed, scattered or reflected by the gases in the atmosphere. DNI is the amount of solar radiation received by a surface that is held normal to the rays from the sun. It needs emphasis because of the sharp shadows that it can extend on the surface of the earth. Irradiance on the surface of a solar cell can be determined with the help of DNI, and thus, proper estimation of DNI is of high importance in Concentrated Solar Power (CSP) systems [24].

The empirical solar radiation models to estimate the sky (DHI) and beam (DNI) components of global solar radiation can be categorized into *parametric* and *decomposition* models [25]. Parametric models require detailed information about atmospheric conditions such as turbidity, cloud cover, precipitable water content, etc. to be able to calculate DHI and DNI. In comparison, the decomposition models formulate empirical equations to estimate DHI and DNI from GHI, based on the correlations between each of the components. The parametric models are a better alternative to decomposition models only in cases where the meteorological data is not available [25][26].

Radiative Transfer Models (RTMs) help in simulating the radiative transfer of electromagnetic radiation through a planetary atmosphere, and thus help in estimating solar irradiance. However, they are computationally expensive to maintain. The clear-sky solar irradiance parametric models provide relatively simple parameterizations to estimate solar irradiance in conditions with less visible clouds [30]. The aerosols and water vapour present in the atmosphere play a significant role in scattering the sunlight, and have an impact on the amount of solar radiation reaching the Earth's surface. Thus, solar forecasting researchers concentrated their efforts towards estimating GHI in clear-sky conditions, i.e, conditions where visible clouds are negligible, and further scaling this parameter across cloud conditions.

Bird and Hulstrom [31] proposed the *Bird Clear Sky Model* based on comparisons of results from the radiative transfer codes, to estimate clear sky direct beam, hemispherical diffuse, and total hemispherical solar radiation on a horizontal surface. However, one of the drawbacks with this model is that various atmospheric parameters such as aerosol optical depth, ozone and water vapour are fixed for an entire year. Gueymard [19] proposed the *REST2 Clear Sky Model* which specifically accounts the effects of aerosols to predict cloudless-sky broadband irradiance. The REST2 model represents broadband components of two different spectra, and incorporates the transmission esti-

mates for each of the spectra separately. Finally, the total diffuse radiation on a horizontal surface is aggregated from the estimates of both the spectra.

Ineichen and Perez [23] proposed a new airmass independent formulation to estimate the Linke Turbidity coefficient, thus removing it's dependency on solar geometry, and used the coefficient to develop two clear-sky models to estimate global and direct normal irradiance. Furthermore, Ineichen [20] modified the original version of the *Solis Clear Sky Model* proposed by Mueller et al [21] to accommodate the circumstances in which spectral computations aren't possible, by introducing a broadband version of the algorithm.

The decomposition models are formulated on the basis of the *clearness index* (k_t) parameter, which is the ratio of the measured solar radiation to the extraterrestrial solar radiation. A higher value ($k_t \rightarrow 1$) of the clearness index parameter indicates that the atmosphere is clear, while a lower value ($k_t \rightarrow 0$) of the clearness index parameter indicates that the atmosphere is cloudy. Chandrasekaran and Kumar [27] collected data in Madras, India to formulate a fourth-order polynomial correlation depending on the clearness index parameter to estimate the irradiance metrics in a tropical setting [28].

By analyzing the data collected across multiple locations in the United States and Canada, *Liu and Jordan* [29] formulated an empirical equation on the basis of the clearness index parameter to estimate the irradiance metrics. Maleki et al [25] reviewed various solar radiation models, and observed that the Liu and Jordan model is very effective in estimating diffuse radiation on inclined surfaces.

Recent Work on Solar Irradiance Forecasting at the University of Georgia

"Georgia Power", a regional utility company, recently partnered with the University of Georgia (UGA) to set up a 1MW solar facility in Athens, GA. The facility consists of multiple fixed and tracking (single-axis and dual-axis) solar arrays. Recent work at UGA has been devoted to analyzing or predicting solar radiation based upon the data provided by the facility and the GAEMN (Georgia Automated Environmental Monitoring Network) weather station network.

In [49], Sanders investigated the importance of different weather variable observations in the prediction of solar irradiance. From among the sixteen weather variables recorded by GAEMN, he obtained current and historical weather information for the following weather variables which

are typically known to affect solar irradiance: air temperature, precipitation rate, visibility, wind speed, wind direction, dew point temperature, air pressure and relative humidity. Utilizing the solar radiation data from GAEMN, they built predictive models including current weather observations, weather forecasts from NWP models for the target location, and additional weather forecasts from NWP models for area surrounding the target location.

In addition, they obtained NWP model predictions for these weather variables at the target locations to analyze the effect of using NWP model predictions for these variables as a means of forecasting solar radiation over one-hour and 24-hour time frames. Upon including the weather forecast data in the predictive models, it was observed that there was a reduction in mean absolute error (MAE) for 1-hour predictions by 7.6% and for 24-hour predictions 40.2%. They noted that the incorporation of weather forecasts from NWP Forecast System is extremely important in solar forecasting, especially over a longer time horizon.

Sanders [49] performed work based on Lorenz et al. [1], who found that expanding the forecast region to approximately $100km$ x $100km$ and performing a spatial averaging across the region resulted in an improvement in day-ahead solar forecasting. Larson et al. [53] noted that the solar radiation predictive models which use NWP data can usually be improved by averaging the GHI forecasts from NWP grid points surrounding the target location. Sanders [49] validated these findings by including weather forecasts from areas surrounding the target location, and found that it was extremely beneficial, especially while forecasting over a longer time horizon, where the weather system is less predictable. This was performed by including forecasted weather variables from the NWP cells lying to the north-west, north, north-east, east, south-east, south, south-west and south, resulting in eight additional parameters for each weather variable. They found that such a methodology led to an increase in predictive accuracy in both one-hour ad 24-hour solar radiation predictions.

Jones [58] extended the work in [49] by developing machine learning models for hourly targets from 1 - 24 hours, using GAEMN observational data and NWP predictions (Rapid Refresh, RAP and North American Mesoscale, NAM). He observed that the forecasted weather variables from NWP models became more important for target hours beyond a very short forecast horizon. In addition, utilizing the solar irradiance observations from the solar farm, they presented a case-study in irradiance prediction by augmenting these observations along with the NAM data. They explored the geographic expansion of forecast coverage by including the NAM weather forecasts from a grid of cells around the NAM data grid representing Athens, towards obtaining a 3 x 3 geographical grid

shape, representing a geographical expanse of 36km x 36km. It was observed that using the 3 x 3 grid shape was optimal for solar irradiance forecasting at the farm, and it improved the accuracy significantly as compared to just considering weather forecasts from the NAM data grid representing Athens. Using the 3 x 3 grid shape, they achieved the best accuracy with an MAE of 47.6 W/m^2 for the fixed-axis solar array, 58.7 W/m^2 for single-axis tracking solar array and 75.4 W/m^2 for dual-axis tracking solar arrays.

Jones [58] attempted to quantify the improvement in accuracy of predictive irradiance models as a result of the expansion of forecast coverage. Two of the best-performing machine learning models for the 3 x 3 grid shape, *k-Nearest Neighbors* and *Random Forests* were retrained for 5 x 5 and 7 x 7 grid shapes as well, representing a geographical expanse of 60km x 60km and 84km x 84km respectively, and their performance in MAE and R^2 were recorded. They realized that the benefits from a wider geographic coverage of forecasted weather variables resulted in diminishing returns as the size grows larger. While the improvement due to using 3 x 3 grid shape over 1 x 1 grid shape (representing Athens) was significant (~16%), those for 5 x 5 and 7 x 7 grid shapes over 3 x 3 was negligible.

CHAPTER 3

SOLAR IRRADIANCE FORECASTING USING NUMERICAL WEATHER PREDICTION MODELS

3.1 OVERVIEW

For a days-ahead forecast horizon, utilizing Numerical Weather Prediction (NWP) models, which predict the evolution of the atmospheric system have been shown to be more useful [58]. The NWP models derive their initial conditions from different ground and airborne sensors from across the world. Based on thermodynamic equations describing the physical processes occurring in atmosphere, they forecast different weather variables into the forecast horizon. The National Oceanic and Atmospheric Administration (NOAA) operates a variety of NWP models with their spatial resolution ranging from approximately 10 km - 50 km, and their temporal resolution typically being 1 hour or 3 hours [40].

Solar forecasting researchers have successfully employed meteorological forecasts from NWP models for forecasting applications for years. The making of a weather forecast involves assessing the current weather situation, assimilating observational information, and projecting this initial state into the future based on the laws of thermodynamics. Weather forecasting employs a set of equations that describe the flow of fluids, being run over a geographic area. Several parameterizations of physical processes are carried out, based on the physical and statistical representations of the physical process. This is useful to approximate the bulk effects of the physical processes.

One of the major challenges faced in this process is determining the range of area to observe. The further the forecasting of the weather conditions, i.e, higher the forecast horizon, wider is the range of area that needs to be observed. Multiple weather prediction models, both global and regional, depending on the spatial domain, are maintained by the National Oceanic and Atmospheric Administration (NOAA). Global Forecast System (GFS) is one of the widely-known global weather prediction models, which represents the atmospheric state as a superposition

of wave functions. It covers the entire globe at a base horizontal resolution of $28km$ between grid points, predicting weather out to 16 days. Within continental United States, North American Mesoscale (NAM), Rapid Refresh (RAP), High Resolution Rapid Refresh (HRRR) are the popular regional weather prediction models, each having it's own advantages. The NAM Forecast System follows a complex cloud prediction scheme accounting for the internal cloud processes, and thus has better cloud parameterizations over RAP and HRRR.

In this work, we developed machine learning models to forecast solar irradiance on multiple dual-axis tracking (array A), fixed-axis (array B) and single-axis tracking (array E) solar arrays. The irradiance predictions were made 24 hours into the future, at a one-hour temporal resolution. The NAM Forecast System, which can predict parameters describing cloudiness [45], was input to these predictive models. A weather forecast dataset spanning NAM data for the years 2017 and 2018 was created, though a few forecasts are missing sporadically.

In order to gauge the effect of different weather variables specified by the NAM Forecast System on the solar irradiance predictions, the following were evaluated: air temperature, geopotential height, cloud cover, visibility, wind speed, dew point temperature, air pressure, downward shortwave radiation flux, downward longwave radiation flux, and humidity. Using *random forests*, the more relevant of these weather variables were identified. It was observed that the irradiance readings from the solar farm for each of the arrays were influenced more by the surface temperature, downward short-wave radiation flux, total cloud cover, and atmospheric height. Thus, the remaining weather variables were discarded. This enabled a cut in computational cost of modeling and also led to an improvement in the performance of the models.

Each of the weather variables in the NAM Forecast System are projected 36 hours into the future, at a 1-hour temporal resolution. Of these, the effect of the first 24 feature projections on the target solar irradiance was analyzed based on the *mutual information* statistical measure. It was found that the weather variable for a particular target hour offset in the forecast horizon depends on only 6 feature projections following the target hour offset, and 6 feature projections preceding the target hour offset. Separate machine learning models were trained, and the efficacy of the proposed input-selection scheme was tested.

The performance of this input-selection scheme was compared with the methodology followed by Jones [58]. It was observed that the input-selection scheme resulted in an average improvement in mean absolute error (MAE) across machine learning models by 19.05%, 19.68% and 10.65% for

the dual-axis tracking, fixed-axis and single-axis solar arrays respectively. *Random forests* achieved a best performance for such predictive models, with an MAE of 72.63 W/m^2, 44.94 W/m^2 and 63.60 W/m^2 for each of the arrays.

The effect of the geographic expansion of weather forecast coverage was analyzed by including the 3 x 3 and 5 x 5 *geo shapes*, wherein weather forecasts data from eight and twenty four NAM data grid cells centered around the NAM data grid representing Athens, Georgia were incorporated respectively. Such a spatial expansion was assessed for the attribute-selected weather forecast data, obtained by incorporating the input-selection scheme. It was found that the geographic expansion had a detrimental effect on the performance of the machine learning models, and had a marginal improvement with respect to the 1 x 1 *geo shape* only for random forests, with the 5 x 5 *geo shape* resulting in MAE of 69.38 W/m^2, 43.62 W/m^2, 61.99 W/m^2 for the dual-axis tracking, fixed-axis and single-axis tracking solar arrays respectively.

3.2 NORTH AMERICAN MESOSCALE (NAM) WEATHER PREDICTION MODEL

The North American Mesoscale (NAM) Forecast System is based on the Weather Research and Forecasting (WRF) model infrastructure, following non-hydrostatic dynamics and thus enabling vertical momentum estimations. It provides high resolution forecasts over North America for a forecast horizon of 84 hours, the first 36 of which are at a one hour temporal resolution, and the remaining thereafter, at a 3 hour temporal resolution. The forecasts are published for a grid spanning approximately $12km$ x $12km$ across the continental United States, which are released four times daily at 00h, 06h, 12h and 18h UTC.

In general, the NWP models cannot realize the physical phenomenon occurring within an individual grid. Vertical redistribution of heat and moisture can easily occur between mesoscale grids resulting in sub grid-scale variations in convection. The NAM Forecast System repeatedly nudge the temperature and moisture profiles in a grid towards decreasing the convective instability. Moreover, the wider forecast horizon of the NAM Forecast System requires it to model the radiative properties within the clouds effectively [47]. The NAM models handle this by implementing columnar Radiative Transfer Models (RTMs), which parameterize the cloud properties for every vertical level individually.

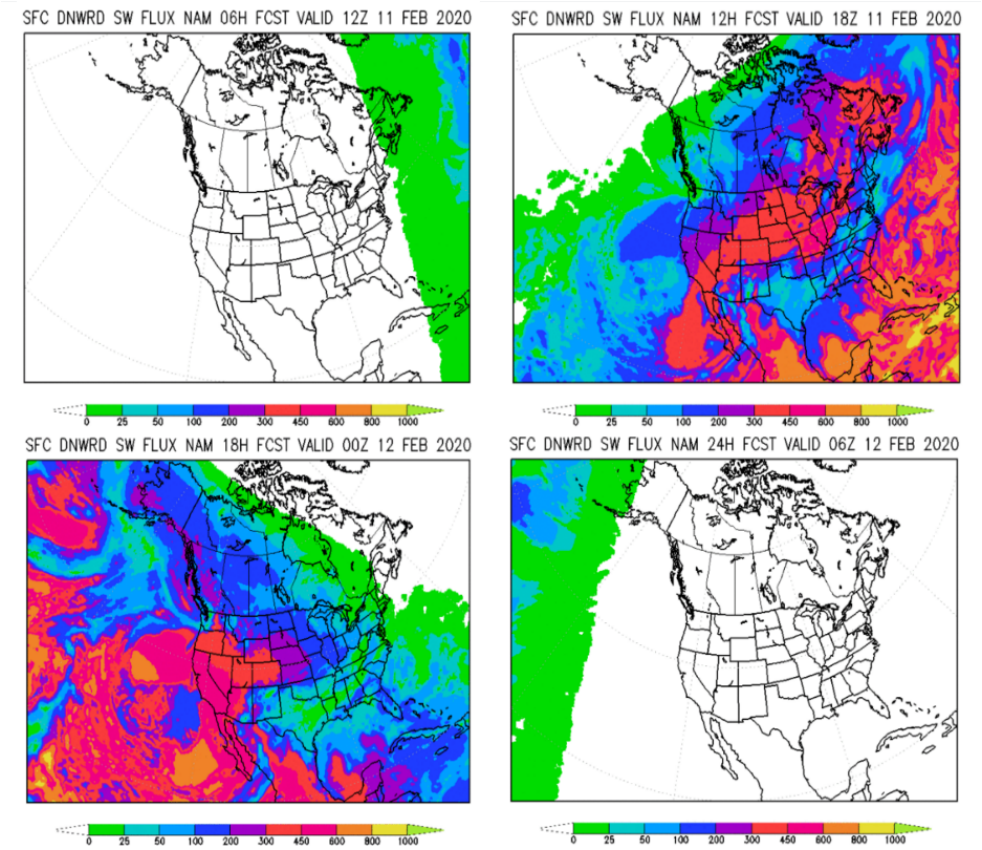


Figure 1: Downward Shortwave Radiation Flux parameter from NAM data over North America domain for 06h forecast (top-left), 12h forecast (top-right), 18h forecast (bottom-left) and 24h forecast (bottom-right) UTC for 11th February, 2020.

Dozens of weather variables are available in a NAM model data grid pertaining to environmental components such as altitude, atmospheric pressure, atmospheric radiation, air temperature, water vapour, atmospheric winds, precipitation, soil properties and cloud cover. Each of these are spread across 60 vertical levels in a 0 - 3 km layer, and across 39 pressure levels from 50mb to 1000mb at 25mb intervals. From among these variables, in Fig. 1[1], the averaged *downward short-wave radiation flux* over North America for the four forecasts on 11th February, 2020 is reported.

3.2.1 DATA COLLECTION

Weather Forecasts

As mentioned in 3.1.1, North American Mesoscale (NAM) Forecast System data was collected from the years 2017 and 2018 for experiments. From this data, surface-level variables as described in

[1]NAM forecast snapshots retrieved from: `https://www.emc.ncep.noaa.gov/mmb/mmbpll/etapll`

Table 1 were retrieved and analyzed. NAM Forecast System projects different weather parameters 84 hours into the future. The first 36 feature projections in the forecast horizon are at a 1-hour temporal resolution, and subsequent 48 hours of the forecast horizon has feature projections at a 3-hour temporal resolution. In this work, we consider the first 24 feature projections (which are at a 1-hour temporal resolution) for each of the weather variables along with their corresponding target pyranometer readings.

