MACHINE LEARNING FOR SOLAR IRRADIANCE FORECASTING

by

ZACHARY DEAN JONES

(Under the Direction of Frederick Maier)

Abstract

Machine learning techniques were investigated to forecast surface-level solar irradiance and, by proxy, the output of a solar farm near Athens, GA, for 1-24 hours into the future. Weather predictions from the National Oceanic and Atmospheric Administration (NOAA) and ground-based weather observations from the Georgia Automated Environmental Monitoring Network (GAEMN) served as inputs to the models. Various learning algorithms were compared, and an analysis of the relative importance of potential model inputs was performed. Also examined was the ability of machine learning models to generalize to unfamiliar weather and climate conditions. In the experiments, random forests outperform several other techniques commonly found in the literature by as much as 26.8%. The random forest models benefit from incorporating weather forecasts from a grid of cells surrounding the target site, resulting in improvements to accuracy up to 4.1% relative to models that do not make use of such information.

Index Words: machine learning, forecasting, solar power, irradiance, random forest, gradient-boosting

MACHINE LEARNING FOR SOLAR IRRADIANCE FORECASTING

By

ZACHARY DEAN JONES

B.S., Marshall University, 2017

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

2019

© 2019

Zachary Dean Jones

All Rights Reserved

MACHINE LEARNING FOR SOLAR IRRADIANCE FORECASTING

By

ZACHARY DEAN JONES

Major Professor: Frederick Maier

Committee: Khaled Rasheed
Shannon Quinn

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2019

Contents

1 Solar Irradiance Forecasting **1**

1.1 Introduction 1

1.2 Literature Review 4

2 Analysis of Feature Set Importance and Geographic Portability **10**

2.1 Introduction 10

2.2 Methodology 12

2.3 Results 16

2.4 Conclusion 31

3 Predictive Models for Stationary & Non-Stationary Solar Arrays **34**

3.1 Introduction 34

3.2 Data Collection & Preprocessing 36

3.3 Experimental Setup 43

3.4 Results 44

3.5 Conclusion 53

4 Conclusion & Future Directions **56**

4.1 Conclusion 56

4.2 Future Directions 57

Bibliography **60**

1 Solar Irradiance Forecasting

1.1 Introduction

Research and development of renewable energy has seen a rapid increase in recent years along with rising public awareness of the impact of energy production on climate change. Although there are many sources of renewable energy, solar photovolataic (PV) energy (the conversion of sunlight into electrical energy) has attracted significant attention in recent decades due to its abundance and environmental friendliness, and many experts believe that solar energy will play a key role in meeting the world's ever-growing energy demands [9, 16, 17]. The International Energy Agency predicts that solar electricity will become the predominant source of renewable energy production by 2050 [1], and the contribution of solar energy production to the power supply grid is constantly increasing. Global PV capacity has grown more than ten-fold in the past decade, exceeding 400 GW of global capacity at the end of 2017 [21].

Despite the rapid growth of renewable energy sources, there are several challenges associated with their integration into the grid. Solar PV power systems in particular suffer from a dependence on environmental conditions. The quantity of energy produced by a solar farm is heavily dependent on the angle and intensity of sunlight that strikes the photo-voltaic panels, which is in turn determined by the meteorological and and environmental conditions of the site. The intermittent and uncontrollable nature of PV systems poses a new challenge for

utility companies: Unlike conventional power plants, PV systems cannot be simply adjusted to meet expected short-term demand. Effective use of solar power thus requires accurate predictions of power generation so that operators can maximize their utilization of solar energy when available and transition to alternative sources of energy as needed.

The purpose of this thesis is to investigate the effectiveness of machine learning techniques to predict observed solar irradiance in the near future. Machine learning techniques have been employed in recent decades to solve a wide variety of challenging real-world regression problems. This thesis investigates the accuracy of several types of machine learning algorithms for future hours ranging from 1 to 24. Several previous studies, including Sanders et al. [28] and Bacher, Madsen, and Nielsen [2], have shown the importance of including local ground-based weather observations together with numerical weather predictions as model inputs, particularly for future horizons beyond an hour. Our experiments attempt to confirm and leverage these findings by incorporating weather data from the Georgia Automated Environmental Monitoring Network (GAEMN) [14] and the United States National Oceanic and Atmospheric Administration (NOAA) [30].

Chapter 2 begins by replicating and then expanding the results from a study published by Sanders et al. [28] that analyzed the usefulness of various potential inputs to machine learning models, including current weather observations, numerical weather predictions (NWPs) for the target site, and NWPs for a region around the target site. The work from [28] is extended by analyzing a wider range of target hours, investigating the viability of models that use *only* numerical weather predictions, and testing the ability of models to generalize to different geographic locations. The experiments in this section use historical weather data from several years ago and irradiance observations collected at

weather monitoring stations throughout the state of Georgia. Results indicate that weather predictions are the most important inputs for any future hour past the first, and accuracy can be improved by up to 8.26% by removing other inputs. It was also found that models trained for a specific geographical location perform relatively poorly when applied to new locations, but including a diverse set of geographical locations in the training set help models generalize more effectively.

Chapter 3 presents a case-study in irradiance prediction using a current dataset of irradiance observations collected from a solar farm in Athens, GA augmented by NOAA numerical weather predictions. The accuracy of a variety of machine learning models is investigated for fixed-axis, single-axis, and dual-axis tracking arrays for future hours 1-24. It was found that random forest tree ensembles exhibit the best performance, achieving a mean absolute error of 47.6 $\mathrm{W\,m^{-2}}$ for fixed-axis arrays, 58.6 $\mathrm{W\,m^{-2}}$ for single-axis tracking arrays, and 75.4 $\mathrm{W\,m^{-2}}$ for dual-axis tracking arrays.

The numerical weather predictions used in this chapter are arranged on a grid covering the south-eastern United States. A subset of this grid centered on the solar farm is used to train irradiance models, and the experiments in Chapter 3 also analyze the effect of broadening the geographical coverage of the grid. Results indicate that expanding the forecast coverage to an 84 km by 84 km area surrounding the target site may improve the predictive accuracy of random forest models by up to 4.1%.

1.2 Literature Review

Many researchers have attempted to develop models which predict solar irradiance in the near future. The SI unit for solar radiation is watts per meter squared ($\mathrm{W\,m^{-2}}$), and error values in the literature are commonly reported in these units. In the domain of solar radiation forecasting, there are several suitable metrics used to gauge the accuracy of solar irradiance point forecasts, including mean absolute error, relative absolute error, root mean squared error, correlation coefficient, and others [6, 7, 15]. None of these metrics stands out as most popular in the literature, which can make direct comparisons of results difficult. A further confounding factor is the lack of well-known benchmark datasets for the problem of irradiance forecasting. The bulk of the researchers use custom datasets which include an unpredictable assortment of weather observations and numerical weather predictions targeting a variety of specific geographic and temporal targets. As such, it is nearly impossible to draw straightforward comparisons between results from the literature and the results obtained in this thesis.

This thesis reports accuracy using the mean absolute error (MAE) metric, and all error rates can be assumed to denote the mean absolute error in $\mathrm{W\,m^{-2}}$ unless stated otherwise. MAE is implicitly resistant to the overweighting of outliers and has straightforward interpretation as the average distance between the value predicted by a model and the ground truth. Another attractive property is that MAE expresses error in the same units as a model's target variable. Furthermore, it was the metric-of-choice for Sanders et al., and its use simplifies the interpretation of the replication study in Section 2.3.1. In some cases, the coefficient of determination (R^2) is also reported to provide another point of comparison

among the models in this thesis and to allow comparison of results in this thesis to a larger portion of the literature.

The most trivial models are *persistence models*, which predict future solar radiation using current or historical observations. Sanders et al. [28] achieved a mean absolute error (MAE) of 67.28 $\mathrm{W\,m^{-2}}$ for 1-hour predictions at a single location in the state of Georgia by using the current value of solar radiation as the predicted value. They achieved an MAE of 60.70 $\mathrm{W\,m^{-2}}$ by using the mean historical solar radiation for the date and time of the prediction. For 24-hour predictions, the same types of persistence models achieved an MAE of 68.51 $\mathrm{W\,m^{-2}}$ and 62.53 $\mathrm{W\,m^{-2}}$, respectively.

Perez et al. [24] evaluated a persistence model on sites in geographically diverse locations including Desert Rock, Nevada, Fort Peck, Montana, Boulder, Colorado, Sioux Falls, South Dakota, Bondville, Illinois, Goodwin Creek, Mississippi, and Penn State, Pennsylvania. Their persistence model achieved an average root mean squared error (RMSE) of 95.29 $\mathrm{W\,m^{-2}}$ for 1-hour predictions and an average RMSE of 182.29 $\mathrm{W\,m^{-2}}$ for 24-hour predictions.

It is possible to create models more sophisticated than persistence models using only historical radiation data. Pedro and Coimbra [23] collected one-hour averaged radiation data from a one-megawatt plant in Central California from November 3, 2009 to August 15, 2011, then used a genetic algorithm to evolve the weights and architecture of an artificial neural network (ANN). Their network achieved an MAE of 42.96 $\mathrm{W\,m^{-2}}$ for 1-hour predictions and an MAE of 62.53 $\mathrm{W\,m^{-2}}$ for 24-hour predictions. The authors note that the purpose of their study was to assess the performance of univariate endogenous models of solar irradiance, and thus they opt not to use available data such as global horizontal irradiance, cloud cover, and wind speed and direction.

For relatively short-term temporal horizons, from the next few minutes up to 2- or 3-hour predictions, statistical approaches based on time series modeling are commonplace. David et al. [8] apply a combination of two linear models, auto-regressive moving average (ARMA) [33] and generalized autoregressive conditional heteroskedasticity (GARCH) [3], to predict solar irradiance for a single location at a future time ranging from 10 minutes to 6 hours. Their model achieved a normalized RMSE of 20.8% for a 10-minute horizon, 30.3% for the 1-hour horizon, and 32.1% for the 6-hour horizon. They noted that nonlinear machine learning models outperform the auto-regressive models past the 2-hour horizon.

Bacher, Madsen, and Nielsen [2] developed both univariate auto-regressive models of solar radiation (AR models) and enhanced auto-regressive models which included weather variables (ARX models). Their models predicted solar radiation for each future hour between 1 and 36 for a small site in Jutland, Denmark. They noted that the inclusion of weather variables is important for achieving accuracy for long-term predictions. The ARX model outperformed the AR model past the 3-hour horizon.

The use of historical irradiance observations is no doubt an important component of predictive models, but for horizons of more than a couple of hours, additional variables are needed to predict irradiance patterns. Along with historical data, it is common to use present- and forward-looking weather data as model inputs.

Several researchers have investigated the use of machine learning techniques to automatically construct predictive models. Sharma et al. [32] trained a support vector machine (SVM) to predict solar radiation at a single site in Massachusetts with a 3-hour future horizon. Their dataset consisted of eight months of solar intensity readings augmented by observations of temperature, dew point, wind speed, sky cover, precipitation, and humidity

from the National Weather Service (NWS). They achieved an RMSE of 128 $\mathrm{W\,m^{-2}}$ using a support vector regressor with an RBF kernel. The authors claimed that their SVR model achieved better accuracy than others found in the literature at that time. Their results indicate that a support vector machine with a nonlinear kernel outperforms simple linear regression.

Lauret et al. [19] compared the predictive accuracy of support vector machines, Gaussian processes, and artificial neural networks (ANNs) with naive persistance models and with linear autoregressive models. The models were trained and validated using irradiance measurements from three Mediterranean islands spanning the years of 2012 and 2013. They found that the nonlinear machine learning models slightly outperform autoregressive models for single-hour forecasts, and the dominance of the machine learning models becomes more pronounced as the forecasting horizon grows beyond one hour. They also noted that the improvement in prediction accuracy is more pronounced in the case of unstable cloud cover.

Feudo et al. [10] found similar correlations between cloud cover and accuracy of solar radiation predictions when making hourly predictions from July 2013 to December 2013 in South Italy. They were able to achieve a mean absolute error of 26.7 $\mathrm{W\,m^{-2}}$ in clear conditions versus 43.1 $\mathrm{W\,m^{-2}}$ in cloudy conditions. They noted that a variability in the optical depth of clouds, suggesting that cloud cover is not a uniform variable which can be expressed as a simple percentage. Rather, different types of cloud cover can affect surface readings differently.

Sfetsos and Coonick [31] evaluated several machine learning techniques for predicting hourly solar radiation, including linear regression, ANNs, recurrent neural networks, radial basis function (RBF) neural networks, and a neuro-fuzzy inference scheme. They achieved

best results of RMSE of 39.32 $\mathrm{W\,m^{-2}}$ using the Levenberg-Marquardt backpropogation algorithm [12] to train a feed-forward neural network. They began by using only lagged solar radiation observations from the French island of Corsica during the late spring and early summer of 1996, but they found greater success after including additional external variables such as temperature and wind direction.

Marquez and Coimbra [20] used a dataset of forecasted weather variables such as sky cover, temperature and chance of precipitation, released by the National Weather Service. The weather variables served as inputs to an artificial neural network which predicted solar irradiance for every future hour up to 6 days ahead-of-time at a site in Merced, California. They used a genetic algorithm to determine the optimal set of weather variables among the many included in a National Weather Service forecast. Their model achieved an average RMSE of 72.00, representing a 42% improvement over a naive persistence model.

Voyant et al. [34] provide a more complete survey of machine learning methods for solar radiation prediction. They compared a wide variety of machine learning and statistical models, including ANNs, ARIMA, SVM, random forests, regression, and ensembles. They also performed a survey of current literature to determine if there are underutilized methods that could be beneficial. They found that ANN and ARIMA had similar performance, but they recommended ANN due to its greater versatility. They also found that SVM, regression trees, and random forests tended to produce promising results and should be investigated further The works surveyed above highlight the importance of including current observations and forecasted future values of weather variables as inputs to solar radiation models, particularly for long-term predictions. Sanders et al. [28] attempted to quantify the contribution of present- and forward-looking approaches in training machine learning models, as opposed to

models which use historical solar radiation observations alone. They found that incorporating forecasted weather variables in a local geographic neighborhood along with a sliding window of recent solar radiation observations improved the accuracy of 1-hour predictions by 5.9% and 24-hour predictions by 37.6%. Sanders et al. also compared a variety of machine learning techniques including linear regression, neural networks, model trees, and tree ensembles. They found that tree ensemble methods, such as random forest, achieved the lowest mean absolute error.

2 Analysis of Feature Set Importance and Geographic Portability

2.1 Introduction

The literature reviewed in Chapter 1 indicates that machine learning techniques can be employed to develop solar irradiance models capable of outperforming naive baseline models as well as most auto-regressive models. Three categories of data are typically used as inputs to models that forecast solar irradiance: The first is accumulated historical data on observed solar irradiance and weather conditions. The second is measured values of weather variables collected at the time when the forecast is generated, hereafter referred to as "current weather observations". The last is comprised of numerical weather predictions (NWPs), which use mathematical models and weather simulations to forecast the values of weather variables such a temperature, cloud cover, and wind at a future time. Sanders et al. [28] attempted to quantify the importance of each of these categories of data. They first developed a variety of machine learning models which were trained using only current weather observations and historical irradiance data. They proceeded to compare these models against a set of models that were trained using both weather observations and weather forecasts. Their work found that the inclusion of weather forecasts for the target site slightly improved prediction accuracy for 1-hour predictions and greatly improved results for longer-term (i.e. 24-hour) predictions.

Including weather forecasts from a small region around the target site offered additional improvements to accuracy for both 1-hour and 24-hour models.

This chapter seeks to extend the work of Sanders et al. [28] to better understand the importance of including various types of data as model inputs. First, a replication study is described to verify the main findings of Sanders et al. The study confirms that the addition of weather forecasts improves model accuracy, but the study casts some doubt on the importance of including historical irradiance readings and current weather observations as model inputs. Second, the experimental results from [28] are extended temporally by analyzing all future prediction hours between 1 and 24. This analysis finds that the importance of NWP inputs increases with the forecast horizon, while historical irradiance readings and current weather observations were found to be unhelpful beyond a very short-term horizon. Finally, the effect of a simple univariate feature selection technique is explored to select an improved set of inputs among those found in current weather observations, historical sliding windows, and numerical weather forecasts. Results indicate that feature selection is an important pre-processing step to achieve optimal accuracy, even for models which implicitly perform feature selection during training.

In addition to analyzing the importance of various feature sets, this chapter investigates the geographical portability of machine learning models for solar irradiance. The irradiance models from the literature surveyed in Chapter 1 are typically trained and evaluated using data from a single geographical site, and the question remains whether machine learning models trained in this fashion are capable of generalizing to new locations. To answer this question, models are trained using weather observations and NWPs for a single weather station from the GAEMN network [14], then evaluated on weather stations throughout the

state of Georgia. It was found that a typical training strategy produces models which are highly specialized to their target site, but models can be made more general by training with data from multiple sites, representing a broader range of climates and weather patterns.

2.2 Methodology

As mentioned in Section 1.2, there are several suitable metrics used to gauge the accuracy of solar irradiance predictions. Results in this chapter report accuracy using the mean absolute error (MAE) metric, and all error rates can be assumed to denote the mean absolute error in $\mathrm{W\,m^{-2}}$ unless stated otherwise.

Sanders et al. [28] used the machine learning software Weka [13] to train and validate models. The experiments in this chapter instead use the open-source Python library scikit-learn [22]. Similar to Weka, scikit-learn includes tools to train linear, tree-based, and ANN models as well as tools to perform parameter tuning and model validation.

Except where specially noted, Sanders et al. used Weka's default set of hyperparameters for each model tested, and in this chapter, the experiments endeavor to match those hyperparameters as closely as possible. Although there is significant overlap in the types of models available in Weka and scikit-learn, the two frameworks are likely to have differences in implementation. Section 2.3.1 replicates the major findings in Sanders's work and includes a more complete discussion of discrepancies.

2.2.1 Data Acquisition

The datasets analyzed in this chapter are composed of one or more of the three following components. The first component is observations of weather variables taken at the *reference*

time (i.e. the time when the irradiance prediction is generated). The second component consists of numerical weather predictions, which forecast the values of weather variables at one or more future hours ahead of the reference time. The final component is parameters describing seasonal variables such as the date and the time of day.

Observations of weather variables are sourced from the Georgia Automated Environmental Monitoring Network (GAEMN) [14]. Every 15 minutes, GAEMN samples 43 meteorological and environmental variables, including air and soil temperature, barometric pressure, wind speed and direction, rainfall, and solar irradiance, from over 80 stations throughout the state of Georgia. GAEMN data was collected for the time period of 2003 to 2013 for five cities in Georgia: Griffin, Jonesboro, Attapulgus, Blairsville, and Brunswick. The machine learning models developed later in this chapter target the irradiance readings from the five sites. Table 1 displays summary statistics on the irradiance readings for each site.

Table 1: Summary statistics for the irradiance readings ($\mathrm{W\,m^{-2}}$) observed at five GAEMN weather stations.

Site	**Mean**	**Std. Dev.**	**Min**	**Max**
Griffin	190.40	277.09	0.00	1200
Jonesboro	170.45	259.76	0.00	1132
Attpulgus	184.96	267.31	0.00	1162
Blairsville	180.29	273.13	0.00	1260
Brunswick	198.06	280.86	3.10	1262

The city of Griffin lies in the central portion of the state and experiences the humid subtropical climate that characterizes most of the state. Jonesboro sits just 20 miles north of Griffin and was used as a validation city for the results obtained in Griffin. The remaining three cities were chosen to represent a broader range of environmental and meteorological

conditions possible throughout the state of Georgia. Figure 1 maps the location of each GAEMN site.

Figure 1: The five cities in Georgia chosen for solar radiation prediction

Forecasted values of weather variables are provided by the National Oceanic and Atmospheric Administration's (NOAA) Numerical Weather Prediction (NWP) models. Each of these models describe a particular geographical region by splitting the region into a grid of cells of consistent size. NOAA releases data in the GRIB file format, a compact binary format commonly used to store historical and forecasted weather data[1]. Each GRIB file lists weather attributes for a particular cell in the region at a particular time, or, in the case of weather forecasts, at a particular future time. Data from two different NOAA models are used in this chapter.

[1]http://www.nco.ncep.noaa.gov/pmb/docs/on388/

The first NWP model is the Rapid Refresh (RAP) model[2]. Every hour, RAP releases a forecast on a 13-km^2 resolution horizontal grid across North America which extends 18 hours into the future with a one-hour resolution. Only a subset of historical NWPs generated using the RAP model is available online at any time due to storage requirements; This chapter uses RAP predictions released between June 22, 2011 to April 30, 2012. Approximately 2% of the RAP forecast files for this time frame were missing, and these periods were not included in any datasets.

The second NWP model is the North American Mesoscale Forecast System (NAM)[3]. NAM is a deployment of the Weather Research and Forecasting Non-hydrostatic Mesoscale Model (WRF-NMM), which forecasts dozens of weather variables with a 12-km^2 spatial resolution covering the continental United States and much of Canada and Mexico. This chapter analyzes NAM forecasts released between June 2003 to May 2005. The forecasts collected during this timeframe were published every 4 hours, and each forecast extended 60 hours into the future with a resolution of 3 hours. Approximately 4% of the NAM forecast files for this time frame were missing, and these periods were not included in any datasets.

2.2.2 Preprocessing

The solar radiation observations collected from GAEMN have a much finer temporal resolution than the RAP or NAM weather forecasts. Sanders et al. [28] created synthetic weather forecasts by linearly interpolating between forecast hours such that the forecast reference time matches the timestamp of the GAEMN observations. For example, to create a synthetic

[2]https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/rapid-refresh-rap

[3]https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/north-american-mesoscale-forecast-system-nam

one-hour weather forecast at 1:15 pm for 2:15 pm, 75% of the one-hour forecast generated at 1:00 pm (for 2:00 pm) and 25% of the two-hour forecast generated at 1:00 pm (for 3:00 pm) were added together. This chapter uses the same method to match GAEMN observations with the NWP weather forecasts.

In the case of Rapid Refresh forecasts, performing such interpolations resulted in 28,896 instances of one-hour forecasts. These instances correspond to 15-minute intervals over 313 different days between June 22, 2011 and April 30, 2012. In the same way, 67,720 instances of 24-hour forecasts were created by interpolating the NAM forecasts over 720 days between June 2, 2003 and May 25, 2005. However, it should be noted that 15-minute interpolations between NAM forecasts, where there is a three-hour gap between forecasted values, are less accurate than those interpolating between RAP forecasts, which have one-hour resolution. Interpolations between three-hour forecasts may be a significant source of noise.

2.3 Results

2.3.1 Replication Study

The work in [28] demonstrated that error rates for both 1-hour and 24-hour predictions can be significantly improved for tree ensemble methods by including a 24-hour window of previous solar radiation observations as model inputs. Accuracy can be improved further by including the values of forecasted weather variables for the NOAA cell containing the target site and from a 3x3 grid of cells surrounding the site. Including both the window and the weather forecasts improves the models more than either alone. A preliminary goal of this chapter is to verify these findings.

Table 2: Comparison of machine learning prediction accuracy for 1-hour and 24-hour solar radiation for Griffin, GA using only historical GAEMN weather observations

Model	**1-hour MAE** ($W\,m^{-2}$)	**1-hour training time (s)**	**24-hour MAE** ($W\,m^{-2}$)	**24-hour training time (s)**
Persistance	53.64	N/A	63.016	N/A
Linear Regression	66.48	0.04	81.69	0.03
Multilayer Perceptron	47.17	12.91	74.03	48.26
M5P Model Tree	39.07	0.33	51.67	1.01
Random Tree	46.33	0.024	65.29	0.04
Bagging M5P	31.93	1.790	45.60	5.92
Bagging Random Tree	37.28	0.13	58.32	0.34
Random Forest	31.12	5.30	44.47	15.51

First, a variety of machine learning methods are compared by their accuracy in predicting 1-hour and 24-hour solar radiation, using *only* historical GAEMN observations as model inputs. The following weather variables are included: air temperature(°C), humidity (%), dew point (°C), vapor pressure (kPa), vapor pressure deficit (kPa), barometric pressure (kPa), wind speed ($m\,s^{-1}$), wind direction (°), maximum wind speed ($m\,s^{-1}$), open pan evaporation (mm), solar radiation ($W\,m^{-2}$), total solar radiation ($kJ\,m^{-2}$), photosynthetically active radiation ($\mu mol\,m^{-2}\,s^{-1}$), and two rainfall measurements (mm). Two additional inputs were included which represent the date and time. The day of the year is encoded as a discrete number between 0 and 365, and time-of-day is represented as continuous value between 0.00 and 0.99 describing the proportion of the day passed (e.g. noon is encoded as 0.50).

Sanders et al. trained a variety of machine learning models using the GAEMN dataset described above. The accuracy of each model was compared and it was found that random forest models outperform all others on this dataset. These results are confirmed by Table 2, which shows the MAE achieved by various learning algorithms in the replication

study. Random forest emerge as the strongest-performing models with other tree ensembles such as Bagging M5P close behind. Multilayer Perceptrons underperform expectations, particularly for 24-hour predictions where they fail to achieve an accuracy better than the persistence model baseline. This surprising result was noted in [28], although additional tuning to the network structure and other hyperparameters may result in better performance.