Table 1: NWP-NAM weather variables used in model development

Label	**Description**	**Unit**
PRES_SFC	Air Pressure	Pa
HGT_SFC	Geopotential Height	gpm
HGT_TOA	Height at Planetary Boundary Layer	gpm
TMP_SFC	Air Temperature	K
VIS_SFC	Visibility	m
UGRD_TOA	U-Component of Wind Speed	m/s
VGRD_TOA	V-Component of Wind Speed	m/s
DSWRF_SFC	Downward Short-Wave Radiation Flux	W/m^2
DLWRF_SFC	Downward Long-Wave Radiation Flux	W/m^2
TCC_EATM	Total Cloud Cover	%

Temporal Features

In this work, temporal features were designed so as to include the *time of day* and *time of year* representations of the forecasts, which incorporate the periodicity in time[2]. The *time of day* was computed by scaling the number of seconds in the reference time with the inverse of $8.64e+4$ (number of seconds in a day); and the *time of year* was computed by scaling the day of the year with the inverse of 365 or 366, depending on whether it is a leap year or not. The sine and cosine values of these measures were added as the temporal features. Such *time of day* representations make the temporal features pertaining to the target hour in the forecast horizon different from that of the reference time. Thus, temporal features representing the target hours in the forecast horizon were also included along with their corresponding predictors.

Irradiance Observations

The target irradiance observations are obtained from three solar arrays in the solar farm, namely array A, array B and array E, representing a dual-axis tracking array, fixed axis array with 200°(SW)

[2]In [58], Jones attempted to use the *time of day* and *time of year* representations by scaling the epoch representing the reference time (in nanoseconds) with the inverse of $8.64e+13$ (number of nanoseconds in a day) and $3.1536e+16$ (number of nanoseconds in a year) respectively, and including their *sine* and *cosine values*. However, these do not appear to be correct as they do not capture the periodicity of the reference time.

azimuth, and a single-axis tracking array respectively. Each of the solar arrays are installed with thermopile pyranometers from different manufacturers such as Kipp & Zonen[3], and LICOR[4]. The thermopile pyranometers have a black absorptive surface which uniformly absorbs the solar radiation across the short-wave solar spectrum, i.e, between 0.2 μm and 3 μm.

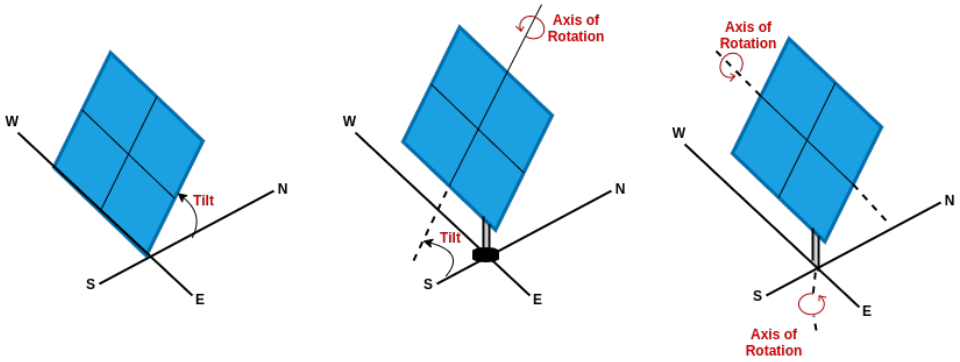


Figure 2: Fixed axis (left), Single-axis tracking (center), Dual-axis tracking (right) Solar Arrays.

The fixed axis solar array has limited exposure to the sun, owing to the change in position of the sun during the day from morning to night. Thus, the solar radiation captured by this solar array is reduced. Though this limitation is minimized by installing the fixed solar array at an optimized tilt angle, the solar radiation captured by solar tracking arrays is still considerably higher. In order to maximize the overall solar energy captured, it is necessary to ensure that the angle of incidence of the sunlight on the solar array is constantly perpendicular. This is achieved with the help of single-axis trackers (horizontal and vertical), which have one degree of freedom acting as an axis of rotation; and dual-axis trackers which have two degrees of freedom acting as axes of rotation normal to one another [8]. This ability to move along the axes enhances the morning and afternoon performance of the solar tracking systems.

While the irradiance observations from the solar arrays are received every five seconds, the NWP NAM model data is available only for four reference times in a day, i.e 00h, 06h, 12h, 18h UTC through 2017 and 2018. Thus, for the target hours in the forecast horizon of all the reference times in 2017 and 2018, for which the NWP NAM model data was collected, the irradiance observations were sampled. In Fig. 3, the average monthly solar radiation captured by the dual-axis tracking, fixed-axis and single-axis tracking solar arrays through 2017 is shown. It can be observed that the solar radiation captured by the tracking arrays (dual-axis and single-axis) is consistently higher than

[3] https://www.kippzonen.com/
[4] https://www.licor.com/

that captured by the fixed axis array.

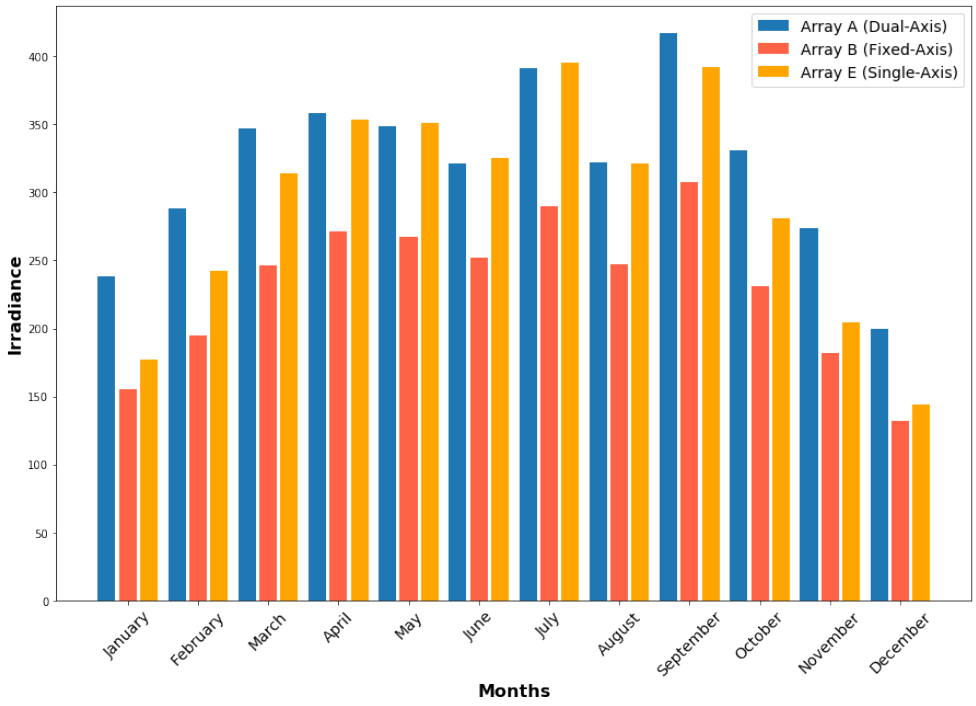


Figure 3: Average monthly solar radiation captured by dual-axis tracking, fixed-axis and single-axis tracking solar arrays through 2017.

3.2.2 EVALUATING IMPACT OF WEATHER VARIABLES ON IRRADIANCE OBSERVATIONS

Mutual information is the measure between two possibly multi-dimensional variables, which quantifies the amount of information obtained from one variable about the other. The relationship detected between the variables can involve either mean, variance or even the higher moments [3]. The most straightforward and widespread approach towards estimating mutual information follows partitioning the supports of X and Y into bins of finite size, and approximating the sum in the following way:

$$I(X,Y) \approx I_{binned}(X,Y) \equiv \sum_{ij} p(i,j).log(\frac{p(i,j)}{p_x(i).p_y(j)}) \tag{1}$$

In this work, the mutual information measure was estimated using the *scikit-learn*[5] machine learning software, which makes use of a non-parametric method based on entropy estimation from the *k-nearest neighbors* as described in [3] and [4]. Mutual information measure was calculated for

[5] `https://scikit-learn.org/stable/`

different weather variables from the NAM data and corresponding irradiance observations from the solar arrays as described in the previous subsections. From among the weather variables, it was observed that downward shortwave radiation flux, air temperature, height at planetary boundary layer and total cloud cover have a mutual information score greater than 0.1, indicating a relatively higher dependency on the irradiance observations.

Downward shortwave radiation flux is the total amount of shortwave radiation that reaches the earth's surface, and is a major component of the total solar radiation on the surface of the earth. Thus, it is the most direct parameter in the estimation solar irradiance, and the high mutual information score between this weather variable and the irradiance observations from the solar farm is understandable. By absorbing the incoming solar radiation, the Earth warms up, and its temperature rises. As long as the amount of incoming radiative flux is greater than the outgoing radiative flux, the Earth will continue to warm. Thus, the air temperature at the surface is essential in estimating the amount of heat absorbed at that particular location, which in turn reveals information about the amount of solar radiation absorbed by the thermopiles in the pyranometers.

The influence of clouds on solar irradiance is significant. In the absence of visible clouds, aerosols, precipitable water and other atmospheric conditions affect the transmission of solar radiation through atmosphere. In cloudy conditions though, the clouds absorb a significant amount of the shortwave radiation, making variables like total cloud cover, which is the fraction of the sky covered by visible clouds essential. The planetary boundary layer (PBL) is the lowest part of the atmosphere which is directly influenced by its contact with the planetary surface. The structure of turbulence within this layer is mainly governed by the PBL height, which is higher during the day, and lower and more stable during nighttime [5]. PBL height characterizes the planetary boundary layer in a fairly integrated manner and affects the weather parameters such as cloud cover and heat flux [6]. This makes PBL height an important parameter in predicting solar irradiance.

For the machine learning models to be able to capture and reconstruct the underlying relationship between input-output data pairs effectively, input selection is essential. By removing the redundant and misleading data, input selection often helps in reducing the computational costs, and improves the accuracy. Several approaches have been defined in literature for the purpose of input selection. In addition to assessing the mutual information scores, we used *random forests* to identify the more important weather variables, as they provide an in-built feature selection.

Random forests employ tree-based strategies, which naturally rank inputs based on how well

they improve the purity of the node. They are an ensemble learning technique constructed over a variety of randomized decision trees, each of which is built over a random extraction of features and data observations. The training of these randomized decision trees is done with the objective of decreasing the *Gini Impurity*, and the features which help in decreasing this measure are selected [7]. Thus, random forests help determine the importance of the features in this manner. It was observed that the weather parameters with higher mutual information scores also received high feature importance scores through this technique, thus validating the dependence of the target irradiance observations on this set of parameters.

In [58], Jones used 37 weather attributes, of which one corresponds to the weather data at the reference time and 36 are feature projections at an one-hour temporal resolution in the forecast horizon, as the predictor variables for machine learning models. In this work, a forecast horizon of 24 hours was selected, and the relationship between the weather attributes in this forecast horizon and the irradiance observations corresponding to the target hours in this forecast horizon was studied.

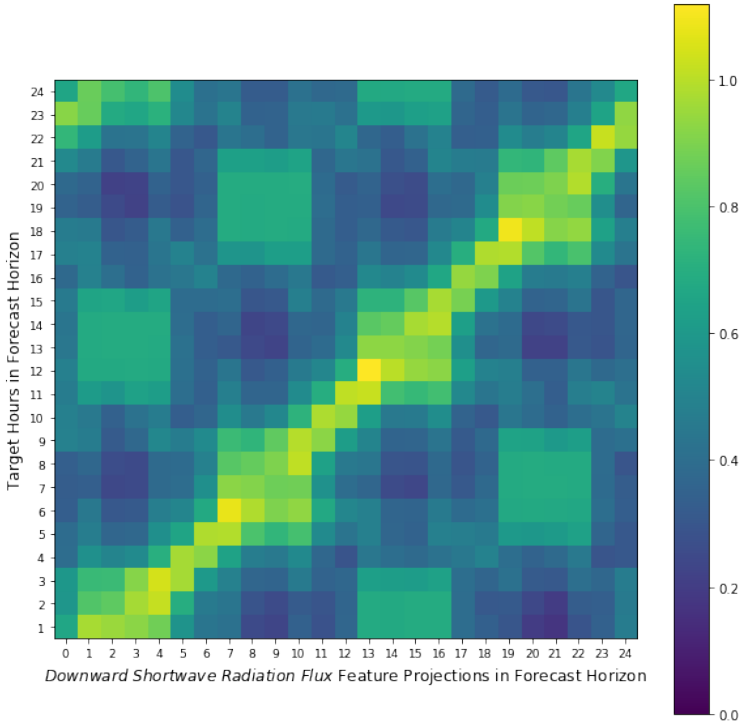


Figure 4: Mutual information between feature projections of Downward Shortwave Radiation Flux in the forecast horizon and irradiance observations for corresponding target hours along fixed-axis solar array.

As was noted earlier, downward shortwave radiation flux was the most important NAM weather

variable with respect to the irradiance observations in the solar farm. In Fig. 4, the mutual information between the feature projections of this weather variable and the corresponding irradiance observations from fixed-axis solar array are represented in a heatmap, wherein, different colours in the colour-bar depict the amount of mutual information measure. It can be observed that the irradiance observations from fixed axis array for a particular target hour are relatively more dependent on only a certain number of feature projections in the forecast horizon. Thus, for the machine learning models trained for each target hour in the forecast horizon, feature projections from six hours ahead, and six hours prior were considered as predictors.

For the first six target hours in the forecast horizon which do not necessarily have six prior feature projections, desired number of feature projections were selected from the end. Similarly, for the last six target hours in the forecast horizon which do not necessarily have six subsequent feature projections, desired number of feature projections were selected from the beginning of the forecast horizon. Such a feature projection selection is justified because it is more likely for the same reference time in two consecutive days to have similar weather conditions. Thus, following this input selections scheme, the NAM model data contributes 13 feature projections for each of the four environmental attributes described earlier, eight temporal features (four for the reference time of the forecast, four for the target hour offset from the reference time) towards the post-processing of solar irradiance from each of the solar arrays using machine learning models.

3.3 EXPERIMENT SETUP

In this chapter, two series of experiments are performed towards predicting solar irradiance on the dual-axis tracking, fixed-axis and single-axis tracking solar arrays. In the first series of experiments, solar irradiance forecasting using Numerical Weather Prediction (NWP) models such as North American Mesoscale (NAM) Forecast System is investigated, replicating the modeling methodology employed by Jones in [58]. This is compared with the processed NAM dataset obtained by incorporating the input selection scheme described in 3.2.2.

In this work, 24-hour *persistence models* were used to set a baseline for the more sophisticated machine learning models. Based on the assumption that conditions remain unchanged between the current time and a future time, 24-hour *persistence models* measure the solar irradiance at a particular time t based on the irradiance measured at $t - 24$. Making use of such a trivial model as baseline provides a reference for improving the machine learning models.

Several machine learning algorithms such as *Least-Squares Linear Regression* (LSLR), *k-Nearest Neighbors* (KNN), *Support Vector Regression* (SVR), *Decision Trees* (DT), *Random Forests* (RF) and *Extreme Gradient Boosted Trees* (XGBT) were used for the purpose of forecasting. Python-based machine learning softwares *scikit-learn* and *xgboost*[6] were used for the implementations of these machine learning algorithms. Randomized cross-validated grid search was employed to identify the optimal set of hyperparameters, ranges for each of which were selected around the default values set for them in the *scikit-learn* implementations.

Weather variables from the NAM Forecast System were used as predictors for the machine learning models, and the irradiance observations from the solar arrays were used as target variables. Each of the weather variables are projected 36 hours into the future at a one-hour temporal resolution. As a part of the input-selection scheme, select feature projections were picked from the important weather variables, depending on the target hour in the forecast horizon. Prediction of target irradiance was done for a day-ahead forecast horizon, i.e, solar irradiance was predicted 24 hours into the future at a one hour temporal resolution. Models were trained on data collected during 2017, and evaluated against data collected during 2018. In the first series of experiments, the performance of the models, with and without employing the input-selection scheme was compared.

Jones [58] reported evaluation metrics such as *mean absolute error* (MAE) and *coefficient of determination* (R^2). MAE is a more natural and unambiguous measure of average error, and is extremely useful in evaluating average-model performance. An evaluation metric such as R^2 helps in providing a reference point for comparing the model results with results from other literature. Owing to these advantages, in this work, both the evaluation metrics were retained to facilitate a consistent comparison. In addition, the correlation coefficient (r) is reported as well.

Model results were analyzed in two schemes: mean of the evaluations for each forecast hour in the forecast horizon (*Overall*); mean of the evaluations for sets of six forecast hours in the forecast horizon, i.e, $1-6$, $7-12$, $13-18$ and $19-24$. MAE was estimated for each of these forecast horizon segments by taking an average of the metric across each of the target hours in the segment. However, for R^2 and r, this was performed by flattening the predictions and ground-truth values for multiple target hours in the forecast horizon segment into single lists, and computing the metrics over these lists. Such an analysis helped in realizing the performance of the models specifically for different periods in the day.

[6] `https://github.com/dmlc/xgboost`

Geographic expansion of forecast coverage by including additional weather forecasts specific to areas surrounding the target location is considered to improve the solar irradiance forecasting capabilities. In the second series of experiments, the effect of such a spatial expansion is investigated, by including the feature projections of weather variables from a grid of cells around the NAM data grid representing Athens, as predictors to the machine learning models. A geographic expansion from the 1 x 1 grid to other *geo shapes* such as 3 x 3 and 5 x 5 is investigated for the *K-Nearest Neighbors*, *Random Forests* and *Extreme Gradient Boosted Trees* algorithms. Each of these methodologies are further explained in finer detail.

3.3.1 IRRADIANCE FORECASTING WITH NAM FORECAST SYSTEM

For the replication study, North American Mesoscale (NAM) weather forecast data and target irradiance data from the solar farm at the University of Georgia were collected for the years 2017 and 2018. For a forecast horizon of 24 hours, planar surface features from the NAM Forecast System such as air pressure, geopotential height, height at planetary boundary layer, air temperature, u-component of wind speed, v-component of wind speed, downward short-wave radiation flux and downward long-wave radiation flux were used.