Random forests achieved an MAE of 31.12 $\mathrm{W\,m^{-2}}$ for predictions 1 hour in the future and an MAE of 44.47 $\mathrm{W\,m^{-2}}$ for predictions 24 hours in the future. This matches the accuracy achieved in the original paper quite closely: Sanders et al. reported an MAE of 29.96 for 1-hour random forest and an MAE of 44.11 for 24-hour random forest. The slight differences are possibly attributable to implementation details that differ between Weka and scikit-learn as well as the stochastic nature of the random forest training algorithm.

Next, we seek to analyze the effect of including sliding windows of historical irradiance and forecasted weather variables as model inputs. In addition to the GAEMN variables, 24 inputs were added to the models, corresponding to the recorded solar radiation for each of the past 24-hours. This dataset is referred to as *GAEMN+window*. To further improve accuracy, forecasted weather variables were added to the model's inputs. In the case of 1-hour predictions, we include eight variables from the RAP weather forecasts: air temperature, precipitation rate, visibility, wind speed, wind direction, dew point temperature, air pressure, and relative humidity. In the case of 24-hour predictions, we include the same eight variables from the NAM weather forecasts. This dataset is referred to as *GAEMN+window+forecast*.

Finally, the eight forecasted weather variables were included from each of the nine cells in a 3x3 grid centered on the target site. This is intended to take into account the possibility of nearby weather conditions moving into the target cell during the prediction window.

The addition of neighboring weather variables may also help reduce the error introduced during the 15-minute interpolations. This dataset is referred to as *GAEMN+window+grid*.

Table 3: Comparison of predictive accuracy for different sets of inputs to random forest models for five Georgia cities

		GAEMN	**GAEMN + window**	**GAEMN + window + forecast**	**GAEMN + window + grid**
Griffin	*1-hour*	31.29	31.04	30.64	29.97
	24-hour	43.84	46.52	35.51	33.22
Jonesboro	*1-hour*	28.84	27.90	26.93	26.74
	24-hour	46.46	49.31	36.74	34.07
Attapulgus	*1-hour*	32.05	31.31	30.21	30.24
	24-hour	49.34	51.60	40.25	37.51
Brunswick	*1-hour*	35.60	35.07	33.98	31.47
	24-hour	45.66	47.93	36.11	33.34
Blairsville	*1-hour*	32.72	32.82	31.76	31.65
	24-hour	48.18	51.95	36.98	33.34
Overall	*1-hour*	32.10	31.63	30.70	30.01
	24-hour	46.70	49.46	37.12	34.30

Table 3 displays the MAE achieved by 1-hour and 24-hour random forest models trained with each set of features. For the most part, these results match the findings presented in Sanders et al. [28]. The addition of forecasted weather variables from the cell containing the target site improved 1-hour prediction accuracy by 2.94%, dropping MAE from 31.63 $\mathrm{W\,m^{-2}}$ to 30.70 $\mathrm{W\,m^{-2}}$. The addition of forecasted weather variables in a 3x3 grid surrounding the target site improved 1-hour prediction accuracy by an additional 2.20%, for a total improvement of 5.12% averaged across the five target sites. This value is slightly lower than the improvement reported by Sanders et al., who observed a 7.60% total improvement.

The effect is much more pronounced for 24-hour predictions. The addition of a single cell of weather variables dropped MAE from 49.46 $\mathrm{W\,m^{-2}}$ to 37.12 $\mathrm{W\,m^{-2}}$, an improvement

of 25.95%. The addition of the 3x3 grid of forecasted variables dropped MAE by an additional 4.70% for a total improvement of 30.65% averaged across target sites.

While the addition of forecasted weather variables improved accuracy for both target hours, the contributions of the sliding window were less consistent. The accuracy of 1-hour models was marginally improved by the addition of the sliding window of recent irradiance, lowering MAE by 1.5% from 32.10 $\mathrm{W\,m^{-2}}$ to 31.63 $\mathrm{W\,m^{-2}}$. But the quality of 24-hour predictions was negatively impacted by the addition of the sliding window, increasing overall MAE from 46.70 $\mathrm{W\,m^{-2}}$ to 49.46 $\mathrm{W\,m^{-2}}$. This observation was not explicitly discussed by Sanders et al. [28], but the trend is consistent with the results they reported. Overall, the results from the original study hold: the accuracy of a predictive model can be improved by the addition of forecasted weather variables for the target site, and the magnitude of the improvement is greater for long-term (24-hour) predictions. Inclusion of forecasted variables in a small region around the target site offers another slight improvement.

2.3.2 Temporal Extension

Sanders et al. [28] restrict their analyses to 1-hour and 24-hour predictions alone. For most applications, however, predictions of solar radiation are needed for every future hour up to some threshold. This section extends the results from [28] by analyzing all forecast hours between 1 and 24.

Random forest models were trained for each target hour using the *GAEMN*, *GAEMN+window*, *GAEMN+window+forecast*, and *GAEMN+window+grid* datasets. RAP forecasts extend 18 hours into the future, so the RAP forecasts were used to train models

targeting target hours 1 through 18. NAM forecasts, each of which extend 60 hours into the future, were used to train models targeting hours 19 through 24.

Table 4: MAE achieved by random forest models targeting each hour between 1 and 24 trained using various combinations of inputs. Results averaged across five sites.

Hour	**GAEMN (Wm-2)**	**GAEMN +window (Wm-2)**	**GAEMN + window + forecast (Wm-2)**	**GAEMN + window + grid (Wm-2)**	**single-cell forecast impact (%)**	**multi-cell forecast impact (%)**
1	32.10	31.63	30.70	30.41	2.91	3.83
2	34.48	35.42	33.62	32.61	2.48	5.40
3	34.09	37.52	34.40	33.41	-0.89	2.02
4	34.09	38.48	34.26	33.50	-0.51	1.71
5	33.13	38.00	33.75	32.81	-1.86	0.96
6	33.03	38.77	34.08	33.15	-3.17	-0.37
7	32.90	38.58	33.39	32.80	-1.49	0.30
8	33.75	39.22	33.93	33.04	-0.52	2.13
9	34.42	39.02	33.77	32.59	1.88	5.32
10	34.64	38.89	33.37	32.09	3.66	7.38
11	34.68	39.09	33.50	32.38	3.39	6.63
12	35.22	39.57	33.80	33.01	4.05	6.29
13	35.77	39.51	33.82	32.70	5.44	8.58
14	36.14	39.82	33.76	32.98	6.59	8.75
15	36.21	39.43	33.26	32.72	8.14	9.64
16	37.64	40.52	34.21	33.28	9.11	11.60
17	38.76	40.65	34.40	33.42	11.25	13.78
18	39.97	41.25	34.81	33.88	12.91	15.24
19	42.41	50.07	37.40	34.57	11.81	18.50
20	42.30	49.84	37.43	33.84	11.51	20.02
21	42.02	49.10	36.58	33.62	12.94	19.99
22	42.94	49.23	36.57	33.33	14.84	22.37
23	44.76	49.36	36.76	33.71	17.86	24.68
24	46.70	49.46	37.11	34.30	20.52	26.54

Table 4 gives the MAE achieved by models trained for each target hour, organized by the dataset used for training, and Figure 2 shows the same results graphically. When using *GAEMN* and *GAEMN+window* as model inputs, the error consistently increases with the forecast hour. This result is intuitive. As the forecast hour increases, the predictive power of

current weather observations is reduced. The addition of forecasted weather variables vastly improves accuracy, and this effect is more pronounced as the forecast horizon increases. The inclusion of a 3x3 grid of forecasts improved accuracy by an average of 2.26% for future hours 1-6, 4.67% for future hours 7-12, 11.27% for future hours 13-18. Future hours 19-24, which used NAM forecasts rather than RAP forecasts, experienced the greatest benefit, improving by 22.02% on average.

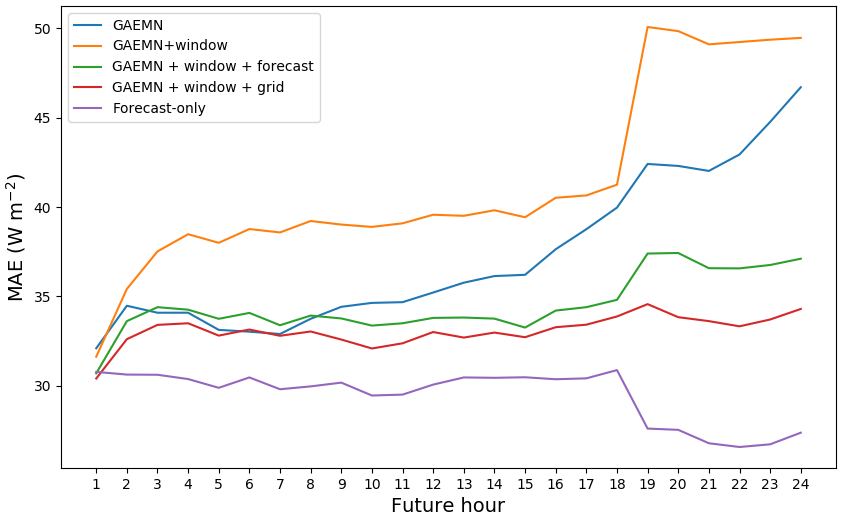

Figure 2: MAE achieved by random forest models targeting each hour between 1 and 24 trained using various combinations of inputs. Results averaged across five sites.

The addition of weather forecasts causes the accuracy achieved for each target hour to become nearly uniform. Models trained with the *GAEMN* data alone exhibited errors ranging from 32.10 W m^{-2} to 46.70 W m^{-2}, whereas models trained with weather forecasts had errors ranging only from 30.41 W m^{-2} to 34.30 W m^{-2}.

This experiment also confirms the suspicion that arose in Section 2.3.1. Sliding windows of historical irradiance were found to be unimportant for longer forecast horizons. For every forecast hour past the first, addition of a sliding window lowered accuracy. As the horizon

increases, the information encoded in the sliding window becomes poorly correlated with irradiance patterns at the target hour. For most tree-based models, which are characterized by implicit feature selection, this would be unlikely to affect results. Random forests, however, build an ensemble of decision trees by selecting a subset of attributes at random to use for each tree. For the *GAEMN+window* dataset, the sliding window comprises 24 of the 39 input variables. When this dataset is used, many of the trees in the random forest ensemble will be trained using features predominantly selected from the sliding window, and the resultant trees will perform relatively poorly for future horizons beyond the first few hours.

2.3.3 Forecast-only Inputs

The work in [28] along with the replication study in Section 2.3.1 show that the addition of forecasted weather variables significantly improves model accuracy, particularly for long-term predictions. Sanders et al. briefly mention that for short-term predictions, the sliding window appears to contribute significantly to prediction accuracy, but for long-term predictions, the sliding window has very little effect.

The temporal extension in the previous section indicates similar trends. The addition of forecasted weather variables had a less dramatic impact on short-term prediction accuracy than for long-term prediction accuracy. The sliding window appears to *harm* accuracy for long-term predictions. Furthermore, the addition of weather forecasts was seen to mostly dissolve the difference in accuracy among target hours. Some questions thus remain regarding the contributions of each set of inputs. Specifically, we hypothesize that forecasted variables from numerical weather predictions may serve as a sufficient set of model inputs to achieve

accuracy comparable to the best models developed so far. That is, machine learning models may not benefit greatly by the addition of weather observations at the reference time.

This hypothesis was tested by training and evaluating random forest models for each target hour using a *forecast-only* dataset, which includes the forecasted weather variables from the 3x3 grid centered on the target site. The forecast-only models are compared against the models trained using the *GAEMN+window+grid* dataset, which achieved the best results in Section 2.3.2. Table 5 shows the MAE, averaged across the five sites, achieved by the *GAEMN+window+forecast* models and the *forecast-only* models.

The accuracy of 1-hour models trained on the *forecast-only* dataset dropped by 1.20% on average, indicating that weather observations are important for very short-term models. It may be the case that 0-hour GAEMN weather observations are a better predictor of weather in the very-near future than the numerical models published by NOAA. However, for any horizon past the first future hour, the *forecast-only* models outperformed the *GAEMN+window+grid* models by a significant margin. The accuracy of *forecast-only* models targeting hours 2-18, which are covered by the Rapid Refresh forecasts, improved by 8.26% on average. Note that models targeting hours 19-24, which use NAM weather forecasts, experienced a more pronounced improvement with MAE dropping by an average of 20.04%. The differences in the magnitude of the improvement may be related to qualitative distinctions between the NAM and RAP weather models, the interplay between the GAEMN inputs and the NAM/RAP inputs, the time periods from which samples were drawn, or some combination of these factors.

Random forests, like most other tree-based models, perform implicit feature selection. Nonetheless, results in this section indicate that, for most future horizons, the accuracy of

Table 5: Comparison of MAE achieved by random forest models trained with *GAEMN+window+grid* and *forecast-only* datasets.