As mentioned in 3.2.2, it was determined that weather variables such as air temperature, total cloud cover, atmospheric height and downward short-wave radiation flux from among the surface-level planar features affected the solar irradiance predictions more. Hence, the other weather variables were omitted. Depending on the target hour offset in the forecast horizon, select feature projections were picked for the weather variables, so as to be included in the NAM dataset. This was done by selecting 13 from among the 37 weather attributes at a one-hour temporal resolution such that, six followed the reference time of the target hour offset, six preceded the reference time of the target hour offset, and one corresponded to the reference time of the target hour offset. In order to choose an ideal set of parameters for the machine learning models, *hyperparameter tuning* was performed with the help of a randomized cross-validated grid search.

The weather forecast data obtained by both methodologies was input to several machine learning models, which were trained on data collected during 2017, and evaluated against data collected during 2018. The results obtained by both methodologies, i.e with and without incorporating the input selection scheme, were compared and analyzed.

3.3.2 GEOGRAPHIC EXPANSION OF FORECAST COVERAGE

Lorenz et al. [1] found that expanding the forecast region to approximately 100km x 100km resulted in an improvement in day-ahead solar forecasting. They performed a spatial averaging across the region, by taking an arithmetic mean of the weather variables from the surrounding weather data grid cells. In contrast, Sanders et. al. [49] and Jones [58] performed a distance-dependent weighted averaging, by including the weather variables from the surrounding weather data grid cells as predictors to the machine learning models.

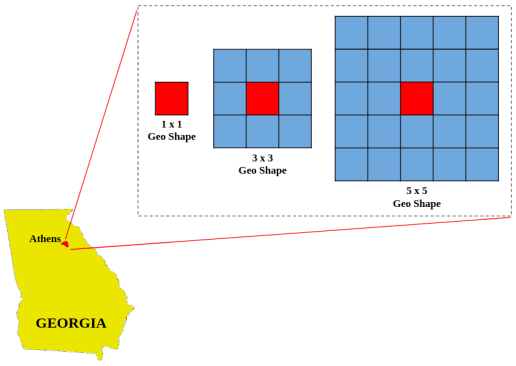


Figure 5: Geographic expansion of forecast coverage with 1 x 1 Geo Shape representing Athens NAM model data grid, 3 x 3 Geo Shape and 5 x 5 Geo Shape representing grid of cells around Athens.

Jones [58] trained *k-nearest neighbors* and *random forest* algorithms for 3 x 3, 5 x 5 and 7 x 7 *geo shapes*, each reflecting a spatial expansion up to 36km x 36km, 60km x 60km and 84km x 84km respectively. As shown in Figure 5, each of the *geo shapes* represent eight, fifteen and forty eight NAM weather forecast data grid cells centered around Athens, Georgia. They realized that including the weather forecasts from the surrounding data grid cells resulted in an improved day-ahead solar forecasting, though it was observed that the improvement diminished as the *geo shape* grew larger. Additionally, Jones [58] also noted that the 3 x 3 *geo shape*, equivalent to a 36km x 36km area was optimal.

In this work both the schemes were compared: one in which GHI from the surrounding grids is averaged, and other in which weather variables from surrounding cells are included as predictors. It was observed that the latter helped in improving the performance of the models more. Thus, a geographic expansion of weather forecast coverage was carried out with 3 x 3 and 5 x 5 *geo shapes*, resulting in a spatial expansion upto 60km x 60km. The dataset set up using the input selection scheme described in 3.2.2, was used to determine the effect of geographic expansion.

3.4 RESULTS AND DISCUSSION

Assessment of Model Performance with and without Input Selection Scheme

In this work, an input selection scheme as described in 3.2.2 was incorporated towards selecting features for the machine learning models. The performance of the machine learning models using both the methodologies, i.e. with and without the input selection scheme were compared for dual-axis tracking solar array, fixed-axis solar array and single-axis tracking solar array. As a part of this scheme, the key differences between Jones' dataset and the one used in this work, towards training with the machine learning models are as follows:

- Jones [58] hadn't considered the *total cloud cover* weather variable in the NAM weather dataset
- from among the other surface-level NAM weather variables used, only air temperature, height at planetary boundary layer and downward shortwave radiation flux were considered
- instead of the 37 weather attributes for each of the weather variables, select feature projections depending on the target hour offset were chosen as predictor variables
- 1 x 1 *geo shape* was selected instead of 3 x 3 (which Jones [58] had found to be optimal)
- *time of day* and *time of year* encodings were modified to incorporate periodicity of the reference time in a particular day or in a particular year

For the dual-axis tracking solar array, all the machine learning models performed exceedingly well with respect to the baseline 24-hour *persistence* models, which was expected. A substantial improvement was observed for all the machine learning models built using weather forecast data without incorporating the input-selection scheme. In particular, using the input selection scheme helped in improving the MAE of simple *linear regression* by 52.17%. There was a considerable improvement in the performance of *support vector regression* and *k-nearest neighbors* algorithms as well, with the MAE reducing by 17.76% and 28.97% respectively. Each of these algorithms is greatly affected by a higher dimensionality and quality of data, and by weeding out weather variables and their feature projections which have a lesser influence on the target irradiance, the improvement in performance of these models can be justified.

Ensemble tree-based methods have the intrinsic ability to calculate feature importance, and account for the possible correlations between the variables. Thus in general, they perform better

Table 2: Comparing performance of machine learning algorithms trained against dual-axis tracking array utilizing NAM Forecast System data: (a) without input selection (upper), (b) with input selection (lower)

	Metric	**Horizon**	**PER**	**LSLR**	**SVR**	**KNN**	**DT**	**RF**	**XGBT**
Without Input Selection	MAE	$1-6$	153.32	231.00	88.38	98.73	98.34	74.38	73.03
		$7-12$	153.91	231.06	89.83	101.77	105.17	77.14	76.97
		$13-18$	154.16	232.15	89.08	106.56	98.31	74.43	75.64
		$19-24$	161.43	243.04	89.88	104.15	98.19	75.02	76.69
		Overall	155.71	234.31	89.29	102.80	100.00	**75.24**	75.58
	R^2	$1-6$	0.46	0.31	0.85	0.78	0.71	0.86	0.86
		$7-12$	0.45	0.29	0.84	0.78	0.68	0.85	0.84
		$13-18$	0.45	0.31	0.84	0.77	0.71	0.86	0.85
		$19-24$	0.41	0.25	0.84	0.77	0.71	0.85	0.85
		Overall	0.44	0.29	0.84	0.77	0.70	0.85	0.85
Relative Imp. in MAE (%)		*Overall*	—	52.17	17.76	28.97	10.91	3.47	1.02
Input Selection	MAE	$1-6$	153.32	107.08	70.63	68.54	85.74	70.21	71.65
		$6-12$	153.91	114.52	73.82	72.35	90.57	72.92	74.6
		$13-18$	154.16	115.03	73.34	73.12	87.65	72.6	75.77
		$19-24$	161.43	111.68	75.93	78.04	92.39	74.78	77.23
		Overall	155.71	112.08	73.43	73.02	89.09	**72.63**	74.81
	R^2	$1-6$	0.46	0.84	0.88	0.88	0.78	0.88	0.87
		$7-12$	0.45	0.83	0.86	0.86	0.76	0.87	0.86
		$13-18$	0.45	0.82	0.86	0.86	0.77	0.87	0.86
		$19-24$	0.41	0.82	0.85	0.85	0.74	0.86	0.85
		Overall	0.44	0.83	0.86	0.86	0.76	0.87	0.86

than the linear regression methods. For the dual-axis tracking array, random forests had the best performance with an MAE of 72.63 W/m^2, and extreme gradient boosted trees recorded an overall MAE of 74.81 W/m^2. The improvement over the performance of the same algorithms without incorporating the input selection scheme was 3.47% and 1.02% respectively.

While the random forests performed the best, considerable improvements in performance as a result of incorporating the input selection scheme were seen in the *support vector regression* and *k-nearest neighbors* algorithms. Overall, there was an average improvement in MAE by 19.05% across the machine learning models, with *random forests* having the best MAE with 72.63 W/m^2 for the dual-axis tracking solar array.

In general, it's expected that the performance of the models degrades as the forecast horizon increases. However, Jones [58] did not observe such a pattern, with sometimes, even the $7-12$ hour forecast horizon performing worse than $13-18$ and $19-24$ hours forecast horizons. Such a trend

Table 3: Comparing performance of machine learning algorithms trained against fixed-axis solar array utilizing NAM Forecast System data: (a) without input selection (upper), (b) with input selection (lower)

	Metric	**Horizon**	**PER**	**LSLR**	**SVR**	**KNN**	**DT**	**RF**	**XGBT**
Without Input Selection	MAE	$1-6$	111.23	151.79	55.018	61.10	62.81	47.09	47.21
		$7-12$	111.51	147.12	56.65	63.78	62.05	48.32	49.56
		$13-18$	111.97	151.09	55.64	68.70	64.64	48.19	49.22
		$19-24$	117.15	159.29	56.11	66.53	62.28	47.65	49.44
		Overall	112.96	152.32	55.85	65.03	62.95	**47.81**	48.86
	R^2	$1-6$	0.59	0.54	0.89	0.85	0.80	0.90	0.89
		$7-12$	0.58	0.55	0.89	0.85	0.80	0.89	0.88
		$13-18$	0.58	0.55	0.89	0.83	0.78	0.89	0.88
		$19-24$	0.55	0.50	0.89	0.83	0.80	0.89	0.88
		Overall	0.58	0.54	0.89	0.84	0.80	0.89	0.88
Relative Imp. in MAE (%)		*Overall*	—	51.92	16.47	28.42	10.90	6.00	4.36
Input Selection	MAE	$1-6$	111.23	69.81	44.26	42.77	55.30	42.94	44.66
		$6-12$	111.51	74.74	47.19	46.72	55.96	45.16	46.77
		$13-18$	111.97	79.63	47.02	47.02	56.65	45.14	46.79
		$19-24$	117.15	68.75	48.13	49.67	56.48	46.51	48.69
		Overall	112.96	73.24	46.65	46.55	56.09	**44.94**	46.73
	R^2	$1-6$	0.59	0.89	0.91	0.92	0.84	0.92	0.91
		$7-12$	0.58	0.88	0.90	0.90	0.84	0.91	0.90
		$13-18$	0.58	0.87	0.90	0.90	0.83	0.91	0.90
		$19-24$	0.55	0.88	0.89	0.89	0.83	0.90	0.89
		Overall	0.58	0.88	0.90	0.90	0.84	0.91	0.90

though, was realized in the results obtained by incorporating the input selection scheme. By and large, most of the models displayed a trend where the error increased (performance degraded) with the target hour in the forecast horizon. Considering the overall improvement in performance, it can be inferred that this indicates an improvement in the short-term forecasting ability of the models.

Similar trends were also observed in the performance of the machine learning models for irradiance predictions on the fixed-axis solar array (in Table 3). The *random forests* performed the best with an MAE of 44.94 W/m^2. This model achieved an improvement of 6% due to the incorporation of the input-selection scheme. In contrast, *random forests*, which were also the best-performing model without incorporating the input-selection methodology (as followed by Jones [58]), achieved a MAE of 47.81 W/m^2. In all, the input-selection scheme achieved an average improvement of 19.68% in MAE across all the machine learning models.

Table 4: Comparing performance of machine learning algorithms trained against single-axis tracking solar array utilizing NAM Forecast System data: (a) without input selection (upper), (b) with input selection (lower)

	Metric	Horizon	PER	LSLR	SVR	KNN	DT	RF	XGBT
Without Input Selection	MAE	$1-6$	128.64	174.61	67.52	83.58	77.53	56.81	57.67
		$7-12$	128.97	176.21	70.13	86.63	82.25	60.32	61.40
		$13-18$	129.25	176.39	68.79	91.08	81.01	58.54	60.15
		$19-24$	135.35	182.16	68.70	87.62	82.42	58.42	60.85
		Overall	130.55	177.34	68.79	87.23	80.80	**58.52**	60.02
	R^2	$1-6$	0.53	0.48	0.88	0.79	0.77	0.89	0.88
		$7-12$	0.53	0.46	0.87	0.79	0.75	0.75	0.87
		$13-18$	0.53	0.48	0.87	0.77	0.75	0.88	0.87
		$19-24$	0.49	0.45	0.87	0.78	0.74	0.88	0.87
		Overall	0.52	0.47	0.87	0.78	0.75	0.88	0.87
Relative Imp. in MAE (%)		*Overall*	—	48.39	5.9	24.65	2.08	-8.68	-8.46
Input Selection	MAE	$1-6$	128.64	87.52	62.29	61.71	75.62	61.37	62.93
		$6-12$	128.97	95.30	65.77	65.98	81.80	64.09	65.55
		$13-18$	129.25	93.47	64.29	65.75	76.77	63.50	65.73
		$19-24$	135.35	89.84	66.57	69.47	82.28	65.44	66.18
		Overall	130.55	91.53	64.73	65.73	79.12	**63.60**	65.10
	R^2	$1-6$	0.53	0.86	0.88	0.88	0.80	0.89	0.88
		$7-12$	0.53	0.85	0.87	0.86	0.77	0.87	0.87
		$13-18$	0.53	0.85	0.87	0.86	0.79	0.88	0.87
		$19-24$	0.49	0.85	0.86	0.86	0.77	0.87	0.87
		Overall	0.52	0.85	0.87	0.86	0.78	0.88	0.87

In the present investigation of incorporating the input-selection scheme, the most interesting results were seen with the single-axis tracking solar array predictions (in Table 4). For the *k-nearest neighbors* algorithm, trends followed the predictions for the other solar arrays reported so far, with a reduction in MAE by 24.65%. For the predictions with *support vector regression*, a reduction in MAE by 5.9% was observed. However, this relative improvement in performance was less in magnitude as compared to those observed for dual-axis tracking array predictions and fixed-axis array predictions. In addition, using the tree-based ensemble methods such as *random forests* and *extreme gradient boosted trees*, a degradation in performance was recorded, with the MAE increasing by 8.68% and 8.46%. Though *random forests* still performed the best with an MAE of 63.6 W/m^2, this performance paled in comparison to that of the *random forests* without incorporating the input-selection scheme, where an MAE of 58.52 W/m^2 was obtained. To explain this, the corresponding data and code were investigated. No errors in data processing were found. Moreover, a systematic error would have affected all the arrays, which wasn't the case.

Evaluating Effect of Geographic Expansion of Forecast Coverage

Better performing algorithms in 3.3.1 such as *k-nearest neighbors*, *random forests* and *extreme gradient boosted trees* were retrained on dual-axis tracking, fixed-axis and single-axis tracking solar arrays and corresponding weather forecast data for the year 2017, and evaluated against data belonging to the year 2018. The performance of these models for each of the *geo shapes* 1 x 1, 3 x 3 and 5 x 5 was compared and analyzed.

Using the *k-nearest neighbors* algorithm to predict the day-ahead solar irradiance on dual-axis tracking solar array, it was observed that the geographic expansion had a slightly detrimental effect on the performance. Expanding to 3 x 3 *geo shape* resulted in increasing the MAE by 1.59%, and increasing the weather forecast coverage to 5 x 5 *geo shape* resulted in increasing the MAE by 0.44%. For these models, the 1 x 1 performed best with a MAE of 73.01 W/m^2.

Table 5: Evaluating effect of geographic expansion of forecast coverage for dual-axis tracking array.

Metric	**Horizon**	**KNN**			**RF**			**XGBT**		
		1x1	3x3	5x5	1x1	3x3	5x5	1x1	3x3	5x5
MAE	$1-6$	68.61	69.68	69.02	68.53	67.90	66.69	71.04	69.23	85.32
	$7-12$	72.52	73.72	72.77	72.40	71.40	69.38	73.51	72.44	87.16
	$13-18$	72.86	74.27	73.06	70.82	69.85	68.39	73.69	71.91	87.47
	$19-24$	78.05	78.99	78.47	74.99	74.46	73.06	77.92	76.02	90.02
	Overall	**73.01**	74.17	73.33	71.68	70.90	**69.38**	74.04	**72.40**	87.49
R^2	$1-6$	0.88	0.87	0.87	0.88	0.88	0.89	0.87	0.88	0.84
	$7-12$	0.86	0.85	0.86	0.87	0.87	0.88	0.86	0.86	0.84
	$13-18$	0.86	0.85	0.85	0.87	0.87	0.88	0.86	0.87	0.83
	$19-24$	0.85	0.84	0.84	0.86	0.86	0.87	0.84	0.85	0.83
	Overall	0.86	0.85	0.86	0.87	0.87	0.88	0.86	0.87	0.83
r	$1-6$	0.94	0.94	0.94	0.94	0.94	0.94	0.93	0.94	0.94
	$7-12$	0.93	0.93	0.93	0.93	0.93	0.94	0.93	0.93	0.93
	$13-18$	0.93	0.92	0.92	0.93	0.94	0.94	0.93	0.93	0.93
	$19-24$	0.92	0.92	0.92	0.93	0.93	0.93	0.92	0.92	0.92
	Overall	0.93	0.93	0.93	0.93	0.93	0.94	0.93	0.93	0.93
Relative Imp. in MAE (%)	*Overall*	—	-1.59	-0.44	—	1.09	3.21	—	2.22	-18.17

However, for the *random forests* trained on weather forecast data and irradiance observations from dual-axis tracking array, it was observed that spatial expansion was beneficial. While expanding from 1 x 1 grid to 3 x 3 and 5 x 5 *geo shapes*, the model performance in MAE improved by 1.09% and 3.21% respectively. An MAE of 70.90 W/m^2 and 69.38 W/m^2 was recorded for each of the

geo shapes. A geographic expansion by including the additional weather forecasts as predictors did not have a negative effect on model performance, possibly due to the better attribute-selection capabilities of the decision-tree based ensemble algorithm.

An improvement in performance of this nature was expected for another decision tree based ensemble algorithm, *extreme gradient boosted trees* as well. However in this case, while expanding to 3 x 3 *geo shape* improved the performance on the dual-axis tracking solar arrays by 2.22%, resulting in an MAE of 72.40 W/m^2, expanding to 5 x 5 *geo shape* had an extremely detrimental performance on the models, subsequently increasing the MAE by 18.17% to 87.49 W/m^2. The best performance for the *extreme gradient boosted trees* models trained on the irradiance observations from the dual-axis tacking solar arrays was recorded for the 3 x 3 *geo shape*, with an MAE of 72.40 W/m^2.

Extreme gradient boosted trees are more sensitive to overfitting if the data is noisy. Because they are built sequentially, training time is generally higher as well. Owing to this, when compared to *random forests*, these models are harder to tune. In this work, for the *extreme gradient boosted trees*, the number of trees, depth of trees and the learning rate were tuned. There was a sharp increase in the number of predictors from 1 x 1 to 5 x 5. It is possible that an insufficient number of trees with lesser depth (than that required for the scale of the dataset) were used in this ensemble technique, resulting in shallow trees being trained for this model.

Similar trends in model performance was observed for irradiance predictions on the fixed-axis and single-axis tracking solar arrays as well. Using the *k-nearest neighbors* algorithms, a best MAE of 46.41 W/m^2 and 65.69 W/m^2 was recorded for the fixed-axis and single-axis tracking solar arrays respectively, utilizing weather forecast data corresponding to 1 x 1 *geo shape*. Expanding to 5 x 5 *geo shape*, and leveraging the weather forecast data to the *random forests* algorithm improved the performance by 2.79% and 2.3% resulting in an MAE of 43.62 W/m^2 and 61.99 W/m^2 for each of the solar arrays.

Using the *extreme gradient boosted trees* algorithm too, similar trends were recorded, with the 3 x 3 *geo shape* performing the best, resulting in an MAE of 45.78 W/m^2 and 63.77 W/m^2 for the fixed-axis and single-axis tracking solar arrays. However, it is to be noted that the *random forests* performed best for the single-axis tracking array, with an MAE of 61.99 W/m^2 for the 5 x 5 *geo shape*. This is still worse than the best performance observed by Jones [58], with an MAE of 58.52 W/m^2 for a 3 x 3 *geo shape* without input selection.

Table 6: Evaluating effect of geographic expansion of forecast coverage for fixed-axis array.

Metric	Horizon	KNN			RF			XGBT		
		1x1	3x3	5x5	1x1	3x3	5x5	1x1	3x3	5x5
MAE	$1-6$	42.62	44.10	44.06	42.33	41.77	41.12	44.46	43.42	56.95
	$7-12$	46.64	48.43	48.03	45.53	44.94	43.82	46.71	45.84	58.85
	$13-18$	46.82	48.77	48.26	44.80	44.13	43.73	46.56	45.56	59.97
	$19-24$	49.59	51.10	50.85	46.82	46.20	45.80	48.63	48.32	60.49
	Overall	**46.41**	48.10	47.80	44.87	44.26	**43.62**	46.59	**45.78**	59.07
R^2	$1-6$	0.92	0.91	0.92	0.92	0.92	0.93	0.91	0.92	0.88
	$7-12$	0.90	0.89	0.89	0.90	0.91	0.91	0.90	0.90	0.87
	$13-18$	0.90	0.89	0.89	0.91	0.91	0.91	0.90	0.90	0.87
	$19-24$	0.89	0.89	0.89	0.90	0.90	0.91	0.89	0.89	0.86
	Overall	0.90	0.90	0.90	0.91	0.91	0.91	0.90	0.90	0.87
r	$1-6$	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
	$7-12$	0.95	0.95	0.95	0.95	0.95	0.96	0.95	0.95	0.95
	$13-18$	0.95	0.94	0.95	0.95	0.95	0.96	0.95	0.95	0.95
	$19-24$	0.95	0.94	0.94	0.95	0.95	0.95	0.95	0.95	0.95
	Overall	0.95	0.95	0.95	0.95	0.96	0.96	0.95	0.95	0.95
Relative Imp. in MAE (%)	*Overall*	—	-3.64	-3.00	—	1.36	2.79	—	1.74	-26.79

Another aspect which needs to be considered in the comparison of methodologies with/without incorporating the input selection scheme is the **computational cost**. For the 1 x 1 *geo shape*, the methodology followed by Jones [58] uses input data for the machine learning models involving 337 predictors (333 weather attributes, 4 temporal attributes). In contrast, the input data generated by incorporating the input-selection scheme for 1 x 1 *geo shape* uses 60 predictors (52 weather attributes, 8 temporal attributes). Similarly, for the 3 x 3 and 5 x 5 *geo shapes*, as compared to 3001 (2997 weather attributes, 4 temporal attributes) and 8329 (8325 weather attributes, 4 temporal attributes) predictors respectively, by incorporating the input-selection scheme, 476 (468 weather attributes, 8 temporal attributes) and 1308 (1300 weather attributes, 8 temporal attributes) predictors respectively were used. This drastic reduction in the number of predictors led to a considerable decrease in the training time of the machine learning models.

Stratified Diurnal Analysis of Performance

NAM forecasts are released at 00h, 06h, 12h and 18h UTC. In order to assess the performance of the machine learning models as a result of incorporating the input-selection scheme better, a stratified analysis was carried out. The mean absolute error (MAE) of the models for each of the

Table 7: Evaluating effect of geographic expansion of forecast coverage for single-axis tracking array

Metric	Horizon	KNN			RF			XGBT		
		1x1	3x3	5x5	1x1	3x3	5x5	1x1	3x3	5x5
MAE	$1-6$	61.68	63.41	63.00	60.83	60.55	60.15	61.32	60.96	75.44
	$7-12$	65.93	67.27	67.03	64.18	63.73	62.52	65.58	64.43	77.09
	$13-18$	65.64	67.99	67.08	62.61	60.91	60.32	64.35	62.66	76.09
	$19-24$	69.53	71.07	70.83	66.19	65.82	64.97	67.30	67.01	80.16
	Overall	**65.69**	67.44	66.99	63.45	62.75	**61.99**	64.63	**63.77**	77.20
R^2	$1-6$	0.88	0.88	0.88	0.89	0.89	0.89	0.88	0.89	0.85
	$7-12$	0.86	0.86	0.86	0.87	0.87	0.88	0.87	0.87	0.84
	$13-18$	0.86	0.85	0.86	0.88	0.88	0.88	0.87	0.87	0.84
	$19-24$	0.86	0.85	0.85	0.87	0.87	0.87	0.86	0.86	0.83
	Overall	0.86	0.86	0.86	0.88	0.88	0.88	0.87	0.87	0.84
r	$1-6$	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
	$7-12$	0.93	0.93	0.93	0.93	0.94	0.94	0.93	0.93	0.93
	$13-18$	0.93	0.92	0.93	0.94	0.94	0.94	0.93	0.93	0.94
	$19-24$	0.93	0.92	0.92	0.93	0.93	0.93	0.93	0.93	0.93
	Overall	0.93	0.93	0.93	0.94	0.94	0.94	0.93	0.93	0.93
Relative Imp. in MAE (%)	*Overall*	—	-2.66	-1.98	—	1.10	2.30	—	1.33	-19.45

target hours in the forecast horizon, i.e. between 1 and 24 was compared for all four types of NAM forecasts individually.

For using weather data from NAM Forecast System for training with the machine learning models, the reference times of the NAM forecasts were made time-zone aware with respect to the target location, i.e. Athens, Georgia. For the reference times corresponding to each of the NAM forecasts, corresponding solar irradiance observations were collected. The target location is -5.00 hours with respect to UTC in the standard time zone, and -4.00 hours with respect to UTC during *daylight saving time*. For the sake of a diurnal analysis, it was assumed that the target location is -4.00 hours with respect to UTC throughout the year.

Thus, 00h, 06h, 12h, 18h NAM forecasts each correspond to 8 P.M, 2 A.M, 8 A.M and 2 P.M locally. In Figure 6, for each of the NAM forecasts, local *time of day* (i.e. between 6.00 AM and 6.00 PM) was identified. In the forecast horizon, i.e. in the consequent 24 hours, for each of the NAM forecasts, such *time of day* was marked in yellow, so as to signify daytime. In Figure 6, it can be seen that the performance of most of the machine learning models is comparable regardless of day-time or night-time. While the *support vector regression* models performed well during daytime,

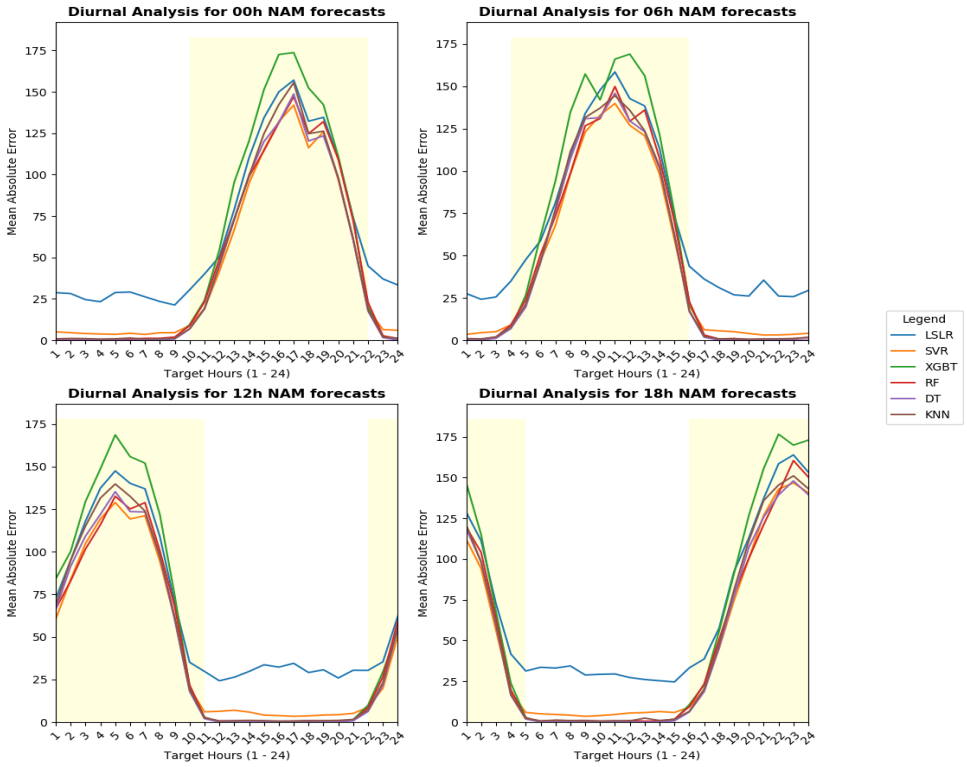
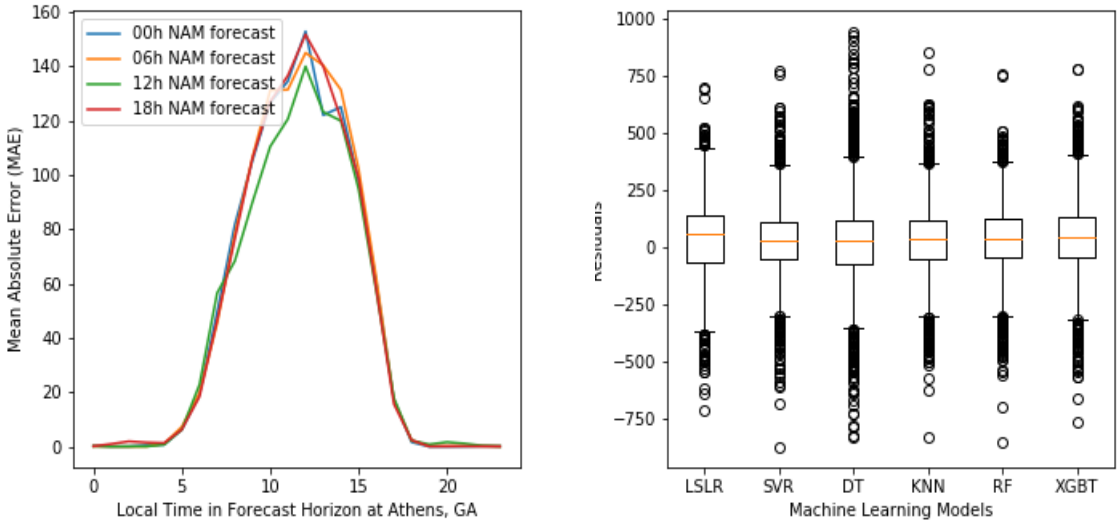


Figure 6: Stratified diurnal analysis of day-ahead irradiance predictions for fixed-axis array: (left-top) 00h NAM forecasts, (right-top) 16h NAM forecasts, (left-bottom) 12h NAM forecasts, (right-bottom) 18h NAM forecasts. Local time of day (6A.M to 6P.M) at the target location for each of the NAM forecasts is indicated in light yellow.

they performed slightly worse during night-time, only performing better than the simple *linear regression*. The stratified diurnal analysis was essential in realizing that the performance of the models in Tables 2, 3 and 4 is misleading, as a much higher MAE can be observed for certain target hours in the *time of day*, for each of the forecasts.

In order to get a better understanding of the performance of the models, a local-time analysis was done. For each of the NAM forecasts, the forecast horizon corresponds to the following time ranges with respect to the target location: 9 P.M - 8 P.M, 3 A.M - 2 A.M, 9 A.M - 8 P.M, 3 P.M - 2 P.M. From these local time ranges, the predictions obtained from the *random forests* algorithm corresponding to target hours representing the same local time were sampled together. MAE was estimated for each of these groups, so as to attain the model performance at a certain time in a day, at the target location. In Fig. 7(a), this was plotted for each of the 00h, 06h, 12h, 18h NAM



(a) Local time analysis of NAM forecasts

(b) Model comparison at noon (local time)

Figure 7: (a) Average performance of different target hours in the forecast horizon corresponding to each of 00h, 06h, 12h, 18h NAM forecasts, adjusted according to local time. (b) Comparison of box-and-whisker plots of residuals from different predictive models utilizing GHI at 12 P.M local time, i.e. noon.

forecasts. The worst performance, expectedly, was observed at 12.00 P.M (local time) or noon. Such an analysis helps realize that for day-ahead solar forecasting with quality weather forecast data, the local time for which the prediction is being made is key to the quality of the prediction, than how farther away the target hour is in the forecast horizon.

In order to ascertain the better machine learning model at the local time most difficult to predict solar irradiance for, the quality of prediction of all machine learning models at 12 P.M (local time) was compared. To enable this, in Fig. 7(b), box-and-whisker plots were drawn for the residuals of the predictions from all the models at this time. It was observed that the *random forests* had the least dispersion beyond the whiskers, indicating its effectiveness in irradiance prediction irrespective of the position of the target hour in the forecast horizon, or the local time of the target hour.

CHAPTER 4

MULTI-MODEL BLENDING APPROACHES TO SOLAR IRRADIANCE FORECASTING

4.1 OVERVIEW

There is a significant variability in the *global horizontal irradiance* (GHI) measured by NWP models with respect to cloud conditions. Mathiesan and Kliessl [46] found that the NAM Forecast System tends to overpredict GHI in clear-sky conditions, i.e. sky conditions in which visible clouds are absent, by up to 40 percent. They proposed a bias-correction scheme to selectively correct the overpredicted GHI. In this scheme, they derived a multivariate fourth-order model-output statistics (MOS) correction function depending on solar zenith angle (θ_z) and clear-sky index (K_c). Based on K_c, sky-conditions for the forecasts were determined. The bias correction function was employed on the forecasts possessing a positive bias in clear-sky conditions.

However, Diagne et al. [17] noted that this bias-correction methodology was not adequate, as even the accurate forecasts were unnecessarily corrected. Thus, a need for a superior bias-correction methodology depending on cloud-conditions was identified, so as to improve the GHI predicted by the NAM Forecast System. The experiments conducted in this chapter are devoted towards exploring theory-driven approaches for this purpose.

There are multiple empirical solar radiation formulations which have been extensively discussed in literature [19][20][21][22][23][31], which compute the different components of solar irradiance from environmental conditions, through experimental observations. These can be broadly classified into *decomposition* and *parametric* models [25]. Using assumptions on solar geometry and transmittance, the former are used to estimate direct beam and diffuse irradiance. The latter are useful for approximating daily solar radiation reaching tilted surfaces. In this chapter, two such empirical solar radiation models, *Clear-Sky Scaling* and *Liu-Jordan Model* are discussed, which help estimate different components of global solar radiation, including GHI.

In climatic research, in order to be able to distinguish between clear-sky and cloudy-sky conditions effectively, measures such as *clear-sky index* (K_c) and *clearness index* (K_t) have been introduced. Clear-sky index is generally described as the ratio of measured global horizontal irradiance in a system to its measure in clear-sky conditions, estimated with the help of a clear-sky solar radiation model. It makes accurate and continuous determination of cloud amount from surface measurements possible [41]. In contrast, clearness index is simply a ratio of the measured irradiance at a location, to the extraterrestrial irradiance calculated at the location. It is extremely useful as it incorporates both light scattering and light absorption, which is beneficial towards estimating the shortwave radiation reaching the surface of the earth.

In this work, a theory-driven multi-model blending approach is explored towards correcting the reported bias in GHI. This is done by combining the NAM Forecast System with the empirical *Clear-Sky Scaling* technique based on K_c estimates, and with the *Liu-Jordan* model based on K_t estimates. In the case of the former, the clear-sky Ineichen Model was used to compute the clear-sky GHI, which goes into estimating K_c. From among the 00h, 06h, 12h, 18h UTC forecasts released by the NAM Forecast System, 18h NAM forecasts with $K_c > 0.85$ were corrected, by substituting the GHI from NAM Forecast System (GHI_{NAM}) with the arithmetic mean of the GHI_{NAM} and GHI estimates from *Clear-Sky Scaling* technique (GHI_{CS}).