Dataset	**Hour**	**GAEMN +window +grid** ($\mathrm{W\,m^{-2}}$)	**Forecast -only** ($\mathrm{W\,m^{-2}}$)	**% improvement (from col. 2 to col. 3)**
2011-2012 RAP NWPs	1	30.41	30.78	*-1.20*
	2	32.61	30.63	*6.09*
	3	33.41	30.62	*8.34*
	4	33.50	30.38	*9.33*
	5	32.81	29.89	*8.89*
	6	33.15	30.47	*8.10*
	7	32.80	29.81	*9.10*
	8	33.04	29.97	*9.28*
	9	32.59	30.18	*7.40*
	10	32.09	29.46	*8.20*
	11	32.38	29.51	*8.86*
	12	33.01	30.07	*8.89*
	13	32.70	30.47	*6.82*
	14	32.98	30.45	*7.69*
	15	32.72	30.48	*6.85*
	16	33.28	30.37	*8.74*
	17	33.42	30.42	*8.97*
	18	33.88	30.88	*8.85*
2003-2005 NAM NWPs	19	34.57	27.61	*20.13*
	20	33.84	27.54	*18.60*
	21	33.62	26.79	*20.31*
	22	33.33	26.58	*20.27*
	23	33.71	26.73	*20.72*
	24	34.30	27.38	*20.19*

random forest models can be improved by dropping certain sets of features such as the sliding window and the GAEMN weather observations. Performing explicit feature selection as a pre-processing step may thus yield results of similar or superior quality than hand-selecting a subset of features. This was tested by re-training random forest models on a modified version of the *GAEMN+window+grid* dataset. The number of features in the dataset was reduced by 50%, keeping the best-scoring features as measured by the *mutual information* criterion [18].

Table 6: Comparison of MAE achieved by random forest models trained with and without feature selection.

Dataset	**Hour**	**GAEMN + window + grid**	**Feature Selection**	**Forecast -only**	**% change (col. 3 to col. 4)**
2011-2012 RAP NWPs	1	30.41	30.54	30.78	*-0.77%*
	2	32.61	30.42	30.63	*-0.68%*
	3	33.41	30.74	30.62	*0.40%*
	4	33.50	29.92	30.38	*-1.55%*
	5	32.81	29.83	29.89	*-0.19%*
	6	33.15	29.51	30.47	*-3.27%*
	7	32.80	30.46	29.81	*2.14%*
	8	33.04	29.68	29.97	*-0.96%*
	9	32.59	30.04	30.18	*-0.46%*
	10	32.09	29.70	29.46	*0.81%*
	11	32.38	30.52	29.51	*3.32%*
	12	33.01	30.39	30.07	*1.05%*
	13	32.70	30.00	30.47	*-1.57%*
	14	32.98	30.16	30.45	*-0.97%*
	15	32.72	29.39	30.48	*-3.71%*
	16	33.28	29.87	30.37	*-1.67%*
	17	33.42	29.61	30.42	*-2.73%*
	18	33.88	29.47	30.88	*-4.77%*
2003-2005 NAM NWPs	19	34.57	29.63	27.61	*6.83%*
	20	33.84	30.08	27.54	*8.45%*
	21	33.62	30.26	26.79	*11.46%*
	22	33.33	30.85	26.58	*13.84%*
	23	33.71	32.55	26.73	*17.89%*
	24	34.30	34.06	27.38	*19.60%*

Table 6 compares the MAE achieved by random forests trained with the full *GAEMN+window+grid* dataset, the *GAEMN+window+grid* dataset reduced using feature selection, and the *forecast-only* dataset. Figures 3 and 4 show the same results graphically. Performing feature selection noticeably improved results for all target hours. Models trained on the *GAEMN+window+grid* dataset achieved an average MAE of 33.09 $\mathrm{W\,m^{-2}}$ across all future hours. Models trained on the same dataset but which use the described feature selection technique achieved an average MAE of 30.32 $\mathrm{W\,m^{-2}}$, representing an 8.5% improvement over

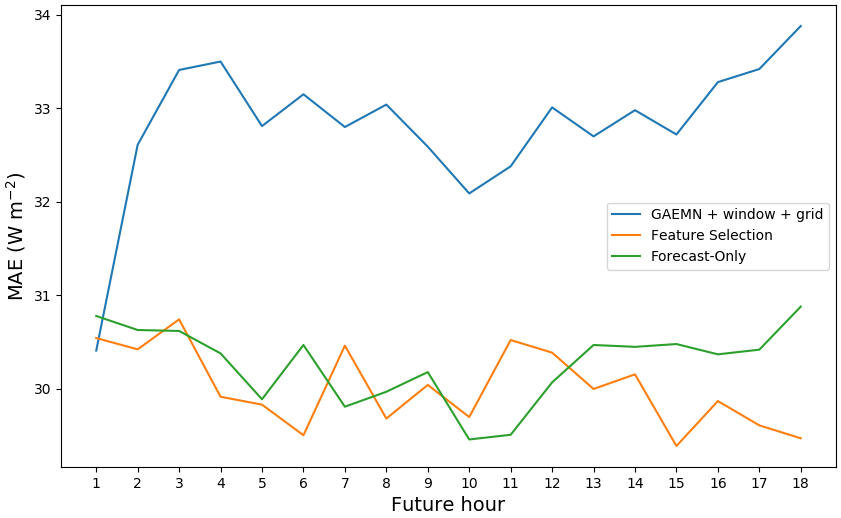

Figure 3: MAE achieved by random forest models trained with and without feature selection for future hours 1-18 using RAP NWPs.

models trained using the full *GAEMN+window+grid* dataset. For future hours 1-18, which use RAP NWPs, models trained with the feature-selection pre-processing step exhibit an average MAE of 30.01 W m^{-2}. This error rate is nearly identical to those trained with the *forecast-only* dataset, which exhibited an average MAE of 30.27 W m^{-2} on the same target hours. However, for future hours 19-24 (which use NAM NWPs), models trained on the *forecast-only* dataset still exhibit superior performance, achieving an average MAE of 27.11 W m^{-2} compared to 31.24 W m^{-2} achieved with the feature selection step.

These results indicate that explicit feature selection is an important pre-processing step, even for learning algorithms which implicitly select features. However, the superior performance of the *forecast-only* dataset on later target hours indicates that the simple univariate feature selection that we performed is insufficient to extract an optimal set of features. More sophisticated feature selection techniques, such as cross-validated recursive

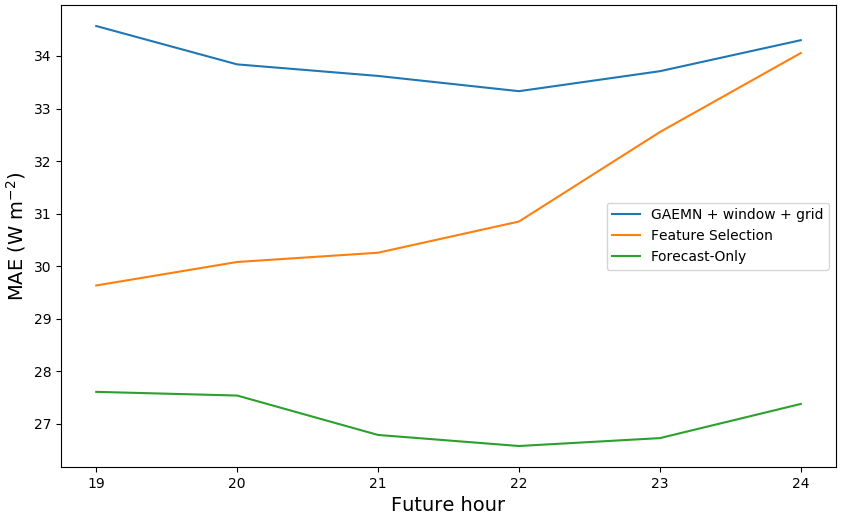

Figure 4: MAE achieved by random forest models trained with and without feature selection for future hours 19-24 using NAM NWPs.

feature elimination [11], may allow models to achieve even better accuracy than what has been observed in this section.

2.3.4 Geographic Portability

An important feature of predictive models is their ability to generalize to geographic and meteorological conditions dissimilar to those encountered during training. Although all target sites in the GAEMN network are located in a single state, they nevertheless represent a range of subclimates. The northern portion of the state is characterized by mountainous terrain and considerable elevation changes. Northern counties commonly receive several inches of precipitation per year, and average temperatures north of the Chattahoochee River are considerably lower than the remainder of the state. In contrast, the southeastern portion of the state borders the Atlantic Ocean and experiences hot, humid weather nearly year-round.

The topology of this area is marked by low-lying plains, and the region typically experiences shorter winters and long, warm summers.

The environmental variations in the state of Georgia can be leveraged to test the portability of predictive models. Models were trained and evaluated in five folds. The first fold trains a predictive model using the data from the city of Griffin, then evaluates the model using the other four cities. The second fold trains on Jonesboro then evaluates on the remaining four cities, and so on. In each case, the *GAEMN+window+grid* dataset was used for training, as it yielded the best results in section 2.3.1. Results are reported in Table 7. The city listed in the left-most column was the city used for training. The remaining four cities were used for testing.

Table 7: Error rates for random forest models trained for a specific site then evaluated on the remaining four sites

	1-hour	**24-hour**
Griffin	76.09	63.89
Jonesboro	50.58	60.43
Attapulgus	51.21	62.28
Brunswick	47.16	61.30
Blairsville	50.90	61.71
MEAN	55.19	61.92

These results indicate that models generalize poorly to locations other than their training site. The mean 1-hour MAE for random forest models over all locations was found to be 55.19 $\mathrm{W\,m^{-2}}$, representing an error 82% higher than the mean error of 30.32$\mathrm{W\,m^{-2}}$ when models were evaluated against their training site (see Table 3). The mean 24-hour MAE for random forest was 61.92 $\mathrm{W\,m^{-2}}$, 80.7% worse than the mean error of 34.26 $\mathrm{W\,m^{-2}}$ when models were trained and evaluated using a single site. Generally speaking, when a

model is trained for a specific site, then applied to a different site it does no better than a naive persistence model (see Table 2).

It may be possible to encourage models to generalize to other subclimates by including a diverse set of sites in the training data. Another experiment was run to test this strategy. Similar to the previous experiment, models were trained and evaluated in five folds, each fold corresponding to one of the five GAEMN sites. This time, however, a single location was held out for testing and the model was trained on data from the other four sites. This training strategy yields a much larger training set for each fold than in the previous experiment, and care must be taken to ensure that any changes in model performance are the result of the modified training strategy and not merely an artefact of the larger quantity of data used to train each model. The training sets for each fold were resampled to have the same number of examples as the training sets in the first experiment. The sample was stratified by target site so that an equal number of instances were taken from each subclimate.

Table 8: Error rates for random forest models evaluated on a specific site after being trained on the other four sites

	1-hour	**24-hour**
Griffin	34.92	43.94
Jonesboro	38.09	43.57
Attapulgus	36.01	52.87
Brunswick	40.36	55.48
Blairsville	42.76	53.78
MEAN	38.43	49.93

The results of this experiment are presented in Table 8. The left-most column corresponds to the site that was held out for testing. These results suggest that training on a variety of sites does indeed help the model generalize to unfamiliar climates. The mean MAE for 1-hour predictions across all sites is 38.43, a 30% improvement to the 1-hour results

observed when models were trained on a single site. The mean MAE for 24-hour predictions is 49.93, a 19.4% improvement over the single-site models.

This finding has practical ramifications. If a model is trained using data from a single location, then it is unlikely to generalize well to a new location where weather patterns are significantly different. The best strategy to maximize the accuracy of irradiance forecasts is to train a separate model for each location, or at least for each subclimate. However, training models can be time-consuming and expensive. In the case where resources are limited, it may be best to train a single model using data from a variety of locations. This strategy will encourage the model to learn more general weather patterns and allow it to generalize better to unfamiliar conditions in the future.

2.4 Conclusion

The replication study at the beginning of this chapter confirms the importance of incorporating current weather information along with weather forecasts for irradiance prediction over both short-term and long-term timeframes. Unlike the results obtained by Sanders et al. [28], little benefit was found when incorporating a sliding window of recent irradiance observations as model inputs. Moreover, when Sanders' experiments were extended to cover all future hours between 1 and 24, it appeared that beyond a very short future horizon, the forecasted weather variables become more important than the other model inputs. This was confirmed by building a set of models using only forecasted weather variables as inputs. Dropping all inputs except the forecasted variables improved predictive accuracy for all models targeting future hours past the 1-hour horizon.