Predictive models implementing the random forest technique were developed utilizing GHI_{NAM} and GHI_{CS} independently. The models were compared, and it was observed that such a bias-correction scheme led to an improvement in performance, reducing the mean absolute error (MAE) by 4.95%, 4.53% and 4.12% for the dual-axis tracking, fixed-axis and single-axis tracking solar arrays respectively. A similar comparison was performed by including other weather variables from NAM Forecast System, and incorporating the input-selection scheme described in 3.2. However, in this case, the model-blending approach did not lead to an improvement in performance, resulting in an MAE of 72.57 W/m^2, 44.91 W/m^2 and 63.6 W/m^2 for each of the solar arrays.

A similar model-blending approach was evaluated for combining the NAM Forecast System with *Liu-Jordan* model, depending on K_t estimates. K_t was computed based on estimations for *extraterrestrial radiation* and *solar zenith angle* for the target location, i.e. Athens, Georgia. For the 18h NAM forecasts with $K_t > 0.35$, GHI_{NAM} was substituted with the arithmetic mean of GHI_{NAM} and GHI from *Liu-Jordan* model (GHI_{LJ}). Predictive models implementing the random forests technique were developed, utilizing GHI_{NAM} and GHI_{LJ} independently as well. It was observed

that the former performed better, with the MAE improving by 4.17%, 4.14% and 3.62% for dual-axis tracking, fixed-axis and single-axis tracking solar arrays respectively. Additionally, predictive models implementing random forests technique were developed by including other weather variables from NAM Forecast System, and incorporating the input-selection scheme. For these models, the model-blending technique *slightly deteriorated* the performance, resulting in an MAE of 72.74 W/m^2, 45.25 W/m^2 and 63.97 W/m^2 for each of the solar arrays.

In order to study the utility of clear-sky index, another model-blending methodology was explored in which clear-sky index was included as a predictor variable to various machine learning models rather than GHI. The intuition behind this approach was to capture the diurnal and seasonal cyclicity capturing ability of this measure [42]. Such an ability can be attributed to the *solar zenith angle*, which goes into the computation of this measure and helps track the position of the Sun. An input-selection scheme in line with that followed in Chapter 3 was used to generate a weather forecast data for the machine learning models, so as to forecast solar irradiance for the dual-axis tracking, fixed-axis and single-axis tracking solar arrays.

The performance of such predictive models was worse than that obtained by utilizing the input-selected weather forecast data from the NAM Forecast System (reported in Tables 2, 3, 4). The MAE increased for the best performing *random forests* by 11.59%, 9.8% and 9.42% along each of the solar arrays, recording an MAE of 79.58 W/m^2, 49.18 W/m^2 and 69.98 W/m^2 respectively. A stratified diurnal and seasonal analysis was performed on the predictions attained through this methodology. The presumed cyclicity-capturing ability of the clear-sky index measure did not translate into improving the performance of the predictive models across corresponding seasons.

4.2 EMPIRICAL SOLAR RADIATION MODELS

Solar researchers have developed various formulations through experimental observations, which help in determining the relation between different components of solar radiation and various meteorological parameters. Of these, *parametric* models require information about atmospheric conditions such as turbidity, cloud cover, etc. to be able to formulate the different components of solar irradiance such as diffuse horizontal irradiance (DHI), direct normal irradiance (DNI) and global horizontal irradiance (GHI). *Decomposition* models formulate equations to estimate the solar irradiance components based on the correlations between them. Such formulations are relevant, especially in cases where meteorological data is not adequately available.

DHI is the amount of solar radiation received by a horizontal surface, which has been scattered by the molecules and particles in the atmosphere. It is the part of solar radiation which does not belong to the 5° field of view concentric around the sun. DNI is the direct radiation received on a plane normal to the sun over the total solar spectrum. It is an essential component of global irradiance, especially in cloudless conditions. GHI is the total amount of such terrestrial irradiance which is received by a surface horizontal to the surface of the earth. It can measured with the help of pyranometers, and in general, can be computed from DHI and DNI using the following equation, where θ_z is the *solar zenith angle* (the angle between sun and the vertical):

$$GHI = DHI + DNI.cos(\theta_z) \tag{2}$$

Holmgrem et al [51] contributed to building *pvlib-python*[7] an open source, python-based software, ported from the PVLIB MATLAB toolbox developed at Sandia National Laboratories. This software provides a set of utilities for simulating the performance of the photovoltaic energy systems, with implementations of algorithms related to solar energy. Specifically, it contains components to obtain weather forecast data from NOAA/NCEP/NWS models including the GFS, NAM, RAP, HRRR, and the NDFD, retrieved from the UNIDATA THREDDS servers; and components to convert this weather forecast data into a PV power forecast.

For our experiments, we retrieved a NAM data product (NAM_{awphys}) from the NCEP servers. This is different from the one retrieved by *pvlib-python* (NAM_{awip}) in that the former is a full complement of both the pressure level fields and surface-based fields, while the latter is a full complement of just the surface-based fields. To be able to extend the solar energy related algorithms to the NAM weather forecast data collected by us, making it compatible with NAM_{awip} becomes essential.

NAM_{awip} data retrieved by *pvlib-python* consists of the following surface-level parameters: air temperature, wind speed, total clouds, low clouds, mid clouds and high clouds. In order to enable the use of *pvlib-python* functionalities on the weather forecast dataset collected by us, equivalent surface-level fields were identified in NAM_{awphys}. In this work, each of the irradiance metrics GHI, DHI and DNI were computed from the *pvlib-python* compatible NAM dataset using two empirical solar radiation models implemented in the software: Clear-sky Scaling, Liu-Jordan Model.

[7] `https://github.com/pvlib/pvlib-python`

4.2.1 CLEAR-SKY SCALING

Global horizontal irradiance can be measured with the help of a pyranometer on a horizontal surface. For this reason, it is typically the most common type of irradiance measurement. Knowledge of clear sky conditions, i.e. where the visible clouds are absent, is a key requirement for forecasting terrestrial solar radiation. Empirical solar radiation models such as clear-sky models estimate the solar radiation under a cloudless sky based on various atmospheric parameters. Such models can generally be validated by comparing the estimated irradiance with the measured irradiance in clear-sky conditions.

Several parametric models have been proposed in literature to compute the different components of solar radiation from environmental conditions such as atmospheric turbidity, fractional sunshine, perceptible water vapor, etc. Ineichen et al [52] formulated a model to compute Linke turbidity independent of the airmass, and clear-sky GHI from this metric. In this technique, the *Ineichen Clear-Sky Model* was used to compute the clear-sky GHI (GHI_{CS}). Going by Larson et al's [53] work, this was scaled on the basis of the total cloud cover (TCC) according to the following equation:

$$GHI = GHI_{CS}.[0.35 + 0.65(1 - TCC)] \tag{3}$$

In addition, the popular *Direct Insolation Simulation Code* (DISC) model introduced by Maxwell et al. [50] was used to compute the direct beam component of global solar radiation, i.e. DNI. The diffuse part of global solar radiation, i.e. DHI was computed by subjecting the GHI (estimated with Eq. 3) and DHI to Eq. 2

4.2.2 LIU-JORDAN MODEL

Decomposition models typically utilize only data pertaining to global solar radiation, to estimate the diffused component. They depend on the atmospheric effects in an isolated place, varying according to time of the year, season and climatic conditions [54]. Liu et al. proposed one of the simplest and earliest models of radiation, the Liu-Jordan model [55], which presumes that diffuse radiation intensity is distributed uniformly over the whole sky. In this model, the diffuse irradiance on a surface tilted towards the equator at an angle θ, where I_D is the diffuse radiation on a horizontal surface is given by the following empirical equation:

$$I_{Dt} = I_D.(\frac{1 + cos\theta}{2}) \tag{4}$$

Liu-Jordan model, though simple, is one of the more accurate models for estimating diffuse radiation on inclined surfaces [56]. This model helps determine DNI, GHI from properties such as extraterrestrial flux, transmittance, and optical air mass number. It has been observed that the Liu-Jordan model provides a good fit to empirical data under overcast skies, but underestimates the solar radiation on tilted surfaces when used for partially-clear and clear-sky days [57].

4.3 EXPERIMENT SETUP

In this chapter, two series of experiments are performed towards predicting solar irradiance on each of the dual-axis tracking, fixed-axis and single-axis tracking solar arrays. They can be summarized as follows:

- contrasting the utilization of GHI (from NAM Forecast System), adjusted GHI (from blending NAM Forecast System with *Clear-Sky Scaling* and *Liu-Jordan* techniques) as a predictor variable
  (a) comparing performance of *random forests* utilizing GHI and adjusted GHI independently as predictors
  (b) comparing performance of *random forests* utilizing other input-selected NAM weather variables along with GHI and adjusted GHI as predictors
- contrasting performance of predictive models utilizing *clear-sky index* rather than GHI

In the first series of experiments, a theory-driven bias-correction scheme combining the NAM Forecast System with empirical solar radiation models such as *Clear-Sky Scaling* and *Liu-Jordan* is explored. In Chapter 3, it was observed that the *random forest* was the best performing machine learning model for the purpose of solar irradiance forecasting. Independent predictive models were developed using this algorithm, utilizing GHI estimates from NAM Forecast System and bias-corrected GHI estimates from model-blending as input. These models were compared and analyzed, so as to gauge the effect of the model-blending scheme. In order to enable a consistent comparison of model performance, evaluation metrics used in Chapter 3, i.e. mean absolute error (MAE) and coefficient of determination (R^2) were used in this chapter as well.

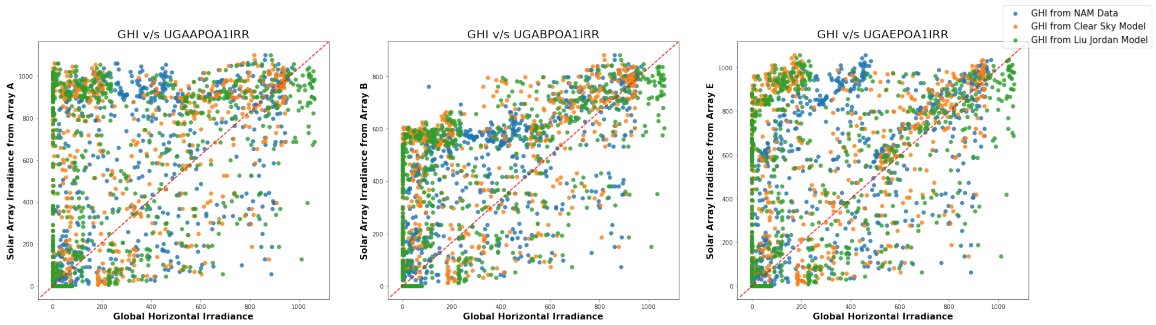


Figure 8: Correlation of the first feature projection corresponding to GHI from NAM data, Clear-sky Scaling and Liu-Jordan Model, with respect to irradiance observations from dual-axis tracking (left), fixed-axis (center) and single-axis tracking (right) solar arrays through 2017.

Furthermore, other NAM weather variables such as air temperature, height at planetary boundary layer, and total cloud cover, which were observed to be effective for solar irradiance prediction were also considered. Predictive models were developed by including these weather variables along with the different variants of GHI described earlier. This weather forecast dataset was input to *random forests* with the input-selection scheme described in 3.2, so as to predict solar irradiance on each of the solar arrays. The performance of these predictive models was compared with models utilizing input-selected NAM weather forecast data, in which the GHI estimates weren't adjusted.

A second series of experiments was performed, where multiple machine learning models were developed using *clear-sky index* rather than GHI, along with the other NAM weather variables. An input-select scheme in line with that in 3.2 was designed to selectively pick relevant feature projections of the weather variables in the forecast horizon. Solar irradiance predictions were made utilizing this weather forecast data on each of the dual-axis tracking, fixed-axis and single-axis tracking solar arrays. A stratified diurnal and seasonal analysis of the performance of these predictive models was performed, to gauge the cyclicity-capturing ability of these models.

4.3.1 BLENDING NAM FORECAST SYSTEM WITH EMPIRICAL MODELS

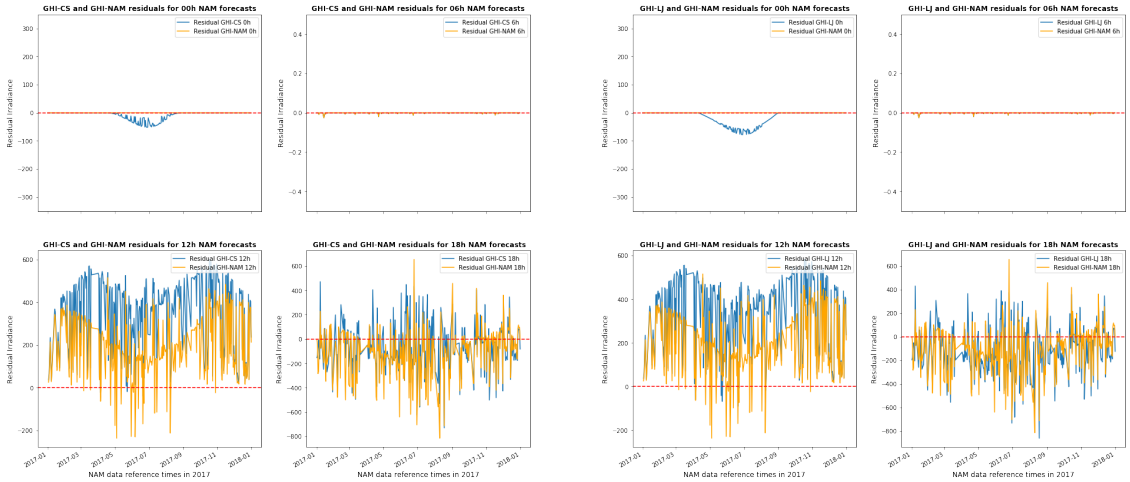
Clear-Sky Scaling and *Liu-Jordan* techniques were used to evaluate GHI (GHI_{CS} and GHI_{LJ} respectively) from the meteorological variables in the NAM Forecast System. Both of these GHI estimates were compared with the GHI from the NAM Forecast System (GHI_{NAM}) for each of the 00h, 06h, 12h and 18h NAM forecasts. In Fig. 9, the residuals of the first feature projection of GHI_{CS} and GHI_{NAM} (left), GHI_{LJ} and GHI_{NAM} (right) with respect to the corresponding irradiance observations from the fixed-axis solar array were plotted for all NAM forecasts in 2017.

Residuals correspond to the difference between the target irradiance observations and the modeled GHI estimates. Thus, in Fig. 9, lying above the X-axis signifies the under-estimation of the corresponding GHI values, and lying under the X-axis signifies the over-estimation of corresponding GHI values. For both the 00h and 06h NAM forecasts, the residuals in both the sub-plots are mostly close to the X-axis, except for a small period for the 00h NAM forecasts. The 00h NAM forecasts correspond to late-evening at the target location. The period for which the residuals corresponding to GHI_{CS} and GHI_{LJ} are under the X-axis signifies the middle of the year where days are longer, thus leading to possible over-estimation of GHI by the empirical techniques. Meanwhile, the 06h NAM forecasts correspond to night-time at the target location, where there is no sun. Thus, the near-zero residual values for these forecasts are justified.

The residuals of GHI estimates from 12h NAM forecasts are consistently above the X-axis, with those corresponding to the GHI estimates from *Clear-Sky Scaling* and *Liu-Jordan* techniques being constantly greater than the residuals corresponding to the GHI estimates from the NAM Forecast System. This illustrates the under-prediction of GHI by all the modeling techniques, with that from the empirical techniques being greater than the NAM Forecast System. Similarly, residuals corresponding to 18h NAM forecasts are studied as well. However in this case, more variability of the GHI estimates was observed, with a considerable number of over-predicted GHI estimates across all the modeling techniques.

To be able to explain the variability in the GHI estimates of the 18h NAM forecasts better, studying measures which determine the amount of cloudiness in the sky is essential. In this regard, *clear-sky index* and *clearness index* were explored. Clear-sky index helps in the removal of diurnal and seasonal signals from a given set of radiation data [2]. This can be attributed to the fact that the *solar zenith angle*, i.e. the elevation angle of the Sun is utilized in the estimation of this measure. In contrast, *Clearness index* helps in estimating the clearness in the sky and can be determined for a specific day based on collected meteorological data and knowledge of extraterrestrial irradiance. It is extremely important in the parameterization of *Liu-Jordan* model.

The *Ineichen Model* was used to compute the clear-sky GHI, which goes into the estimation of *clear-sky index*. For the computation of clearness index, estimation of extraterrestrial radiation for a day of the year, and the estimation of solar zenith angle are necessary. Reda et al. [48] proposed *Solar Position Algorithm*, an implementation of which was used in the determination of the solar zenith angle. The extraterrestrial radiation was determined using an implementation of



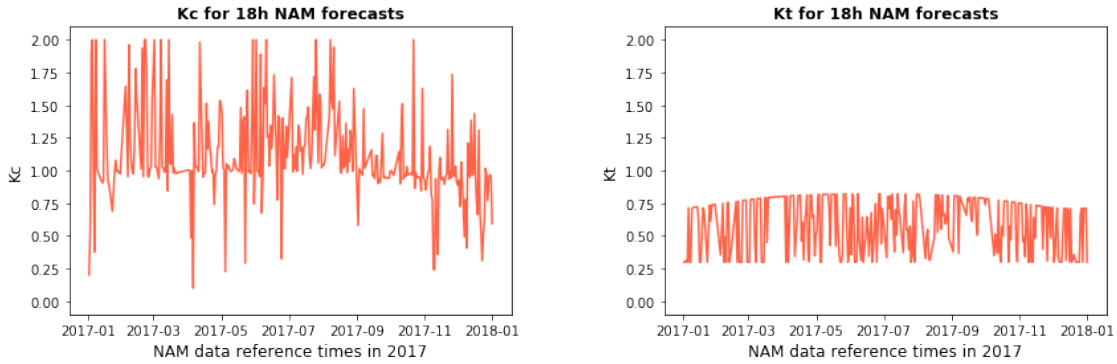
(a) NAM Forecast System + Clear-Sky Scaling (b) NAM Forecast Sytem + Liu Jordan Model

Figure 9: Comparing GHI estimates from Clearsky-Scaling (GHI_{CS}) and Liu-Jordan (GHI_{LJ} techniques with NAM Forecast System (GHI_{NAM} for 00h, 06h, 12h, 18h NAM forecasts: Residuals of GHI_{CS} (blue) and GHI_{NAM} (orange) with respect to fixed-axis solar array irradiance observations for individual weather forecasts in 2017 (left); Residuals of GHI_{LJ} (blue) and GHI_{NAM} (orange) with respect to fixed-axis solar array irradiance observations for individual weather forecasts in 2017 (right).

the algorithm described by Spencer. J in [43]. Both of these measures were formulated such that the negative and non-finite values are truncated to zero, and the maximum value is 2, allowing the over-irradiance events typically seen in sub-hourly data. In Fig. 10, to further study the variability in GHI_{CS} and GHI_{LJ}, clear-sky index and clearness index were plotted for the 18h NAM forecasts.

Generally, clear-sky index values greater than 1 indicate higher solar irradiance observations. It was observed that the clear-sky index for the 18h NAM forecasts were highly variable, and identifying clear-sky periods based on the clear-sky index was not as straightforward. Thus, a randomized cross-validated grid search was performed to find a threshold for the clear-sky index, above which clear-sky periods could be presumed, and GHI_{NAM} could be corrected. The proposed bias-correction involved substituting GHI_{NAM} with the arithmetic mean of GHI_{NAM} and GHI_{CS}. Based on such an approach, it was found that K_c could be thresholded at 0.85. For the sake of convenience, in this work, the GHI estimates corresponding to the model-blending techniques will be referred to as *adjusted GHI*, and those corresponding to model-blending between the NAM Forecast System and *Clear-Sky Scaling* as *adjusted* $GHI_{\mathrm{NAM+CS}}$.

A similar randomized cross-validated grid search was performed to identify a threshold for the clearness index as well. For NAM forecasts with clearness index greater than this threshold,



(a) (b)

Figure 10: Clear-Sky Index (K_c, left) estimates and Clearness Index (K_t, right) estimates for 18h NAM forecasts in 2017.

sunny conditions could be presumed, and below which, cloudy sky conditions could be presumed. Such a threshold was identified at 0.35. Thus, for all 18h NAM forecasts with $K_t > 0.35$ (i.e. possessing sunny sky conditions), GHI_{NAM} was adjusted by substituting it with the arithmetic mean of GHI_{NAM} and GHI_{LJ}. For the sake of convenience, such bias-corrected GHI estimates will be referred to as *adjusted* $GHI_{\mathrm{NAM+LJ}}$.

Random forests were the better performing machine learning models for solar irradiance prediction in Chapter 3. Separate predictive models were developed using this algorithm, utilizing GHI_{NAM}, *adjusted* $GHI_{\mathrm{NAM+CS}}$ and *adjusted* $GHI_{\mathrm{NAM+LJ}}$ alone. The performance of each of these models forecasting solar irradiance on the dual-axis tracking, fixed-axis and single-axis tracking solar arrays was compared.

In addition, such predictive models were also developed by including additional NAM weather variables along with GHI_{NAM}, *adjusted* $GHI_{\mathrm{NAM+CS}}$ and *adjusted* $GHI_{\mathrm{NAM+LJ}}$ respectively. Relevant feature projections were chosen for each of them based on the input-selection scheme described in 3.2. *Liu-Jordan* model has been shown to be effective in predicting diffuse irradiance on inclined surfaces. This was also verified as the DHI estimated by this technique was highly correlated with ground-based solar irradiance observations. Thus, DHI estimated by *Liu-Jordan* model was included as a predictor to the models involving *adjusted* $GHI_{\mathrm{NAM+LJ}}$. Performance of models using these three variants of weather forecast data was compared for the dual-axis tracking, fixed-axis and single-axis tracking solar arrays.

4.3.2 PREDICTIVE MODELING USING CLEAR-SKY INDEX

For each of the 25 GHI feature projections in the NAM Forecast System, corresponding clear-sky GHI estimates were computed using the *Ineichen Clear-Sky Model*. These estimates were used to determine the clear-sky index (K_c) values, resulting in 25 feature projections for this measure. A higher dependency was observed between K_c and solar irradiance observations from the solar farm, as compared to the other NAM weather variables. Thus, following the input-selection scheme described in 3.2, a mutual information matrix was computed for each of the feature projections of K_c with respect to the solar irradiance observations from fixed-axis solar array in the forecast horizon.

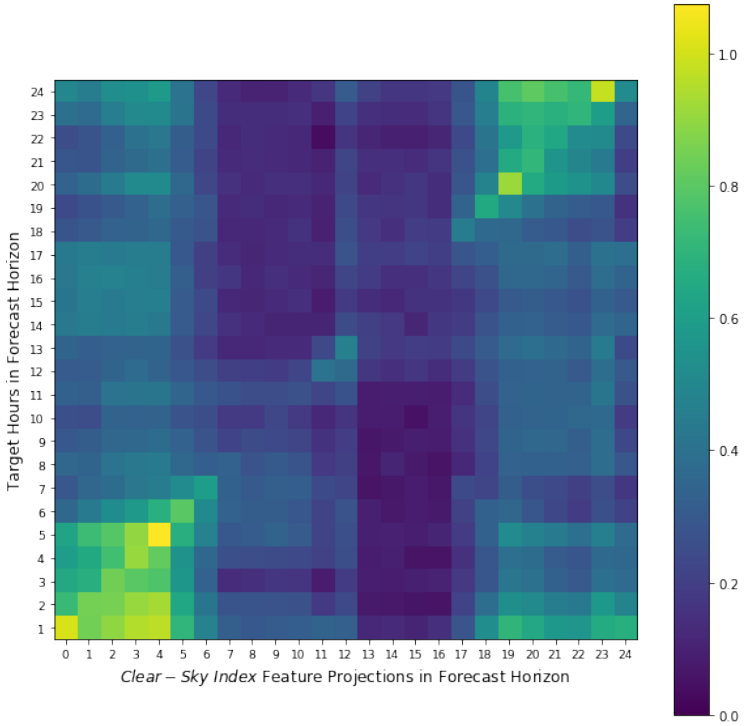


Figure 11: Mutual information between feature projections of Clear-Sky Index (K_c) in the forecast horizon and irradiance observations for corresponding target hours along fixed-axis solar array.

In Fig. 11, it can be seen that the solar irradiance observations are not dependent on all of the K_c feature projections in the forecast horizon. In contrast to the mutual information matrix between the GHI feature projections and solar irradiance observations in the forecast horizon (as in Fig. 8), here, the irradiance for a particlar target hour offset in the forecast horizon doesn't necessarily depend on the K_c feature projection corresponding to the same reference time either. As a matter of fact, a lower dependency was observed between the clear-sky index feature projections and target solar irradiance at the center of the forecast horizon.

In order to find one such optimal range, a cross-validated grid search was performed. It was identified that the feature projections ranging from 13 hours to 17 hours in the forecast horizon were less relevant. Hence, for the machine learning models trained for each of the target hours, these feature projections were omitted. As a consequence of such an input-selection scheme, 20 feature projections corresponding to clear-sky index and the other NAM weather variables such as air temperature, total cloud cover and height at planetary boundary layer were included as predictors for the machine learning models. Eight temporal encodings (four representing the reference time of the observation, four representing the target hour offset from the reference time) depicting the *time of day* and *time of year* were included as well.

Using each of these features as predictors, machine learning models such as *Least-Squares Linear Regression* (LSLR), *k-Nearest Neighbors* (KNN), *Support Vector Regression* (SVR), *Decision Trees* (DT), *Random Forests* (RF) and *Extreme Gradient Boosted Trees* (XGBT) were utilized towards post-processing the solar irradiance from each of the solar arrays. The trivial 24-hour *persistence models* described in Chapter 3 were used as a baseline for this set of experiments as well. For recording the performance of these models, evaluation metrics such as mean absolute error (MAE), coefficient of determination (R^2) and Pearson's correlation coefficient (r) were used.

4.4 RESULTS AND DISCUSSION

Assessing Effect of Multi-Model Blending Approaches

In this chapter, the effectiveness of the multi-model blending approaches, i.e. methodologies which combine the NAM Forecast System with the empirical solar radiation models was assessed. Firstly, the performance of the *random forests* algorithm was compared by taking three versions of weather data as input:

- GHI from NAM Forecast System
- adjusted GHI, from blending NAM Forecast System with Clear-Sky Scaling technique ($adjusted\ GHI_{\mathrm{NAM+CS}}$)
- adjusted GHI, from blending NAM Forecast System with Liu-Jordan model ($adjusted\ GHI_{\mathrm{NAM+LJ}}$)

In Table 8, the performance of these predictive models is reported as methodologies *I*, *II* and

III respectively for each of the dual-axis tracking, fixed-axis and single-axis tracking solar arrays. To enable a smooth comparison with the results reported in chapter 3, and also with the model results from other literature, evaluation metrics such as mean absolute error (MAE), coefficient of determination (R^2) and Pearson's correlation coefficient (r) are reported. The evaluation scheme from Chapter 3, in which the averaged results over 6-hour partitions in the forecast horizon, and the averaged results over the entire forecast horizon is reported, is continued in this chapter too. Both R^2 and r over a forecast horizon are computed in a similar fashion as well.

Random forests utilizing *adjusted* $GHI_{\text{NAM+CS}}$ performed the best, improving upon those using GHI_{NAM} by 4.95%, 4.53% and 4.12% for the dual-axis tracking, fixed axis and single-axis tracking solar arrays respectively. An averaged MAE of 80.61 W/m^2, 49.32 W/m^2 and 67.52 W/m^2 over the entire forecast horizon was reported for each of the solar arrays. An improvement was observed for methodology *III* as well, in which *adjusted* $GHI_{\text{NAM+LJ}}$ was utilized. Using such a weather forecast dataset resulted in a decrease in MAE by 4.17%, 4.14% and 3.62% respectively, with an MAE of 81.27 W/m^2, 49.52 W/m^2 and 67.52 W/m^2 for the solar arrays.

Table 8: Evaluating effect of multi-model blending approaches on irradiance observations from dual-axis tracking, fixed-axis and single-axis tracking solar arrays, using random forests algorithm: *Methodology I* - Using GHI from NAM Forecast System; *Methodology II* - GHI from blending NAM Forecast System and *Clear-Sky Scaling*; *Methodology III* - GHI from blending NAM Forecast System and *Liu-Jordan*.

Metric	**Horizon**	**Dual-Axis Tracking**			**Fixed-Axis**			**Single-Axis Tracking**		
		I	*II*	*III*	*I*	*II*	*III*	*I*	*II*	*III*
MAE	$1-6$	77.57	75.95	76.13	48.12	47.31	46.88	66.07	64.85	64.96
	$6-12$	86.97	83.45	83.78	52.5	51.23	51.46	72.04	69.73	69.61
	$12-18$	86.62	80.43	81.05	53.13	49.42	49.89	71.04	66.1	66.79
	$18-24$	88.06	82.61	84.14	52.87	49.33	49.85	72.52	69.4	70.13
	Overall	84.81	**80.61**	81.27	51.66	**49.32**	49.52	70.42	**67.52**	67.87
R^2	1 - 6	0.86	0.86	0.86	0.91	0.91	0.91	0.87	0.87	0.87
	$6-12$	0.83	0.84	0.83	0.89	0.89	0.89	0.85	0.85	0.85
	12 - 18	0.83	0.84	0.84	0.89	0.9	0.89	0.85	0.87	0.86
	18 - 24	0.82	0.84	0.83	0.88	0.9	0.89	0.85	0.86	0.86
	Overall	0.83	0.84	0.84	0.89	0.9	0.9	0.86	0.86	0.86
r	1 - 6	0.93	0.93	0.93	0.95	0.95	0.95	0.94	0.94	0.93
	$6-12$	0.91	0.91	0.91	0.94	0.94	0.94	0.92	0.92	0.92
	12 - 18	0.91	0.92	0.92	0.94	0.95	0.95	0.93	0.93	0.93
	18 - 24	0.91	0.92	0.91	0.94	0.95	0.95	0.92	0.93	0.93
	Overall	0.91	0.92	0.92	0.94	0.95	0.95	0.93	0.93	0.93
Relative Imp. in MAE (%)	*Overall*	—	4.95%	4.17%	—	4.53%	4.14%	—	4.12%	3.62%

For the next set of experiments, the *random forests* algorithm was combined with the input-selection scheme described in 3.2. This resulted in three versions of input data for the models, in which NAM weather variables such as air temperature, height at planetary boundary layer and total cloud cover were utilized along with the three variants of GHI described earlier. Select feature projections were chosen for each of these meteorological parameters in the predictive models, depending on the target hour offset in the forecast horizon. The results from each of these methodologies is described under *I*, *II* and *III* in Table 9.

It was observed that both the model-blending approaches observed a slight deterioration in performance. The methodology blending NAM Forecast System with *Clear-Sky Scaling* technique recorded an MAE of 72.57 W/m^2, 44.91 W/m^2 and 63.56 W/m^2 for the dual-axis tracking, fixed-axis and single-axis tracking solar arrays respectively. Meanwhile, combining the NAM Forecast System with Liu-Jordan model recorded an MAE of 72.74 W/m^2, 45.25 W/m^2 and 63.97 W/m^2 for each of the solar arrays.

Table 9: Evaluating effect of multi-model blending approaches on irradiance predictions along dual-axis tracking, fixed-axis and single-axis tracking solar arrays, using random forests algorithm: *Methodology I* - Using input-selected NAM Forecast System; *Methodology II* - Blending input-selected NAM Forecast System and *Clear-Sky Scaling*; *Methodology III* - Blending input-selected NAM Forecast System and *Liu-Jordan*.

Metric	**Horizon**	**Dual-Axis Tracking**			**Fixed-Axis**			**Single-Axis Tracking**		
		I	*II*	*III*	*I*	*II*	*III*	*I*	*II*	*III*
MAE	1 - 6	68.5	70.11	70.56	42.26	42.87	43.34	60.5	61.53	62.11
	$6-12$	72.55	72.82	73.05	45.49	45.16	45.37	63.98	63.88	64.41
	12 - 18	70.65	72.59	72.71	44.75	45.22	45.22	62.75	63.67	63.38
	18 - 24	74.97	74.77	74.66	46.65	46.39	47.06	66.34	65.15	65.98
	Overall	**71.67**	72.57	72.74	**44.79**	44.91	45.25	**63.39**	63.56	63.97
R^2	1 - 6	0.88	0.88	0.88	0.92	0.92	0.92	0.89	0.89	0.89
	$6-12$	0.87	0.87	0.87	0.91	0.91	0.91	0.87	0.87	0.87
	12 - 18	0.87	0.87	0.87	0.91	0.91	0.91	0.88	0.88	0.88
	18 - 24	0.86	0.86	0.86	0.9	0.91	0.9	0.87	0.87	0.87
	Overall	0.87	0.87	0.87	0.91	0.91	0.91	0.88	0.88	0.88
r	1 - 6	0.94	0.94	0.94	0.96	0.96	0.96	0.94	0.94	0.94
	$6-12$	0.93	0.93	0.93	0.95	0.96	0.95	0.93	0.93	0.93
	12 - 18	0.93	0.93	0.93	0.95	0.96	0.96	0.94	0.94	0.94
	18 - 24	0.93	0.93	0.93	0.95	0.95	0.95	0.93	0.93	0.93
	Overall	0.93	0.93	0.93	0.95	0.96	0.96	0.94	0.94	0.94

In general, the addition of weather variables has shown to incorporate site-specific information into the predictive models. In Chapter 3, the selective picking of feature projections of NAM weather variables in the forecast horizon improved the performance of the predictive models considerably.

A similar improvement in performance was expected by including additional weather data with the model-blended GHI as well. However, this wasn't the case. This can possibly be attributed to the misidentification of clear-sky conditions by clear-sky index and clearness index respectively. Such a misidentification could have led to an adjustment in GHI, where it wasn't needed.

Performance Evaluation for Predictive Modeling using Clear-Sky Index

In this series of experiments, meteorological variables including relevant NAM weather variables and clear-sky index was utilized. The machine learning models were trained on such weather data corresponding to 2017, and evaluated against that corresponding to 2018. For the dual-axis tracking solar array, all the models performed better than the baseline $t-24$ persistence models, which was expected. *Random forests* performed the best, recording an MAE of 79.58 W/m^2. However, this was worse in comparison to the performance of this model utilizing input-selected NAM forecast data (as in Table 2), for which an MAE of 72.63 W/m^2 was observed. In addition, a considerable degradation in performance was observed in the performance of *support vector regression*, the MAE of which increased from 73.43 W/m^2 to 83.2 W/m^2, and for *k-nearest neighbors*, whose MAE increased from 73.02 W/m^2 to 98.52 W/m^2.

Table 10: Evaluating performance of predictive models using Clear-Sky Index, for irradiance predictions along dual-axis tracking solar array.

Metric	Horizon	PER	LSLR	SVR	KNN	DT	RF	XGBT
MAE	$1-6$	153.32	119.41	80.21	99.59	75.96	78.52	78.52
	$7-12$	153.91	134.82	82.3	94.59	79.9	77.8	79.98
	$13-18$	154.16	118.45	84.08	98.35	82.29	80.44	80.63
	$19-24$	161.43	134.98	86.19	101.54	83.76	81.55	81.36
	Overall	155.71	126.92	83.2	98.52	80.48	**79.58**	80.12
R^2	$1-6$	0.46	0.82	0.87	0.71	0.86	0.86	0.86
	$7-12$	0.45	0.78	0.86	0.74	0.85	0.86	0.85
	$13-18$	0.45	0.82	0.86	0.72	0.84	0.85	0.85
	$19-24$	0.41	0.76	0.84	0.71	0.84	0.85	0.85
	Overall	0.44	0.79	0.85	0.72	0.85	0.85	0.85
r	$1-6$	0.73	0.91	0.93	0.86	0.93	0.93	0.93
	$7-12$	0.73	0.88	0.93	0.87	0.92	0.93	0.92
	$13-18$	0.73	0.91	0.93	0.86	0.92	0.92	0.92
	$19-24$	0.71	0.87	0.92	0.85	0.91	0.92	0.92
	Overall	0.72	0.89	0.93	0.86	0.92	0.93	0.92

k-Nearest Neighbors algorithm does not make any assumptions about the distribution of the data. In this methodology, the feature projections corresponding to GHI are replaced with those

of clear-sky index, which have a lesser bias and variance as compared to the former. Thus, a more rigorous cross-validated grid search can be performed to find a better set of hyperparameters. Doing so might improve the model complexity and help in attaining a better model fit.