This chapter also analyzed the geographic portability of models trained for a specific location. Under ideal circumstances, a model would learn to interpret a variety of weather conditions and generalize well to other sites. However, it was found that models trained for a specific site generalize poorly to other locations with dissimilar weather conditions. This effect can be partially mitigated by training models using data from a diverse set of locations. In practical scenarios where training time or computational resources are limited, it may be desirable to train a generalizable model using data drawn from a variety of sites. However, in scenarios where raw accuracy is a high priority, the best strategy is to train a single, highly-focused model for each deployment site.

There are several directions in which the work from this chapter could be extended. First, it would be helpful to analyze a broader range of weather stations than are found in the GAEMN dataset. There are many weather conditions and weather patterns which impact solar irradiance but are not regularly found within the state of Georgia. It may be possible to build a single model to predict irradiance for any climate given a sufficiently diverse set of training data.

Second, this chapter has only considered weather observations and predictions from a few meteorological stations and two different numerical weather models. The analysis is limited in that there are many data sources which may be helpful in predicting irradiance, such as satellite imagery, Doppler radar animations, and weather forecasts from a wider range of sources. More sophisticated machine learning techniques may be able to leverage information from these sources and improve predictive accuracy even further.

Finally, this chapter does not analyze the qualitative differences between the RAP forecasts from 2011-2012 which were used to train models for future hours 1-18 and the NAM

forecasts from 2003-2005 that were used to train models for future hours 19-24. Significant differences were observed in irradiance models trained using these datasets, and a better understanding of these differences may inform an improved data acquisition and model development process.

3 Predictive Models for Stationary & Non-Stationary Solar Arrays

3.1 Introduction

The intermittent and uncontrollable nature of PV systems poses a new challenge for utility companies: Unlike conventional power plants, PV systems cannot be simply adjusted to meet the expected short-term demand. Effective use of solar power thus requires accurate predictions of power generation so that operators can maximize their utilization of solar energy when available and transition to alternative sources of energy as needed.

The analyses performed in this thesis have so far considered solar irradiance as measured by various weather stations in the state of Georgia. While these models of solar irradiance may be useful, the data collected by these stations often differ qualitatively from irradiance observed at an actual solar farm. Many solar farms make use of "tracking" arrays, which automatically rotate around one or two axes to maximize exposure to the sun, and the irradiance collected by these tracking arrays will differ significantly from the irradiance measured by a fixed sensor.

This chapter presents a case study in developing predictive models of solar irradiance using observations collected from an actual solar farm in Athens, GA. Unlike the bulk of the literature concerning solar irradiance prediction, the experiments in this chapter investigate the accuracy of various machine learning techniques for predicting irradiance as measured by

tracking arrays, which rotate along one or two axes in order to maximize exposure to the sun, in addition to stationary arrays. Model development is informed by the results obtained in Chapter 2. Models are trained and evaluated using datasets composed entirely of forecasted weather variables, leaving out observations from weather stations and sliding windows of past irradiance readings.

To capture the geographical and meteorological information required to make accurate predictions, we use historical weather forecasts released by the National Weather Service to train several models using a variety of machine learning algorithms. The models are compared by their ability to predict solar irradiance in a 24-hour future window measured by fixed-axis solar arrays, single-axis tracking arrays, and dual-axis tracking arrays. Random forest models achieved the best accuracy with an MAE of 47.6 $\mathrm{W\,m^{-2}}$ for fixed arrays, 58.7 $\mathrm{W\,m^{-2}}$ for single-axis arrays, and 75.4 $\mathrm{W\,m^{-2}}$ for dual-axis arrays.

This chapter also analyzes the importance of the geographical coverage of the numerical weather predictions which serve as inputs to the models. Models are built which consider only the NWPs for the NAM cell containing the solar farm, then sequentially expanded to consider wider regions. It was found that including weather forecasts from a wider area, such as a 36 km by 36 km grid around the target site, resulted in a 1.9% decrease in MAE for random forest models and a 16.0% decrease in MAE for k-Nearest Neighbors. Returns diminish, however, as the grid grows larger.

3.2 Data Collection & Preprocessing

3.2.1 Weather Forecasts

As discussed in Chapter 3, the North American Mesoscale Forecast System (NAM) is one of the major numerical weather models administered by the National Centers for Environmental Prediction (NCEP), a division of the National Oceanic and Atmospheric Administration (NOAA) of the United States government. NAM is a deployment of the Weather Research and Forecasting Non-hydrostatic Mesoscale Model (WRF-NMM) which forecasts dozens of geospatial weather variables, including cloud cover, humidity, air and soil temperature, and precipitation. These forecasts also include a measure of downwelling shortwave radiation flux, a measure of irradiance, that is known to be biased towards overestimates.

We collected these forecasts for the years of 2017 and 2018, though a few forecasts are sporadically missing. Figure 5 shows the temporal coverage of the NAM forecasts in our dataset. Lighter colors indicate missing or partially missing data. The NAM model was updated in June 2006, and thus the NAM numerical weather predictions used in this chapter differ from those analyzed in Chapter 2. The updated model releases forecasts every six hours, and each forecast includes predictions for the following 84 future hours. The predicted weather variables have a 1-hour temporal resolution for the first 36 forecast hours and a 3-hour resolution for the remaining forecast hours. Many weather variables include volumetric data at various vertical resolutions (known as *isobaric layers*).

The NAM forecasts are released by NOAA in GRIB [4] format. As part of the collection process, we converted these into the netCDF [5] format, which provides a much

[4]http://www.nco.ncep.noaa.gov/pmb/docs/on388/
[5]https://www.unidata.ucar.edu/software/netcdf/

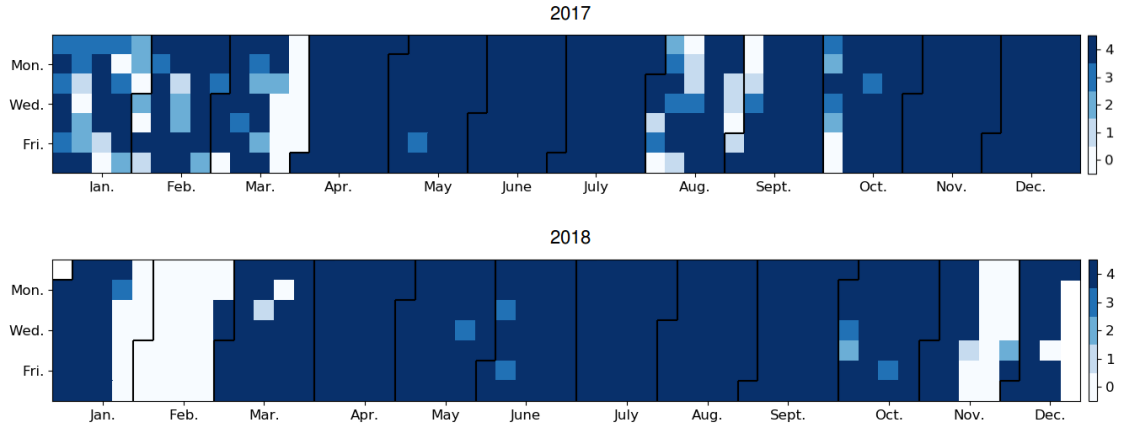

Figure 5: Heatmap of NAM forecasts collected between 2017 and 2018.

more convenient API for interacting with the data as a single object. The resulting database occupies over 2 TB of disk space.

The NAM forecasts are quite large and exhibit high dimensionality. To alleviate storage requirements, several variables which are unlikely to be directly related to solar irradiance, such as soil temperature, were discarded. The resulting dataset contains the following weather variables: downwelling shortwave and longwave flux ($\mathrm{W\,m^{-2}}$), geopotential height (gpm), air pressure (Pa), air temperature(K), water vapor content ($\mathrm{kg\,m^{-2}}$), equivalent reflectivity factor (dBZ), relative humidity (%), upward latent heat flux ($\mathrm{W\,m^{-2}}$), upward sensible heat flux ($\mathrm{W\,m^{-2}}$), cloud cover (%), atmospheric kinetic energy ($\mathrm{J\,kg^{-1}}$), visibility (m), and wind speed and direction ($\mathrm{m\,s^{-1}}$).

Since 2006, the NAM forecasts have a 12 $\mathrm{km^2}$ spatial resolution covering the contiguous United States and much of Mexico. The geographic coverage of each forecast was cropped to a 1,920 $\mathrm{km^2}$ area centered around Macon, Georgia and forecast hours beyond 36 were discarded. The resulting data has a consistent 1-hour forecast resolution and encompasses the southeast United States. The longitudinal range extends from the western border of

Louisiana to beyond the east coast of North Carolina. The latitudinal range extends from Key West, Florida to mid-Ohio. Figure 6 shows the geographic coverage of the NAM forecasts in our dataset.

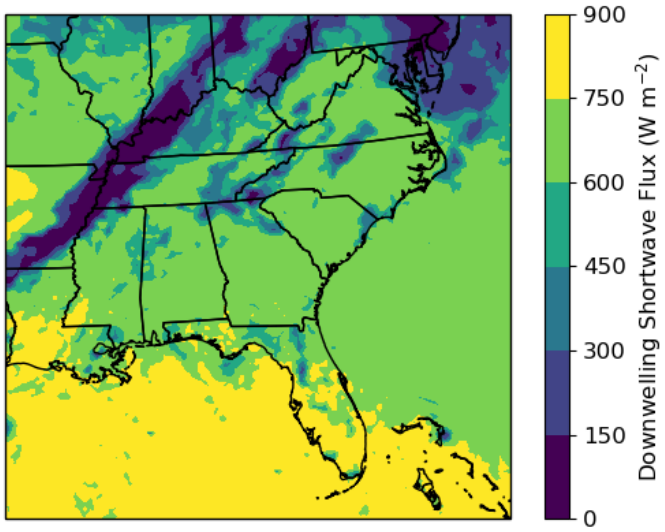

Figure 6: The geographic region encompassed by our dataset.

For all experiments, we standardize weather variables to have zero mean and unit variance. NAM forecasts include a "reference time" attribute, which is a textual representation of the date and time when the forecast was released. We convert this into a numerical representation of the day and time by taking the sine and cosine of the reference time with periods of one-day and one-year. Together, these four features parameterize the time of day and time of year of the forecasts.

3.2.2 Irradiance Observations

The models analyzed in this chapter target irradiance observations from a 1 MW solar farm in Athens, Georgia. This farm is operated by the Georgia Power company, and it includes

both fixed arrays, which are unable to change their position, as well as tracking arrays, which are designed to automatically rotate around one or two axes to maximize exposure to sunlight. The quantity of solar irradiance that strikes a planar surface can be measured by a *pyranometer*, and a variety of pyranometers from multiple manufacturers are installed on the solar arrays. Each pyranometer samples irradiance every five seconds and automatically uploads the reading to an online data store. We collected these readings for the years of 2017 and 2018 (though some readings are sporadically missing).

The irradiance observations have a much higher temporal resolution than the NAM forecasts. We aggregate the observations such that they match the temporal resolution of the NAM forecasts by truncating the timestamps of each observation to the previous hour and taking the mean of the resulting one-hour windows. The ground-truth observations can then be appended to the NAM attributes by matching the observation time to the forecast reference time.

This chapter considers three solar arrays, labelled A, B, and E. Array A is a dual-axis tracking array , Array B is a fixed array facing 20 degrees west of south, and Array E is a single-axis tracking array. Each array is outfitted with three *thermopile pyranometers*. The first pyranometer is model SMP-11 manufactured by Kipp & Zonan[6], and the other two pyranometers are model LI-200R manufactured by LICOR[7].

The readings obtained from a thermopile pyranometer are subject to uncertainty arising from calibration error, change in instrument sensitivity (non-stability), temperature response, and directional response. Reda [25] describes a practical method to estimate

[6] `https://www.kippzonen.com/`
[7] `https://www.licor.com/`

the uncertainty in the readings obtained from the pyranometers: To estimate uncertainty, compute the root sum square of the major uncertainty sources listed above. Table 9 compares the uncertainty in SMP-11 and LI-200R pyranometers. In both cases, uncertainty in calibration, non-stability, temperature response, and directional response are taken from the product manual published by the pyranometer manufacturer, assuming a maximum observed temperature of 50 degrees Celsius.

Table 9: Estimated uncertainty of pyranometers

Model	**Calibration**	**Non-Stability**	**Temp. Response**	**Directional Response**	**Uncertainty**
K&Z SMP-11	1.5%	0.5%	1.0%	0.2%	1.88%
LICOR LI-200R	3.0%	Not Reported	7.5%	1.0%	4.21%

The estimated uncertainty of the SMP-11 pyranometer is significantly lower than the LI-200R. Unless otherwise stated, the models developed in this chapter target solar irradiance as measured by the SMP-11 sensor.

3.2.3 Data Visualization

Figure 7 shows the mean irradiance readings from arrays A, B, and E between January 2017 to December 2018, grouped by hour of the day. For all arrays, irradiance peaks during the early afternoon, when the sun is at its highest, then drops to near-zero during nighttime hours.

Somewhat surprisingly for this installation, the single-axis tracking array (Array E) receives less irradiance at its peak when compared to the fixed array (Array B). However, the single-axis tracking array reaches its peak irradiance earlier than the fixed array and sustains

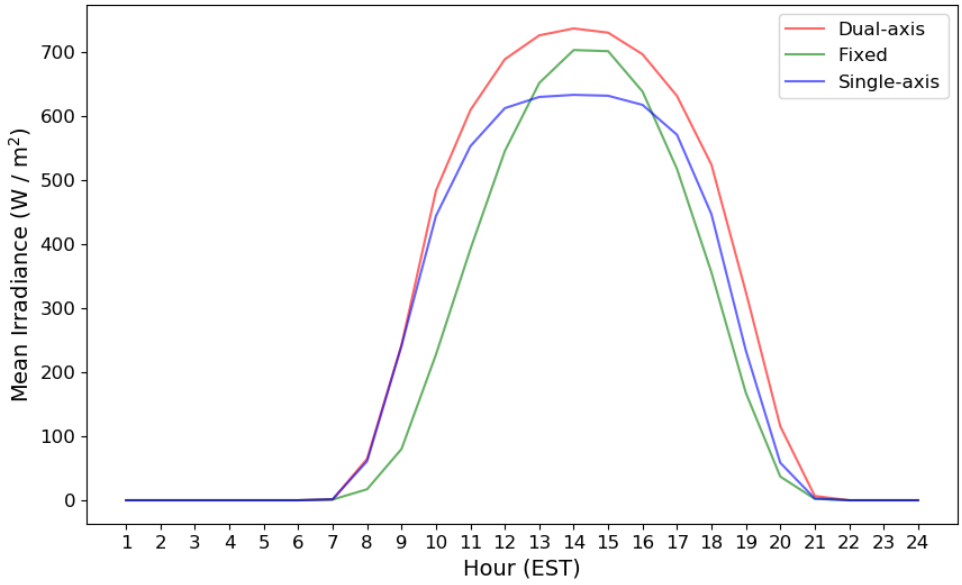

Figure 7: Mean irradiance versus hour-of-day

the peak for longer. As evidenced by Table 10, which displays the area-under-the-curve computed for each array in Figure 7, the average total irradiance received by the single-axis tracking array exceeds that of the fixed array by 13.9%. The dual-axis tracking array reaches and sustains significantly higher irradiance readings than either the single-axis or fixed array. On average, it receives 30.6% greater irradiance per day than the fixed array and 14.7% greater irradiance per day than the single-axis tracking array.

Table 10: Area-under-curve from Figure 7

Average Daily Irradiance	
Dual-Axis	6576.5
Fixed	5033.4
Single-Axis	5733.2

Figure 8 shows the mean irradiance observations from Arrays A, B, and E grouped by month of the year. Irradiance peaks during mid-summer when Earth's northern hemisphere is tilted closer to the sun. The dual-axis tracking array (Array A) significantly outperforms

the fixed array for all months of the year, and the single-axis tracking array (Array E) significantly outperforms the fixed array for most spring and summer months.

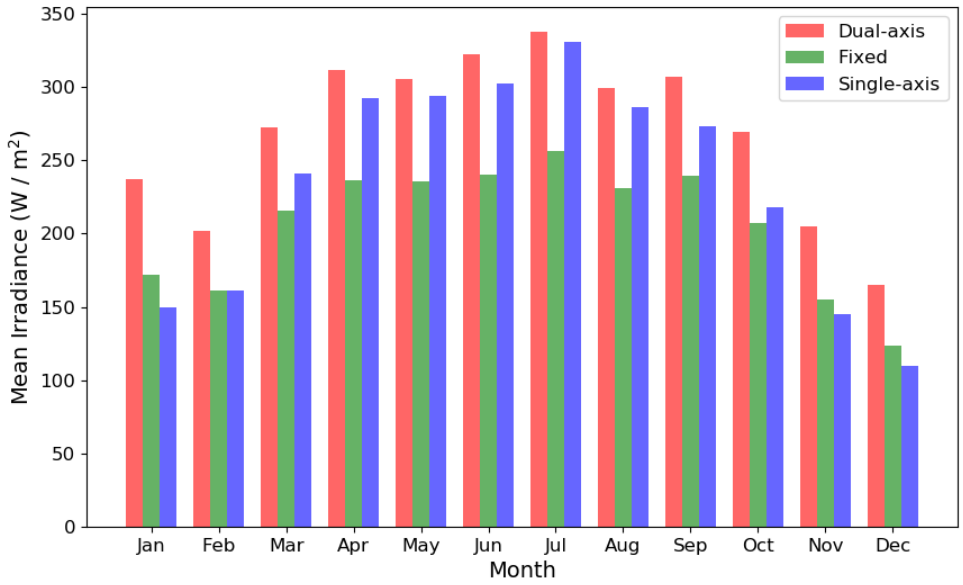

Figure 8: Mean irradiance versus month

As previously mentioned, the NAM forecasts include a prediction of downwelling shortwave radiation flux (DSRF), the quantity of power passed through a given area which is attributable to visible, near-visible, and near-infrared light. The majority of shortwave radiation is comprised of solar radiation, so DSRF is highly correlated with solar irradiance. Figure 9 shows a scatterplot of DSRF versus the irradiance readings from arrays A, B, and E. Although DSRF and irradiance are correlated, DSRF is clearly insufficient to explain the variance in observed solar irradiance. Several other environmental variables significantly impact the quantity of irradiance that reaches the surface of a panel.

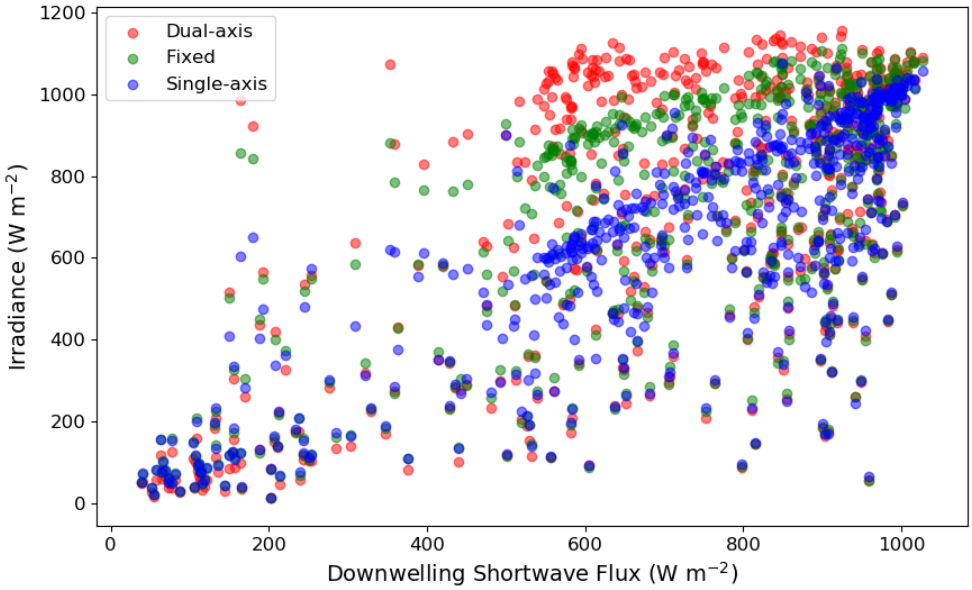

Figure 9: Irradiance Readings versus Downwelling Shortwave Radiation Flux

3.3 Experimental Setup

The goals of this chapter are two-fold. First, a variety of machine learning techniques are compared to determine their relative usefulness for predicting solar irradiance. The differences in model performance are analyzed with respect to both forecast hour and type of solar array. Second, the importance of the geographic coverage of the weather predictions is analyzed. Results from Sanders et al. [28] and from Chapter 2.3.1 suggest that the inclusion of weather predictions from areas surrounding the target site can improve predictive accuracy. The second experiment in this chapter seeks to quantify the improvement obtained for various scopes of coverage.

As in Chapter 2, the models considered in this chapter are trained and evaluated using the open-source Python library *scikit-learn*. The standard technique for evaluating machine learning models is cross-validation. The dataset is split into several *folds*, and the

model is evaluated on each fold in turn after being trained against the other folds. The mean score across all folds provides a robust estimate of the model's performance. However, traditional cross-validation assumes that there is no relationship in the ordering of the data. This assumption is clearly violated by weather observations and predictions, which exhibit significant temporal correlation. Models in this chapter are instead evaluated by training on the data collected during 2017 and testing against the data collected during 2018.

The performance exhibited by many of the models varies based on the values of various hyperparameters. Finding a good set of hyperparameters is paramount to extracting maximum performance. A cross-validated grid search over the parameter space of each model was used to find good hyperparameters. A cross-validated grid search was performed over the parameter space of each model, using parameter regions that are local to the default values as listed in the scikit-learn documentation.

3.4 Results

3.4.1 Model Comparison

The simplest predictive models of solar irradiance incorporate only historical irradiance readings. One naive approach to predict future irradiance is a 24-hour *persistence model*. Such a model predicts irradiance at target hour t using the irradiance value measured at time $t - 24$. This type of model can be easily implemented and serves as a useful baseline when evaluating the performance of more sophisticated models.

The other models considered in this chapter make use of the NAM weather forecasts described in section 3.2, which include several forecasted weather variables at various vertical,

horizontal, and temporal resolutions along with parameters describing the date and time. Forecasted weather variables are included from a three-by-three grid of NAM cells, centered on Athens, GA. For each irradiance observation from January 2017 to January 2018, the matching NAM forecasts from all cells were flattened into a single-dimensional vector, then fed into the following machine learning algorithms: least-squares linear regression (LSLR), Support-Vector Machine (SVM), k-Nearest Neighbors[8] (KNN), M5P Model Tree (M5P), Random Forest (RF), and Extreme Gradient-Boosted Trees (XGBT) [5]. Each model was evaluated using data collected from January 2018 to December 2018.

For each algorithm, separate models were trained to predict each future hour between 1 and 24. To simplify the analysis, the evaluations are split into categories by prediction hour: *Short-term* (future hours 1-6), *Medium-Short* (future hours 7-12), *Medium-Long* (Future hours 13-18), *Long-term* (19-24), and *Overall* (future hours 1-24). The evaluation metrics for each category are computed by taking the mean of evaluations for each target hour included in the category.

Models were first trained against the fixed-axis solar array (Array B). Each model's performance was measured by the criteria of mean absolute error (MAE) and coefficient of determination (R^2). Table 11 lists these metrics for each learning algorithm, and Figure 10 shows a graphical representation of the MAE achieved by each learning algorithm. A model's MAE expresses the average absolute distance between the prediction made by a model and the true value, while the R^2 score represents what proportion of the variance in the target variable is accounted for by the model. The R^2 metric may be useful for comparing results from this thesis to those found elsewhere in the literature.

[8] with the number of neighbors, k, set to 5

Table 11: MAE and R^2 coefficient of several learning algorithms trained against Array B using a 36 km by 36 km grid of numerical weather predictions.

Metric	**Horizon**	**PER**	**LSLR**	**SVM**	**KNN**	**DT**	**RF**	**XGBT**
MAE ($\mathrm{W\,m^{-2}}$)	1-6	209.6	151.8	55.3	61.1	63.3	46.9	47.2
	7-12	209.2	147.2	57.0	63.8	63.7	48.1	49.6
	13-18	209.1	151.1	56.3	68.7	64.3	47.6	49.2
	19-24	208.9	159.3	56.5	66.5	62.3	47.7	49.4
	Overall	209.2	152.3	56.3	65.0	63.4	47.6	48.9
Relative improvement over PER	Overall	-	27.2%	73.1%	68.9%	69.7%	77.2%	76.6%
R^2	1-6	-0.494	0.550	0.902	0.855	0.799	0.901	0.898
	7-12	-0.493	0.559	0.897	0.853	0.794	0.895	0.887
	13-18	-0.493	0.557	0.899	0.839	0.791	0.897	0.887
	19-24	-0.493	0.504	0.897	0.837	0.805	0.897	0.888
	Overall	-0.493	0.542	0.899	0.846	0.797	0.897	0.890

All models exceeded the performance of the persistence baseline. Unsurprisingly, linear regression was unable to capture the cyclical nature of solar irradiance and performed quite poorly with respect to both MAE and R^2 . The k-Nearest-Neighbors model performed reasonably well with k = 5. This result is intuitive: days with similar weather and similar forecasts often exhibit similar irradiance readings. Nonetheless, kNN is outperformed by support vector machines and tree-based models. Support vector machines achieve the third-best MAE and the best R^2 of all learning algorithms trialed. The SVMs in these experiments use a radial basis function (RBF) kernel, which the literature suggests is the strongest choice of kernel function in the domain of time-series forecasting [27, 29]. A single regression tree failed to achieve better accuracy than the support vector machine, but combining multiple regression trees into an ensemble resulted in the best performance among all learning algorithms. Random forest ensembles and gradient-boosted tree ensembles exhibited very similar performance, achieving the lowest MAE and highest R^2.