The decrease in performance of *support vector regression* can posssibly be attributed to the C hyperparameter, which represents the strength of regularization. With the change in data distribution due to the substitution of GHI feature projections with those corresponding to clear-sky index, the decrease in bias should be compensated with an increase in C. Doing so would decrease the regularization penalty on the predictive models, and might improve the performance. This was not accounted for during the training of the predictive models using clear-sky index.

Table 11: Evaluating performance of predictive models using Clear-Sky Index, for irradiance predictions along fixed-axis solar array.

Metric	**Horizon**	**PER**	**LSLR**	**SVR**	**KNN**	**DT**	**RF**	**XGBT**
	$1-6$	111.23	76.68	50.51	59.50	48.07	48.25	50.62
	$7-12$	111.51	84.61	52.89	60.35	52.34	47.76	51.36
MAE	$13-18$	111.97	81.33	54.87	64.85	53.68	49.55	50.92
	$19-24$	117.15	88.48	56.06	63.35	53.55	51.17	51.87
	Overall	112.96	82.77	53.58	62.01	51.91	**49.18**	51.20
	$1-6$	0.59	0.88	0.91	0.83	0.90	0.91	0.90
	$7-12$	0.58	0.86	0.90	0.82	0.89	0.90	0.89
R^2	$13-18$	0.58	0.87	0.90	0.80	0.89	0.90	0.90
	$19-24$	0.55	0.83	0.88	0.81	0.88	0.90	0.89
	Overall	0.58	0.86	0.90	0.81	0.89	0.90	0.90
	$1-6$	0.79	0.94	0.96	0.91	0.95	0.96	0.95
	$7-12$	0.79	0.93	0.95	0.91	0.95	0.95	0.95
r	$13-18$	0.79	0.93	0.95	0.90	0.94	0.95	0.95
	$19-24$	0.78	0.91	0.94	0.90	0.94	0.95	0.95
	Overall	0.79	0.93	0.95	0.90	0.94	0.95	0.95

Similar trends were observed for the predictions along fixed-axis and single-axis tracking solar arrays as well. *Random forests* performed the best, recording an MAE of 49.18 W/m^2 and 69.98 W/m^2 respectively for each of the solar arrays. This performance was worse when compared with the best accuracy obtained by machine learning models utilizing input-selected NAM weather dataset (as in Table 3 and Table 4). There, an MAE of 44.94 W/m^2 and 63.60 W/m^2 was recorded for both the solar arrays respectively. Though the overall performance of predictive models using clear-sky index worsened, in order to examine the cyclicity-capturing ability of these models, a stratified diurnal and seasonal analysis of the performance of the predictive models with respect to target irradiance predictions on the fixed-axis solar array is performed.

Table 12: Evaluating performance of predictive models using Clear-Sky Index, for irradiance predictions along single-axis tracking solar array.

Metric	Horizon	PER	LSLR	SVR	KNN	DT	RF	XGBT
	$1-6$	128.64	94.40	68.81	85.56	68.16	69.6	70.55
	$7-12$	128.97	107.51	72.16	81.62	72.58	68.46	69.09
MAE	$13-18$	129.25	98.92	72.48	87.56	72.49	69.39	69.50
	$19-24$	135.35	102.31	73.88	86.56	74.25	72.45	72.84
	Overall	130.55	100.78	71.83	85.33	71.87	**69.98**	70.50
	$1-6$	0.53	0.85	0.87	0.75	0.86	0.87	0.86
	$7-12$	0.53	0.81	0.86	0.77	0.85	0.86	0.86
R^2	$13-18$	0.53	0.84	0.86	0.74	0.85	0.86	0.86
	$19-24$	0.49	0.83	0.86	0.75	0.85	0.86	0.86
	Overall	0.52	0.82	0.85	0.74	0.84	0.85	0.85
	$1-6$	0.77	0.92	0.94	0.88	0.93	0.93	0.93
	$7-12$	0.77	0.90	0.93	0.89	0.92	0.93	0.92
r	$13-18$	0.75	0.92	0.93	0.87	0.92	0.93	0.93
	$19-24$	0.75	0.91	0.92	0.87	0.92	0.92	0.92
	Overall	0.76	0.91	0.93	0.87	0.92	0.93	0.93

Stratified Diurnal and Seasonal Analysis of Performance

The stratified diurnal analysis conducted in Chapter 3 was extended to the predictive models utilizing clear-sky index as well. In Fig. 12, the performance of the machine learning models for all target hours in the forecast horizon, were plotted for each of the 00h, 06h, 12h and 18h UTC NAM forecasts individually. The time of day was presumed to be between 6 A.M and 6 P.M at the target location, and was highlighted in yellow. In general, as was expected, it was observed that most of the models were able to detect the period of darkness, i.e. night-time relatively well.

For the period of day, the more sophisticated machine learning algorithms like *k-nearest neighbors*, *support vector regression* and *random forests* performed better than *decision trees*, and understandably, *linear regression*. However, interestingly enough, *extreme gradient boosted trees* failed to capture day-time well for all the NAM forecasts. This can possibly be attributed to a lesser number of decision trees being used in the ensemble technique. In addition, the diurnal performance of the best performing *random forests* with respect to its performance when utilizing the input-selected NAM weather forecast data (as shown in Fig. 6) can be summarized in the following way:

- performed worse for all the target hours in the day-time for the 00h NAM forecasts
- performed better for the target hours 8 through 12 in the forecast horizon, for the 06h NAM forecasts

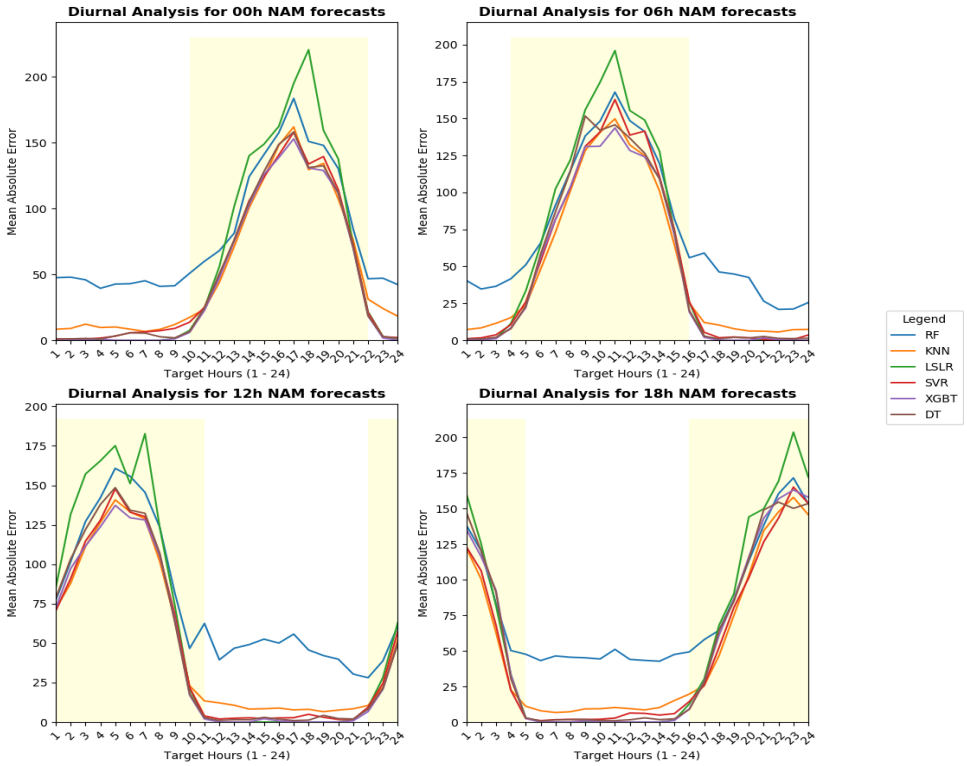


Figure 12: Stratified diurnal analysis of day-ahead irradiance predictions using Clear-Sky Index for fixed-axis solar array: (left-top) 00h NAM forecasts, (right-top) 16h NAM forecasts, (left-bottom) 12h NAM forecasts, (right-bottom) 18h NAM forecasts. Local time of day (6A.M to 6P.M) at the target location for each of the NAM forecasts is indicated in light yellow.

- performed better for the target hours 22 through 24 in the forecast horizon, for the 12h NAM forecasts
- performed worse for all the target hours in the day-time for the 18h NAM forecasts

The target location, i.e. Athens, Georgia is -5.00 hours with respect to UTC in the standard time zone, and -4.00 hours with respect to UTC during *daylight saving time*. To be able to conduct a uniform analysis of the individual NAM forecasts, it was assumed that the local time at the target location is constantly -4.00 hours relative to UTC.

Based on this assumption, as was done in Chapter 3, the performance of different predictive models was compared by analyzing their residuals corresponding to the target hour in the forecast horizon, representing noon, i.e. 12 P.M locally at Athens, Georgia. In Fig. 13, box-and-whisker

plots were drawn corresponding to the residuals from each of the predictive models, so as to study their distributional characteristics. In the figure, the size of the box-plots for most of the models is comparable. In addition, the spread of the residuals beyond the whiskers is minimum for *random forests*, indicating that it is the better machine learning technique for this variant of weather data.

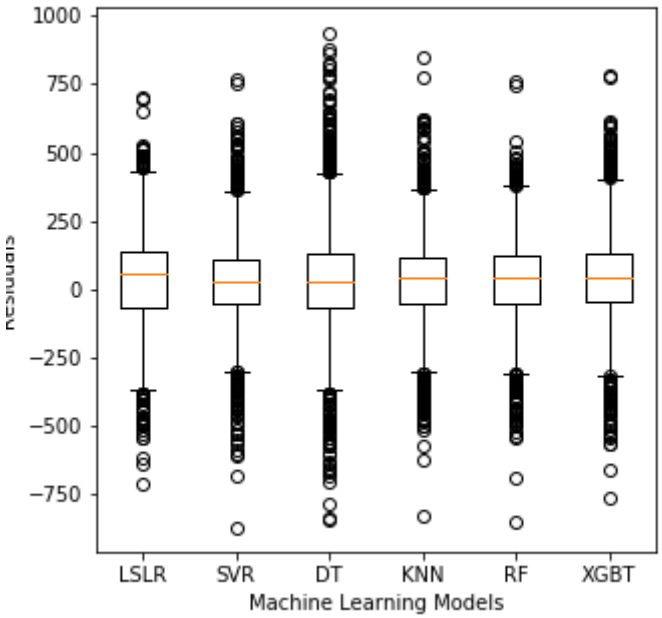


Figure 13: Comparison of box-and-whisker plots of residuals from different predictive models utilizing clear-sky index at 12 P.M local time, i.e. noon.

Furthermore, a stratified seasonal analysis was conducted for NAM forecasts. It was extended to the 12h and 18h NAM forecasts, which, under the above assumption, represent 8 A.M and 12 P.M locally. Based on the general seasonal trends in the target location i.e. Athens, Georgia, the periods in a year were divided into four seasons: *summer* (May - July), *autumn* (August - October), *winter* (November - January) and *spring* (February - April).

In Table 13, the performance of the better-performing *random forests* across each of these seasons was compared. The MAE corresponding to predictions of the models utilizing NAM data involving both GHI and clear-sky index was included. It can be observed that the predictive models using clear-sky index performed poorly for both the 12h and 18h NAM forecasts across all the seasons. Moreover, as was noted earlier in the diurnal analysis, the 18h NAM forecasts performed poorly for all the target hours in the forecast horizon. Owing to the relatively poor performance of

Table 13: Comparing seasonal performance (in MAE) of random forests using input-selected NAM data with GHI (left) and clear-sky index (right) for 12h, 18h NAM forecasts.

Model	**Hour**	**NAM using GHI / NAM using Clear-Sky Index**			
		Summer	*Autumn*	*Winter*	*Spring*
LSLR	12h	60.21 / 67.41	86.35 / 106.9	99.71 / 121.9	119.4 / 138.1
	18h	84.93 / 112.9	95.13 / 121.0	70.82 / 91.61	38.68 / 49.09
SVR	12h	46.95 / 64.11	64.01 / 74.95	72.02 / 88.05	84.52 / 102.6
	18h	76.11 / 102.4	78.75 / 94.32	49.03 / 60.49	12.39 / 19.49
DT	12h	74.60 / 90.51	104.4 / 107.9	101.8 / 116.4	128.3 / 171.6
	18h	110.8 / 127.7	109.2 / 124.8	57.45 / 85.56	24.34 / 36.59
KNN	12h	54.16 / 58.54	81.13 / 83.87	89.93 / 90.19	98.44 / 87.89
	18h	87.39 / 102.5	87.52 / 102.4	62.32 / 74.12	16.25 / 18.76
RF	12h	56.15 / 75.93	82.91 / 101.9	95.77 / 114.0	107.5 / 123.0
	18h	80.52 / 139.3	81.73 / 119.8	56.84 / 94.25	10.47 / 29.03
XGBT	12h	54.07 / 74.41	86.37 / 98.35	101.1 / 123.3	119.2 / 126.0
	18h	80.66 / 134.9	85.68 / 128.1	56.24 / 96.10	10.50 / 27.75

the NAM forecasts involving clear-sky index, as against those involving GHI, it can be concluded that using the clear-sky index as a predictor in the machine learning models failed to improve the diurnal and seasonal trend capturing ability of the models.

CHAPTER 5

CONCLUSION & FUTURE DIRECTIONS

The main purpose of this thesis was to develop machine learning models to effectively predict surface-level solar irradiance 24 hours into the future at multiple fixed and tracking solar arrays located at a solar farm near the University of Georgia. Towards this end, firstly, a study was conducted where the work done by Jones [58] was replicated. An input-selection scheme was designed to weed out the less relevant weather variables, and corresponding feature projections. This scheme helped in improving the performance (mean absolute error, MAE) in the replication study by 19.05%, 19.68% and 10.65% (average across machine learning models) for the dual-axis tracking, fixed-axis and single-axis tracking solar arrays respectively. A best performance, i.e. least MAE of 72.63 W/m^2, 44.94 W/m^2 and 63.60 W/m^2 was recorded for each of the arrays.

The effect of geographic expansion, i.e. including weather forecasts around the target location was evaluated. This was extended to 3 x 3 and 5 x 5 *geo shapes*, by including the input-selected NAM weather forecasts from these additional cells. It was observed that an improvement in performance (marginal) with an increase in the *geo shape* was seen for only the *random forests* algorithm. By utilizing the weather forecast data corresponding to the 5 x 5 *geo shape* as input for this machine learning technique, an MAE of 69.38 W/m^2, 43.62 W/m^2 and 61.99 W/m^2 was recorded for the dual-axis tracking, fixed-axis and single-axis tracking solar arrays.

A few theory-driven bias-correction methodologies (multi-model blending approaches) were explored in Chapter 4. The motivation behind these approaches was to selectively correct the bias in global horizontal irradiance (GHI) reported in literature, by identifying the clear-sky conditions effectively. For this purpose, measures such as *clear-sky index* and *clearness index* were used depending on the empirical solar radiation model that the NAM Forecast System was being combined with. An exhaustive grid search was performed to identify a threshold for these measures, based on which the clear-sky conditions could be distinguished from cloudy-sky conditions easily. Based on these thresholds, the GHI weather variable was selectively corrected for the 18h NAM forecasts, for which an over-prediction of this parameter was observed. This was carried out by substituting the

GHI for such observations with the arithmetic mean of GHI from the NAM Forecast System and the corresponding empirical solar radiation model.

Such a bias-correction scheme resulted in an improvement in performance for the predictive models using the random forests technique, utilizing adjusted GHI from the model-blending techniques, with respect to just using the GHI from NAM Forecast System. A reduction in MAE by 4.95%, 4.53% and 4.12% for the dual-axis tracking, fixed axis and single-axis tracking solar arrays was observed for the model-blending methodology combining the NAM Forecast System and Clear-Sky Scaling technique, over using GHI from just the NAM Forecast System. Additionally, the model-blending methodology combining the NAM Forecast System with the Liu-Jordan model reduced the MAE by 4.17%, 4.14% and 3.62% for each of the solar arrays.

To further evaluate the model-blending approach, other NAM weather variables which were identified to be relevant, i.e. air temperature, height at planetary boundary layer and total cloud cover were included along with the adjusted GHI (obtained through the model-blending approaches). Select feature projections were chosen for each of the weather variables depending on the target hour offset in the forecast horizon, in line with the input-selection scheme described in 3.2. However, it was observed that such an input-selection scheme slightly depreciated the performance of the model-blending approaches combining the NAM Forecast System with both *Clear-Sky Scaling* and *Liu-Jordan* techniques.

The lack of improvement possibly indicates an inability to adequately identify sky conditions effectively, which in turn prevents accurate bias correction. In this work, the Ineichen model was used to determine the clear-sky GHI. It would be interesting to see if utilizing other empirical clear-sky models for this purpose will improve the ability of the *clear-sky index* measure to identify the sky conditions, and in turn, improve model performance. This also leaves a scope for exploring superior techniques for distinguishing between sky conditions, and in turn, for identifying the over-prediction in the GHI weather variable.

In the theory-based bias correction methodology described in this work, a simple bias-correction function was used, wherein, the GHI from the NAM Forecast System was substituted with the arithmetic mean of GHI from the NAM Forecast System and the corresponding empirical solar radiation model. This can be improved upon by subjecting both of these GHI estimates to statistical post-processing, and determining a superior bias-correction function. This can be investigated in future work.