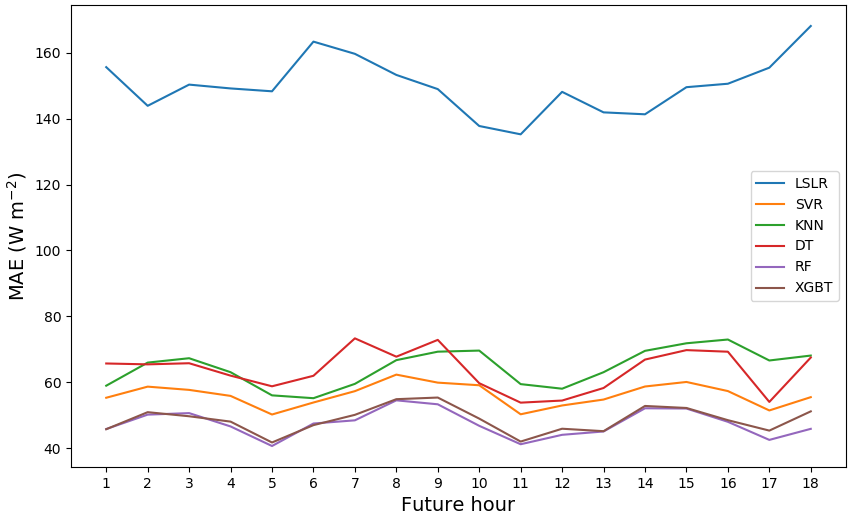

Figure 10: MAE achieved by several learning algorithms trained against Array B

Interestingly, a model's accuracy seems to be nearly independent of its target hour. Intuitively one would expect error to increase with the future horizon, but KNN, random forest, and gradient-boosted trees exhibit nearly identical error for short-term, medium-term, and long-term predictions. Though unexpected, this aligns with similar findings from Chapter 2, where forecast-only models achieved similar error rates for each target hour.

It is also notable that the best accuracy achieved for models targeting irradiance measured by the fixed array is significantly worse than the best accuracy achieved by models in Chapter 2 which targeted stationary pyranometers. Table 12 shows summary statistics for the irradiance targets considered in this paper. Notably, the standard deviation of the irradiance samples considered in this chapter is much higher than that of the GAEMN irradiance samples. There are a couple of possible explanations for the observed differences in variance.

Table 12: Statistical Summary of Irradiance Targets from Chapters 1 & 2

Target	**Mean**	**Std. Dev.**	**Min**	**Max**
GAEMN irradiance	176.76	264.20	0.00	1260.00
Array A	273.80	379.14	-0.03	1156.71
Array B	209.51	308.03	-0.04	1111.69
Array E	238.66	332.86	-0.03	1082.39

The first and most obvious explanation is that differing weather patterns occurred during the sampling periods. It is possible that the weather from 2011-2012 simply exhibited less variation than the weather during 2017-2018. Another possibility is that different models of sensors exhibit different levels of variance and sensitivity to changes in irradiance.

Another explanation is the difference in sampling rates in the GAEMN readings and the solar array readings. The GAEMN irradiance readings from Chapter 2 were sampled at 15 minute intervals, whereas the irradiance readings used in this chapter are sampled every 5 seconds. The higher sampling rate implicitly captures a more detailed, high-variance irradiance signal. Moreover, the comparatively sparse sampling rate for the GAEMN targets may result in *aliasing*, where the irradiance many of the days from Chapter 2 look similar when, in reality, there were details in the irradiance patterns that were not captured. The lower sampling rate of the GAEMN readings may thus artificially reduce the difficulty of the regression problem.

Tables 13 and 14 present results for models trained against the fixed-axis tracking array (Array E) and dual-axis tracking array (Array A). Many of the same trends hold for these two arrays. Similar patterns hold for these arrays. KNN and SVM models perform

moderately well, but tree ensembles achieve significantly better accuracy than any other learning algorithm.

The accuracy of all models targeting irradiance measured by tracking arrays suffered from much lower accuracy in comparison to the non-tracking target. This result is unsurprising given the statistics found in Table 12; both tracking arrays exhibit higher variance that the stationary array. The best overall MAE for the non-tracking array was 47.6 $\mathrm{W\,m^{-2}}$, achieved by a random forest model. Random forest also achieved the best MAE for the single-axis tracking array, but the error increased by a 23.1% to 58.6 $\mathrm{W\,m^{-2}}$. This effect was exacerbated for the dual-axis target, where the random forest's MAE was found to be 75.4 $\mathrm{W\,m^{-2}}$, an increase of 58% of the error in the fixed array. It may be the case that irradiance from a fixed array is easier to predict due to a larger portion of the variance being explained by the time-of-day features and day-of-year features. If the temporal variation is reduced, the model must rely more heavily on weather forecasts, which are unreliable by comparison.

3.4.2 Expansion of Forecast Coverage

Sanders et al. [28] showed that including weather forecasts from an area around the target site improves the accuracy of predictive irradiance models, and this result was confirmed in Section 2.3.1. So far the models developed in this chapter have been trained using a 3x3 grid of NAM forecasts centered on Athens, GA. This is equivalent to a 36 km x 36 km area. This section attempts to quantify the improvements to accuracy gained by incorporating weather forecasts from surrounding grids of various size. The experiments also attempt to determine whether a grid of size 3x3 is optimal. Better accuracy may be achievable if the geographic coverage is widened even further.

Table 13: MAE and R^2 coefficient of several models trained against Array E using a 36 km by 36 km grid of numerical weather predictions.

Metric	**Horizon**	**PER**	**LSLR**	**SVM**	**KNN**	**DT**	**RF**	**XGBT**
MAE (Wm-2)	1-6	239.5	174.4	71.6	83.6	77.4	57.4	57.7
	7-12	239.2	176.2	73.5	86.6	81.2	60.5	61.4
	13-18	239.0	176.6	72.2	91.1	80.7	58.3	60.2
	19-24	238.8	182.2	72.1	87.6	80.9	58.7	60.9
	Overall	239.1	177.3	72.4	87.2	80.1	58.7	60.0
Relative improvement over PER	Overall	-	25.8%	69.7%	63.5%	66.5%	75.4%	74.9%
R^2	1-6	-0.539	0.491	0.879	0.799	0.774	0.891	0.887
	7-12	-0.538	0.466	0.872	0.794	0.754	0.877	0.871
	13-18	-0.537	0.481	0.873	0.777	0.756	0.883	0.874
	19-24	-0.537	0.456	0.873	0.784	0.753	0.883	0.873
	Overall	-0.538	0.473	0.874	0.788	0.759	0.884	0.876

Table 14: MAE and R^2 coefficient of several models trained against Array A using a 36 km by 36 km grid of numerical weather predictions.

Metric	**Horizon**	**PER**	**LSLR**	**SVM**	**KNN**	**DT**	**RF**	**XGBT**
MAE	1-6	273.0	230.8	90.5	98.7	98.0	74.7	73.0
	7-12	272.7	231.1	91.5	101.8	105.5	77.3	77.0
	13-18	272.5	232.3	90.9	106.6	99.6	74.7	75.6
	19-24	272.1	243.0	90.7	104.2	97.7	75.0	76.7
	Overall	272.6	234.3	90.9	102.8	100.2	75.4	75.6
Relative improvement over PER	Overall	-	14.0%	66.7%	62.3%	63.2%	72.3%	72.3%
R^2	1-6	-0.546	0.318	0.851	0.787	0.724	0.863	0.864
	7-12	-0.545	0.298	0.848	0.785	0.687	0.853	0.850
	13-18	-0.544	0.309	0.849	0.774	0.708	0.860	0.852
	19-24	-0.544	0.254	0.846	0.773	0.719	0.860	0.850
	Overall	-0.545	0.295	0.848	0.780	0.710	0.859	0.854

Two of the best-performing models from section 3.4.1, k-nearest neighbors (KNN) and random forest (RF), were retrained on fixed-array targets (Array B) from 2017 using grids of size 1x1, 3x3, 5x5, and 7x7, then each model was evaluated on the data from 2018. Like the previous experiment, the models cover future hours 1-24 by training a separate model for each future hour, and results are grouped by *Short-term*, *Medium-Short*, *Medium-Long*, and *Long-term* predictions.

Table 15: MAE and R^2 coefficient of models trained against Array B using a variety of NWP grid sizes.

Metric	**Horizon**	**KNN**				**RF**			
		1x1	*3x3*	*5x5*	*7x7*	*1x1*	*3x3*	*5x5*	*7x7*
MAE	*1-6*	64.9	61.1	60.0	59.3	53.6	52.9	53.1	51.7
	7-12	68.8	63.8	63.4	61.9	54.0	52.6	51.6	50.9
	13-18	95.7	68.7	66.9	66.2	52.6	51.1	49.8	49.6
	19-24	80.4	66.5	64.7	64.3	52.8	52.0	51.3	50.7
	Overall	77.4	65.0	63.8	62.9	53.2	52.2	51.5	51.0
Relative improvement over 1x1	*Overall*	-	16.0%	17.6%	18.7%	-	1.9%	3.2%	4.1%
R^2	*1-6*	0.832	0.855	0.858	0.861	0.891	0.890	0.889	0.895
	7-12	0.827	0.853	0.853	0.860	0.882	0.887	0.892	0.896
	13-18	0.710	0.839	0.844	0.848	0.890	0.895	0.899	0.900
	19-24	0.765	0.837	0.844	0.846	0.891	0.894	0.897	0.899
	Overall	0.783	0.846	0.850	0.854	0.888	0.891	0.894	0.898

As evidenced by table 15, both models benefit from wider geographic coverage of forecasted weather variables, although the improvement diminishes as the size grows larger. The k-Nearest Neighbor model experienced the most dramatic improvement, with overall MAE dropping from 77.4 $\mathrm{W\,m^{-2}}$ using a 1x1 grid to 65.0 $\mathrm{W\,m^{-2}}$ when using a 3x3 grid, then 62.9 $\mathrm{W\,m^{-2}}$ using a 7x7 grid. This represents a 18.7% improvement in accuracy. The most dramatic improvements were seen in *Medium-long* and *Long-term* predictions, where overall MAE dropped by 30.8% and 20.0% respectively. The effect was less pronounced but still

significant for short-term forecasts, where MAE dropped by 8.6% from the 1x1 model to the 7x7 model.

The effect on the random forest model was much less dramatic, with overall MAE dropping only 4.1% from 53.2 $\mathrm{W\,m^{-2}}$ using a 1x1 grid to 51.0 $\mathrm{W\,m^{-2}}$ using a 7x7 grid. Unlike the KNN model, the magnitude of the improvement was nearly the same for all target hours. The MAE for *Long-term* hours dropped by 4.0% from 52.8 $\mathrm{W\,m^{-2}}$ to 50.7 $\mathrm{W\,m^{-2}}$, while the MAE for short-term hours dropped by 3.5% from 53.6 $\mathrm{W\,m^{-2}}$ to 51.7 $\mathrm{W\,m^{-2}}$.

The difference in the magnitude of improvement observed between the two models may be explainable by the inherent properties of the algorithm used to train the model. The k-Nearest Neighbors model predicts irradiance using the five instances of weather conditions which are most similar to the one being predicted. KNN thus draws significant benefit from the added weather information, since it can more accurately determine which examples from the dataset most closely match the conditions of the instance under consideration. Random forest, on the other hand, is a tree-based method which makes a hierarchical series of decisions about specific weather features. The most important features are likely the weather variables observed in the same geographic region as the target array. These features will appear high in the hierarchy and will heavily impact the decisions made by the model, while the less-significant information found in the surrounding cells will have comparatively little impact.

Although error rates decrease as the size of the grid increases, this effect suffers from diminishing returns for grids of size larger than 3x3. KNN models, for example, experience a 16.0% error reduction in MAE moving from a 1x1 grid to a 3x3 grid, but the additional error reduction when moving from a 3x3 grid to a 7x7 grid is only 2.7%. Similarly, the error of

random forest models is reduced by 1.9% when the grid size increases from 1x1 to 3x3, but only 1.3% as the size increases from 3x3 to 5x5 and only 0.9% as the size increases from 5x5 to 7x7.

It is reasonable to conclude that the trend of diminishing gains in accuracy holds for grids of size larger than 7x7. A naive interpretation of this finding would suggest training irradiance models using the maximum-sized grid that is possible given constraints on memory, storage, and CPU time. However, this strategy is unlikely to succeed. The purpose of the grid is to account for the possibility of nearby weather conditions moving into the target region. As the grid size grows, it will inevitably contain information that will have very little bearing on the weather of the target site. Both instance-based learning algorithms, such as KNN, and random forest models will suffer under conditions where many spurious attributes are present in the data.