By utilizing the meteorological projections in NAM Forecast System, *clear-sky index* was projected into the future as well. Predictive models were developed utilizing this measure rather than GHI, in order to exploit it's presumed ability to capture seasonality. A best performance, i.e. least MAE of 79.58 W/m^2, 49.18 W/m^2 and 69.98 W/m^2 was recorded for the dual-axis tracking, fixed-axis and single-axis tracking solar arrays respectively. In order to assess the seasonality-capturing ability, a stratified seasonal analysis was performed, where the performance of individual forecasts across seasons (summer, spring, winter, autumn) was compared with that of the models utilizing GHI. It was observed that the former performed poorly across seasons, when compared to that of the latter. Consequently, it can be concluded that the presumed cyclicity-capturing ability of *clear-sky index* did not translate into improving the performance of the predictive models. It would be interesting to explore other clear-sky models in literature towards determining this measure, and reviewing their performance in such a framework. This can be looked into in further work.

REFERENCES

[1] Elke Lorenz et al. "Irradiance forecasting for the power prediction of grid-connected photovoltaic systems". In: *IEEE Journal of selected topics in applied earth observations and remote sensing* 2.1 (2009), pp. 2–10.

[2] NA Engerer and FP Mills. "KPV: A clear-sky index for photovoltaics". In: *Solar energy* 105 (2014), pp. 679–693.

[3] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. "Estimating mutual information". In: *Physical review E* 69.6 (2004), p. 066138.

[4] Brian C Ross. "Mutual information between discrete and continuous data sets". In: *PloS one* 9.2 (2014).

[5] Ram Hashmonay, Ariel Cohen, and Uri Dayan. "Lidar observation of the atmospheric boundary layer in Jerusalem". In: *Journal of Applied Meteorology* 30.8 (1991), pp. 1228–1236.

[6] Axel von Engeln and João Teixeira. "A planetary boundary layer height climatology derived from ECMWF reanalysis data". In: *Journal of Climate* 26.17 (2013), pp. 6575–6590.

[7] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[8] Suneetha Racharla and K. Rajan. "Solar tracking system – a review". In: *International Journal of Sustainable Engineering* 10.2 (2017), pp. 72–81. DOI: 10.1080/19397038.2016.1267816. eprint: https://www.tandfonline.com/doi/pdf/10.1080/19397038.2016.1267816. URL: https://www.tandfonline.com/doi/abs/10.1080/19397038.2016.1267816.

[9] Chi Wai Chow et al. "Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed". In: *Solar Energy* 85.11 (2011), pp. 2881–2893.

[10] Ricardo Marquez and Carlos FM Coimbra. "Intra-hour DNI forecasting based on cloud tracking image analysis". In: *Solar Energy* 91 (2013), pp. 327–336.

[11] *NCEP Global Forecast System (GFS) Analyses and Forecasts*. Boulder CO, 2007. URL: https://doi.org/10.5065/D65Q4TSG.

[12] *NCEP North American Mesoscale (NAM) 12 km Analysis*. Boulder CO, 2015. URL: https://doi.org/10.5065/G4RC-1N91.

[13] Patrick Mathiesen and Jan Kleissl. “Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States”. In: *Solar Energy* 85.5 (2011), pp. 967–977.

[14] José A Ruiz-Arias et al. “Optimal combination of gridded and ground-observed solar radiation data for regional solar resource assessment”. In: *Solar Energy* 112 (2015), pp. 411–424.

[15] Elke Lorenz et al. “Benchmarking of different approaches to forecast solar irradiance”. In: *24th European Photovoltaic Solar Energy Conference* (Jan. 2009).

[16] Richard Perez et al. “Validation of short and medium term operational solar radiation forecasts in the US”. In: *Solar Energy* 84.12 (2010), pp. 2161–2172.

[17] Maimouna Diagne et al. “Review of solar irradiance forecasting methods and a proposition for small-scale insular grids”. In: *Renewable and Sustainable Energy Reviews* 27 (2013), pp. 65–76.

[18] Matthew J Reno and Clifford W Hansen. “Identification of periods of clear sky irradiance in time series of GHI measurements”. In: *Renewable Energy* 90 (2016), pp. 520–531.

[19] Christian A Gueymard. “REST2: High-performance solar radiation model for cloudless-sky irradiance, illuminance, and photosynthetically active radiation–Validation with a benchmark dataset”. In: *Solar Energy* 82.3 (2008), pp. 272–285.

[20] Pierre Ineichen. “A broadband simplified version of the Solis clear sky model”. In: *Solar Energy* 82.8 (2008), pp. 758–762.

[21] RW Mueller et al. “Rethinking satellite-based solar irradiance modelling: The SOLIS clear-sky module”. In: *Remote sensing of Environment* 91.2 (2004), pp. 160–174.

[22] Matthew J Reno, Clifford W Hansen, and Joshua S Stein. “Global horizontal irradiance clear sky models: Implementation and analysis”. In: ().

[23] Pierre Ineichen and Richard Perez. “A new airmass independent formulation for the Linke turbidity coefficient”. In: *Solar Energy* 73.3 (2002), pp. 151–157.

[24] Hyun-Jin Lee, Shin-Young Kim, and Chang-Yeol Yun. “Comparison of solar radiation models to estimate direct normal irradiance for Korea”. In: *Energies* 10.5 (2017), p. 594.

[25] LT Wong and WK Chow. “Solar radiation model”. In: *Applied energy* 69.3 (2001), pp. 191–224.

[26] Christian A Gueymard. “Direct solar transmittance and irradiance predictions with broadband models. Part I: detailed theoretical performance assessment”. In: *Solar Energy* 74.5 (2003), pp. 355–379.

[27] J Chandrasekaran and S Kumar. “Hourly diffuse fraction correlation at a tropical location”. In: *Solar Energy* 53.6 (1994), pp. 505–510.

[28] Seyed Abbas Mousavi Maleki, H Hizam, and Chandima Gomes. “Estimation of hourly, daily and monthly global solar radiation on inclined surfaces: Models re-visited”. In: *Energies* 10.1 (2017), p. 134.

[29] Benjamin YH Liu and Richard C Jordan. “The interrelationship and characteristic distribution of direct, diffuse and total solar radiation”. In: *Solar energy* 4.3 (1960), pp. 1–19.

[30] F Antonanzas-Torres et al. “Clear sky solar irradiance models: A review of seventy models”. In: *Renewable and Sustainable Energy Reviews* 107 (2019), pp. 374–387.

[31] Richard E Bird and Roland L Hulstrom. *Simplified clear sky model for direct and diffuse insolation on horizontal surfaces*. Tech. rep. Solar Energy Research Inst., Golden, CO (USA), 1981.

[32] Daniel Cano et al. “A method for the determination of the global solar radiation from meteorological satellite data”. In: *Solar energy* 37.1 (1986), pp. 31–39.

[33] Christelle Rigollier, Mireille Lefèvre, and Lucien Wald. “The method Heliosat-2 for deriving shortwave solar radiation from satellite images”. In: *Solar energy* 77.2 (2004), pp. 159–169.

[34] Jethro Betcke et al. “Energy-Specific Solar Radiation Data from Meteosat Second Generation (MSG): The Heliosat-3 Project, Final Report”. In: (Jan. 2006). DOI: 10.13140/RG.2.1.2054.6406.

[35] Annette Hammer et al. “Solar energy assessment using remote sensing technologies”. In: *Remote Sensing of Environment* 86.3 (2003), pp. 423–432.

[36] Jan Kleissl, Carlos Coimbra, and Ricardo Marquez. “Overview of Solar Forecasting Methods and a Metric for Accuracy Evaluation”. In: July 2013. ISBN: 9780123971777. DOI: 10.1016/B978-0-12-397177-7.00008-5.

[37] Ricardo Marquez and Carlos FM Coimbra. “Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the NWS database”. In: *Solar Energy* 85.5 (2011), pp. 746–756.

[38] Gordon Reikard. “Predicting solar radiation at high resolutions: A comparison of time series forecasts”. In: *Solar Energy* 83.3 (2009), pp. 342–349.

[39] LL Mora-Lopez and M Sidrach-de-Cardona. “Multiplicative ARMA models to generate hourly series of global irradiation”. In: *Solar Energy* 63.5 (1998), pp. 283–291.

[40] Manajit Sengupta et al. “Best practices handbook for the collection and use of solar resource data for solar energy applications”. In: (2015), pp. 7–11.

[41] Christoph Marty and Rolf Philipona. “The clear-sky index to separate clear-sky from cloudy-sky situations in climate research”. In: *Geophysical Research Letters* 27.17 (2000), pp. 2649–2652.

[42] Dazhi Yang. “Choice of clear-sky model in solar forecasting”. In: *Journal of Renewable and Sustainable Energy* 12.2 (2020), p. 026101.

[43] J. W. Spencer. “Fourier series representation of the position of the sun”. In: *Search* 2.5 (May 1971), pp. 172+. URL: http://www.mail-archive.com/sundial@uni-koeln.de/msg01050.html.

[44] Manajit Sengupta et al. “Best practices handbook for the collection and use of solar resource data for solar energy applications”. In: (2015), pp. 7–4.

[45] *NCEP North American Mesoscale (NAM) 12 km Analysis*. Boulder CO, 2015. URL: https://doi.org/10.5065/G4RC-1N91.

[46] Patrick Mathiesen and Jan Kleissl. “Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States”. In: *Solar Energy* 85 (May 2011), pp. 967–977. DOI: 10.1016/j.solener.2011.02.013.

[47] William Cotton, George Bryan, and Susan van den Heever. “Chapter 5. Radiative Transfer in a Cloudy Atmosphere and Its Parameterization”. In: *International Geophysics* 99 (Dec. 2011), pp. 143–175. DOI: 10.1016/S0074-6142(10)09911-0.

[48] Ibrahim Reda and Afshin Andreas. “Solar position algorithm for solar radiation applications”. In: *Solar energy* 76.5 (2004), pp. 577–589.

[49] Sam Sanders et al. “Solar Radiation Prediction Improvement Using Weather Forecasts”. In: Dec. 2017, pp. 499–504. DOI: 10.1109/ICMLA.2017.0-112.

[50] Eugene L Maxwell. *A quasi-physical model for converting hourly global horizontal to direct normal insolation*. Tech. rep. Solar Energy Research Inst., Golden, CO (USA), 1987.

[51] William Holmgren, Clifford Hansen, and Mark Mikofski. “pvlib python: a python package for modeling solar energy systems”. In: *Journal of Open Source Software* 3.29 (2018), p. 884. DOI: 10.21105/joss.00884. URL: https://doi.org/10.21105/joss.00884.

[52] Pierre Ineichen and Richard Perez. "A new airmass independent formulation for the Linke turbidity coefficient". In: *Solar Energy* 73.3 (2002), pp. 151–157. ISSN: 0038-092X. DOI: https://doi.org/10.1016/S0038-092X(02)00045-2. URL: http://www.sciencedirect.com/science/article/pii/S0038092X02000452.

[53] David P Larson, Lukas Nonnenmacher, and Carlos FM Coimbra. "Day-ahead forecasting of solar power output from photovoltaic plants in the American Southwest". In: *Renewable Energy* 91 (2016), pp. 11–20.

[54] Joseph C Lam and Danny HW Li. "Correlation between global solar radiation and its direct and diffuse components". In: *Building and environment* 31.6 (1996), pp. 527–535.

[55] Benjamin YH Liu and Richard C Jordan. "The interrelationship and characteristic distribution of direct, diffuse and total solar radiation". In: *Solar energy* 4.3 (1960), pp. 1–19.

[56] Seyed Abbas Mousavi Maleki, H Hizam, and Chandima Gomes. "Estimation of hourly, daily and monthly global solar radiation on inclined surfaces: Models re-visited". In: *Energies* 10.1 (2017), p. 134.

[57] Thomas M Klucher. "Evaluation of models to predict insolation on tilted surfaces". In: *Solar energy* 23.2 (1979), pp. 111–114.

[58] Zachary Dean Jones. "Machine Learning For Solar Irradiance Forecasting". MA thesis. Athens, GA: The University of Georgia, 2019.

APPENDIX A

MODEL HYPERPARAMETERS

A.1 1 x 1 Grid Size

A.1.1 Dual-Axis Tracking Solar Array

- Support Vector Regression
  - C : 1000, *epsilon* : 1, *gamma* : 0.01
- Decision Tree
  - *criterion* : 'mae', *splitter* : 'random', *max_depth* : 10, *min_impurity_decrease* : 0.25
- k-Nearest Neighbors
  - *n_neighbors* : 11, *leaf_size* : 20, p : 1
- Random Forests
  - *n_estimators* : 650, *max_depth* : 20, *min_impurity_decrease* : 0.1
- Extreme Gradient Boosted Trees
  - *n_estimators* : 700, *max_depth* : 5, *learning_rate* : 0.01

A.1.2 Fixed-Axis Solar Array

- Support Vector Regression
  - C : 500, *epsilon* : 1, *gamma* : 0.01
- Decision Tree
  - *criterion* : 'mae', *splitter* : 'random', *max_depth* : 10, *min_impurity_decrease* : 0.25
- k-Nearest Neighbors
  - *n_neighbors* : 11, *leaf_size* : 25, p : 1
- Random Forests
  - *n_estimators* : 500, *max_depth* : 40, *min_impurity_decrease* : 0.2
- Extreme Gradient Boosted Trees
  - *n_estimators* : 600, *max_depth* : 5, *learning_rate* : 0.01

A.1.3 Single-Axis Tracking Solar Array

- Support Vector Regression
  - C : 500, *epsilon* : 1, *gamma* : 0.01
- Decision Tree
  - *criterion* 'mae': , *splitter* : 'random', *max_depth* : 10, *min_impurity_decrease* : 0.25
- k-Nearest Neighbors
  - *n_neighbors* : 11, *leaf_size* : 5, p : 1
- Random Forests
  - *n_estimators* : 300, *max_depth* : 20, *min_impurity_decrease* : 0.1
- Extreme Gradient Boosted Trees
  - *n_estimators* : 600, *max_depth* : 5, *learning_rate* : 0.01

A.2 3 x 3 Grid Size

A.2.1 Dual-Axis Tracking Solar Array

- Support Vector Regression
  - C : 500, *epsilon* : 2, *gamma* : 0.01
- Decision Tree
  - *criterion* : 'mae', *splitter* : 'random', *max_depth* : 10, *min_impurity_decrease* : 0.25
- k-Nearest Neighbors
  - *n_neighbors* : 9, *leaf_size* : 40, p : 1
- Random Forests
  - *n_estimators* : 500, *max_depth* : 20, *min_impurity_decrease* : 0.2
- Extreme Gradient Boosted Trees
  - *n_estimators* : 550, *max_depth* : 5, *learning_rate* : 0.01

A.2.2 Fixed-Axis Solar Array

- Support Vector Regression
  - C : 500, *epsilon* : 3, *gamma* : 0.01
- Decision Tree
  - *criterion* : 'mae', *splitter* : 'random', *max_depth* : 10, *min_impurity_decrease* : 0.25

- k-Nearest Neighbors
  - *n_neighbors* : 9, *leaf_size* : 60, *p* : 1
- Random Forests
  - *n_estimators* : 550, *max_depth* : 40, *min_impurity_decrease* : 0.1
- Extreme Gradient Boosted Trees
  - *n_estimators* : 700, *max_depth* : 5, *learning_rate* : 0.01

A.2.3 Single-Axis Tracking Solar Array

- Support Vector Regression
  - *C* : 1000, *epsilon* : 1, *gamma* : 0.01
- Decision Tree
  - *criterion* : 'mae', *splitter* : 'random', *max_depth* : 10, *min_impurity_decrease* : 0.25
- k-Nearest Neighbors
  - *n_neighbors* : 9, *leaf_size* : 60, *p* : 1
- Random Forests
  - *n_estimators* : 300, *max_depth* : 20, *min_impurity_decrease* : 0.1
- Extreme Gradient Boosted Trees
  - *n_estimators* : 600, *max_depth* : 5, *learning_rate* : 0.01

A.3 5 x 5 Grid Size

A.3.1 Dual-Axis Tracking Solar Array

- Support Vector Regression
  - *C* : 500, *epsilon* : 3, *gamma* : 0.01
- Decision Tree
  - *criterion* : 'mae', *splitter* : 'random', *max_depth* : 10, *min_impurity_decrease* : 0.25
- k-Nearest Neighbors
  - *n_neighbors* : 9, *leaf_size* : 30, *p* : 1
- Random Forests
  - *n_estimators* : 550, *max_depth* : 60, *min_impurity_decrease* : 0.4
- Extreme Gradient Boosted Trees
  - *n_estimators* : 200, *max_depth* : 5, *learning_rate* : 0.01

A.3.2 Fixed-Axis Solar Array

- Support Vector Regression
  - C : 500, *epsilon* : 2, *gamma* : 0.01
- Decision Tree
  - *criterion* : 'mae', *splitter* : 'random', *max_depth* : 10, *min_impurity_decrease* : 0.25
- k-Nearest Neighbors
  - *n_neighbors* : 9, *leaf_size* : 20, p : 1
- Random Forests
  - *n_estimators* : 700, *max_depth* : 50, *min_impurity_decrease* : 0.3
- Extreme Gradient Boosted Trees
  - *n_estimators* : 200, *max_depth* : 5, *learning_rate* : 0.01

A.3.3 Single-Axis Tracking Solar Array

- Support Vector Regression
  - C : 1000, *epsilon* : 2, *gamma* : 0.01
- Decision Tree
  - *criterion* : 'mae', *splitter* : 'random', *max_depth* : 10, *min_impurity_decrease* : 0.25
- k-Nearest Neighbors
  - *n_neighbors* : 10, *leaf_size* : 20, p : 1
- Random Forests
  - *n_estimators* : 300, *max_depth* : 20, *min_impurity_decrease* : 0.1
- Extreme Gradient Boosted Trees
  - *n_estimators* : 200, *max_depth* : 5, *learning_rate* : 0.01