3.5 Conclusion

The goal of this chapter was to investigate predictive techniques for irradiance observations taken from solar arrays of various types in Athens, GA. The results from this chapter indicate that tree-based machine learning methods are capable of more robust and accurate irradiance predictions than regressive models popular in the literature such as least-squares linear regression and support-vector machines. This appears to be the case particularly for tree ensemble models, which average the predictions from several tree-based models. This result holds for models that predict irradiance for any future hours from 1 through 24, and for each of the types of arrays commonly used on solar farms.

Although single- and dual-axis tracking arrays are capable of higher energy yields than their stationary counterparts, the quantity of irradiance collected by the tracking arrays is more difficult to predict with the techniques considered in this chapter. Further work may be necessary to improve the accuracy of models targeting tracking arrays. One possible avenue of improvement would be to add features describing the position of the array to the model's input.

This chapter also analyzed the effect of increasing the geographic coverage of the weather forecasts which are used as model inputs. Results suggest that greater predictive accuracy is possible when weather information from a wide local area is included, but non-parametric regression models such as k-Nearest Neighbors experience a greater benefit than tree-based models. The magnitude of this effect was found to diminish quickly as the size of the forecast coverage increases beyond a 36 km by 36 km grid of surrounding cells.

There are several avenues by which the work in this chapter could be extended. The NAM weather forecasts include dozens of variables across many dimensions which are spatially and temporally correlated. In recent years, deep neural networks that utilize convolutional layers and recurrent mechanisms have achieved state-of-the-art performance on similarly challenging problems. Deep learning techniques may yield improved performance if sufficient training data could be collected.

A second potential avenue of improvement was identified by Sanders et al. [28], who showed that the accuracy of weather forecasts can be improved by post-processing to remove historically observed bias. The numerical weather predictions used to train models could be post-processed to remove bias in the forecasted weather variables before feeding them into

the models. This may result in improved accuracy, but it would require the training of two sets of models - one set to post-process NWPs and one set to forecast irradiance.

Finally, the analyses in this chapter consider only a single site in central Georgia. It may be informative to conduct similar experiments for a broader range of climates, particularly regions which experience more sudden and volatile meteorological changes. It is possible that the importance of certain model inputs fluctuates with the predictability of the local weather.

4 Conclusion & Future Directions

4.1 Conclusion

The results presented in this thesis indicate that properly-chosen machine learning techniques can be used to generate models of solar irradiance, and there are several techniques by which the accuracy of such models may be improved. Chapter 2 affirms previous findings from the literature which demonstrate the importance of incorporating predictive data such as weather forecasts in models of irradiance for timeframes between 1 and 24 hours. It was also shown that the importance of predictive data overtakes that of historical or current data for any predictive model beyond the very short-term. Results from Chapter 2 also indicate that machine learning models of solar irradiance are, by default, highly specialized to the climate or subclimate matching the data used to train the models. The capability of models to generalize to new climates can be improved by including a diversity of climates in the training data, but this technique still fails to match the accuracy of models specialized to a single climate.

Chapter 3 presents a case-study in building irradiance models for a solar farm which take only forecasted weather variables as inputs. Models built using the random forest algorithm achieved superior performance over many other contemporary techniques for all future hours between 1 and 24. It was found that the models developed in this thesis are significantly more successful at predicting the irradiance observed by a stationary solar array

than predicting the irradiance observed by 1- or 2-axis tracking arrays. This chapter also investigated the relationship between the accuracy attained by a model and the geographic coverage of the forecasts used to train the model. It was found that including forecasted variables from a 36 km by 36 km region centered on the target site improved accuracy significantly compared to models that consider only a 12 km by 12 km region. Wider regions were also analyzed, and further improvements in accuracy are possible as the area covered by the forecast expands, but the magnitude of the improvements quickly diminish.

As mentioned at the beginning of this thesis, solar energy is a promising power source and will likely play a key role in the transition to a renewable, environmentally-friendly power grid. Improvements in the accuracy of irradiance models will help smooth this transition by allowing energy providers to rely more heavily on solar plants and ensure uninterrupted service to their customers. The results reported in this thesis suggest that machine learning techniques such as random forest are well-suited to build accurate models of future irradiance, particularly when predictive data such as numerical weather predictions are available. Widespread dissemination of these findings may aid in the expansion of solar systems and their integration into the grid.

4.2 Future Directions

One important avenue of research extension identified by Sanders et al. [28] would be to analyze a broader variety of climates and macroclimates. Section 2.3.4 has taken the first steps in this direction by analyzing the portability of models across a variety of weather stations in Georgia, but a much wider range of weather conditions can be found throughout

the world. It remains to be shown that the results found in this thesis hold for weather systems which are dissimilar to those found in Georgia.

The accuracy of machine learning models may also benefit from the inclusion of a wider variety of data sources not explored in this thesis. The models considered so far have incorporated historical irradiance data, current weather observations, numerical weather predictions from two NOAA models, date & time parameters, or some combinations of these inputs. Incorporating other sets of inputs such as satellite imagery and Doppler radar may further increase accuracy. As mentioned in Chapter 3, the inclusion of additional non-meteorological parameters describing the position of tracking arrays may help achieve accuracy comparable to their fixed-axis counterparts. It also may be interesting to experiment with a larger quantity of data. Most of the models described so far were trained using only a year or two of numerical weather predictions and historical observations.

Results from both Chapter 2 and Chapter 3 discuss some of the available trade-offs between model efficiency and accuracy. For example, it was found that increasing the size of the area covered by numerical weather predictions may increase accuracy, but at the cost of storage space and training time. One interesting trade-off that has not been considered is the development of different models for each target hour versus the use of a single model that predicts all target hours at once. Many machine learning techniques, including kNN, random forest, gradient-boosted trees, and neural networks, allow models to have an arbitrary number of outputs. It would be more efficient to train a single model which predicts irradiance for all future hours of interest. It would be interesting to compare the accuracy of such a model with the accuracy achieved in this thesis, where a separate model was trained for each future hour.

Finally, the accuracy of models could likely be enhanced by improving the quality of weather predictions which are used an inputs. There are several ways such an improvement might be accomplished. One simple approach would be to combine weather predictions from multiple sources. This may be difficult, however, given the differences in coverage, geographic granularity, temporal granularity, and types of forecasted variables among the available models. Another approach would be to post-process weather predictions to remove historical bias. Many results from the literature suggest that the accuracy of numerical weather forecasts can be improved using statistical or machine learning post-processing techniques [4, 26]. Finally, it may be helpful to include weather forecasts with finer spatial resolution. The NAM numerical weather predictions used throughout this thesis are released on a 12 km by 12 km grid. A more granular model may provide more detailed local weather information and may allow for better tuning of the grid size described in Chapter 3.

Bibliography

[1] International Energy Agency. *World Energy Outlook 2018*. 2018, p. 661. DOI: https://doi.org/https://doi.org/10.1787/weo-2018-en. URL: https://www.oecd-ilibrary.org/content/publication/weo-2018-en.

[2] Peder Bacher, Henrik Madsen, and Henrik Aalborg Nielsen. "Online short-term solar power forecasting". In: *Solar Energy* 83.10 (2009), pp. 1772–1783.

[3] Tim Bollerslev. "Generalized autoregressive conditional heteroskedasticity". In: *Journal of econometrics* 31.3 (1986), pp. 307–327.

[4] Jie Chen, François P Brissette, and Zhi Li. "Postprocessing of ensemble weather forecasts using a stochastic weather generator". In: *Monthly Weather Review* 142.3 (2014), pp. 1106–1124.

[5] Tianqi Chen et al. "Xgboost: extreme gradient boosting". In: *R package version 0.4-2* (2015), pp. 1–4.

[6] Carlos FM Coimbra, Jan Kleissl, and Ricardo Marquez. "Overview of solar-forecasting methods and a metric for accuracy evaluation". In: *Solar energy forecasting and resource assessment* (2013), pp. 171–194.

[7] Mathieu David, M Diagne, and Philippe Lauret. "Outputs and error indicators for solar forecasting models". In: *Proceedings of the World Renewable Energy Forum (WREF)*. 2012, pp. 13–17.

[8] M David et al. "Probabilistic forecasting of the solar irradiance with recursive ARMA and GARCH models". In: *Solar Energy* 133 (2016), pp. 55–72.

[9] Furkan Dincer. "The analysis on photovoltaic electricity generation status, potential and policies of the leading countries in solar energy". In: *Renewable and Sustainable Energy Reviews* 15.1 (2011), pp. 713–720.

[10] Teresa Lo Feudo et al. "Comparison of hourly solar radiation from a ground–based station, remote sensing and weather forecast models at a coastal site of South Italy (Lamezia Terme)". In: *Energy Procedia* 76 (2015), pp. 148–155.

[11] Baptiste Gregorutti, Bertrand Michel, and Philippe Saint-Pierre. "Correlation and variable importance in random forests". In: *Statistics and Computing* 27.3 (2017), pp. 659–678.

[12] Martin T Hagan and Mohammad B Menhaj. "Training feedforward networks with the Marquardt algorithm". In: *IEEE transactions on Neural Networks* 5.6 (1994), pp. 989–993.

[13] Geoffrey Holmes, Andrew Donkin, and Ian H Witten. "Weka: A machine learning workbench". In: *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*. IEEE. 1994, pp. 357–361.

[14] Gerrit Hoogenboom, BP Verma, and ED Threadgill. "The development of the Georgia automated environmental monitoring network". In: Georgia Institute of Technology. 1991.

[15] T Jensen et al. *Metrics for evaluation of solar energy forecasts*. Tech. rep. NCAR, Tech. Rep., 2016.[Online]. Available: http://dx. doi. org/10.5065/D6RX99GG, 2016.

[16] Ehsanul Kabir et al. "Solar energy: Potential and future prospects". In: *Renewable and Sustainable Energy Reviews* 82 (2018), pp. 894–900.

[17] Athina Korfiati et al. “Estimation of the Global Solar Energy Potential and Photovoltaic Cost with the use of Open Data”. In: *International Journal of Sustainable Energy Planning and Management* 9 (2016), pp. 17–30.

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

[19] Philippe Lauret et al. “A benchmarking of machine learning techniques for solar radiation forecasting in an insular context”. In: *Solar Energy* 112 (2015), pp. 446–457.

[20] 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.

[21] Gaëtan Masson et al. *Snapshot of Global Photovoltaic Markets (1992-2017)*. Tech. rep. T1-33:2018. 2018. URL: `http://www.iea-pvps.org/fileadmin/dam/public/report/statistics/IEA-PVPS_-_A_Snapshot_of_Global_PV_-_1992-2017.pdf`.

[22] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *Journal of machine learning research* 12.Oct (2011), pp. 2825–2830.

[23] Hugo TC Pedro and Carlos FM Coimbra. “Assessment of forecasting techniques for solar power production with no exogenous inputs”. In: *Solar Energy* 86.7 (2012), pp. 2017–2028.

[24] 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.

[25] Ibrahim Reda. "Method to calculate uncertainties in measuring shortwave solar irradiance using thermopile and semiconductor solar radiometers". In: *Contract* 303 (2011), pp. 275–3000.

[26] DE Robertson, DL Shrestha, and QJ Wang. "Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting". In: *Hydrology and Earth System Sciences* 17.9 (2013), pp. 3587–3603.

[27] Stefan Rüping. *SVM kernels for time series analysis*. Tech. rep. Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten ..., 2001.

[28] Sam Sanders et al. "Solar Radiation Prediction Improvement Using Weather Forecasts". In: *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*. IEEE. 2017, pp. 499–504.

[29] Nicholas I Sapankevych and Ravi Sankar. "Time series prediction using support vector machines: a survey". In: *IEEE Computational Intelligence Magazine* 4.2 (2009), pp. 24–38.

[30] NOAA National Weather Service. *The North American Mesoscale Forecast System*. 2019. URL: `https://www.emc.ncep.noaa.gov/index.php?branch=NAM` (visited on 04/09/2019).

[31] A Sfetsos and AH Coonick. "Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques". In: *Solar Energy* 68.2 (2000), pp. 169–178.

[32] Navin Sharma et al. "Predicting solar generation from weather forecasts using machine learning". In: *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on*. IEEE. 2011, pp. 528–533.

[33] Ruey S Tsay. *Analysis of financial time series*. Vol. 543. John Wiley & Sons, 2005.

[34] Cyril Voyant et al. "Machine learning methods for solar radiation forecasting: A review". In: *Renewable Energy* 105 (2017), pp. 569–582.

[35] LT Wong and WK Chow. "Solar radiation model". In: *Applied Energy* 69.3 (2001), pp. 191–224.