Optimizing Solar Capacity Forecasting in the U.S

Caden R. Truelick, Jayant Babu, Jacob Helgason, Kishan A. Bhakta
Arizona State University

May 1, 2024

1 Abstract

This study employs an ensemble of advanced regression models to predict county-level solar capacity in Megawatts of Alternating Current (MW-AC) across the United States, identifying regions where solar energy resources are underutilized. Solar capacity is a metric used in solar power generation that identifies the maximum generation of a plant. By analyzing Utility-Scale Plant Data from the National Renewable Energy Laboratory[3] with a variety of machine learning techniques, we aim to find counties which have the natural conditions such as weather trends, solar irradiance, precipitation trends, and land area to have high solar capacity, but significantly under utilize those resources, suggesting areas for targeted energy policy and infrastructure investment. We also use similar data to conduct a time-series analysis with the intention of creating a 7-day forecast for global horizontal irradiance (GHI) in a given area. This approach not only helps mitigate climate change but also supports informed decision-making in both public and private sectors, promoting a more efficient and responsible energy landscape.

2 Introduction

As global concerns about sustainability and climate change drive the shift from fossil fuels to renewable energy sources, solar energy emerges as a pivotal component of this transformative landscape. Offering a clean and sustainable alternative, solar power has witnessed exponential growth field by technological advancements, supportive policy incentives, and dynamic market forces. This surge not only underscores solar energy's vast potential but also illuminates the challenges and opportunities its adoption presents.[4]

Central to the renewable energy transition are solar technologies, ranging from photovoltaic cells to advanced solar thermal systems[6]. These innovations are crucial in reducing carbon emissions and enhancing the efficiency of energy systems globally. Significant investments in research and development have propelled these technologies forward, reducing costs and improving the practicality of solar solutions. Moreover, national and international policy frameworks have increasingly sought to integrate solar energy more fully into the mainstream energy mix.[1]

However, the widespread adoption of solar energy is fraught with challenges. The intermittent nature of solar energy, exacerbated by fluctuating weather conditions, poses significant hurdles in energy production forecasting, which is critical for grid integration and effective energy management. Reliable forecasting models are essential, as highlighted by Sankhe (2023)[8], who emphasizes the ongoing struggle to maximize prediction accuracy amidst environmental unpredictabilities. Despite advancements in machine learning models, there remains a considerable gap in achieving optimal predictive performance, suggesting the need for inovative approaches that can handle the complexity of solar outputs.

Studies like that of Abdulai et al.[2] in Ghana have shown how advanced predictive models using techniques such as random forests and gradient boosting can address these challenges by enhancing the accuracy of solar generation predictions, thereby stabilizing energy supply and supporting grid management.

This study contributes to this effort by developing predictive models to accurately forecast solar capacity at the county level across the United States, measured in Megawatts of Alternating Current (MW-AC). Utilizing data from the National Renewable Energy Laboratory[3] and employing an ensemble of regression models, including Random Forest, SVM, Gradient Boosting, and Decision Trees, this research aims to identify regions where solar resources are underutilized. The ultimate goal of our project is to enhance the technical understanding and practical application of solar capacity forecasting, thereby contributing to the stability and sustainability of the energy supply.

3 Methods

3.1 Data Collection and Preparation

We utilized the National Solar Radiation Database (NSRDB)[3], renowned for its comprehensive, high-resolution solar irradiance data which is crucial for assessing solar power generation potential. The dataset includes measurements such as global horizontal (GHI), direct normal (DNI), and diffuse horizontal irradiance (DHI). We enhanced this dataset with additional environmental data from NASA's Goddard Institute for Space Studies[7, 10] and utility-scale solar project information from the Lawrence Berkeley National Laboratory[9] to ensure a robust dataset. The data preparation phase involved several key steps: transforming various data formats, such as raster and tabular data, into county-specific formats to suit our geographical analysis needs; aggregating the capacities of solar projects to their respective counties; and conducting extensive data cleaning to ensure the precision of our mapping and the integrity of our data, correcting any discrepancies identified in the original sources.

3.2 Exploratory Data Analysis (EDA) and Visualization

Our exploratory data analysis primarily focused on visually understanding the distribution and characteristics of solar power generation, weather, and temperature data across the U.S. Our data from the Lawrence Berkeley National Laboratory[9] was able to be shown as a scatter plot of Utility Scale Solar plants. Our data from GISS[7, 10] provided us with monthly averages for GHI and temperature in raster data which would be some of the primary data we used. We were able to condense our raster data by grouping data points using county shape files and then averaging points within each county group. We also explored the correlation between solar generation as a percentage of in-state load and solar generation as a percentage of in-state generation to provide a quantitative understanding of their relationship. A bar chart provided a comparative view of solar generation percentages across states, effectively highlighting how various states stack up in utilizing solar energy relative to their total energy mix.

3.3 Predictive Modeling

3.3.1 Ensemble Regression Models

To predict solar capacity at the county-level, we utilized a simple averaging ensemble approach combining Random Forest, SVM, Gradient Boosting, and Decision Trees. Each model made predictions on the same sample set and the predictions of all models were averaged to create

our ensemble model predictions. This method as chosen to effectively handle the variations and complexities inherent in the geographic data of solar capacities. The ensemble approach allowed for a balanced analysis, minimizing the weaknesses of using any single model while improving the reliability of the predictions.

3.3.2 SARIMA Model for GHI Forecasting

The SARIMA (Seasonal AutoRegressive Integrated Moving Average) model was developed to address the time-series prediction of GHI for counties, focusing on short-term (seven-day) forecasts. This model takes into account both non-seasonal and seasonal factors, identified through preliminary analyses like the Auto-correlation Function (ACF) and Partial Auto-correlation Function (PACF) plots. The SARIMA model is specifically tailored to capture the cyclical nature of solar irradiance, which is crucial for accurate forecasting in solar capacity planning and energy management.

4 Results

4.1 Exploratory Data Analysis (EDA) Outcomes

Our exploratory data analysis was able to help understand general trends in the data like GHI patterns throughout the country and the year, as well as distribution of solar plants across the US. Through these visualizations we can see the Southwest U.S. has a fairly dense population of solar plants and high GHI throughout the year. We also see that the East coast has a very dense population of solar plants, with a much less variable GHI throughout the year. Both of these regions are represented well in the bar plot of Solar Generation Percentage of states. Because all three of these data sources have similar trends, we have reason to believe our data is leading us in the right direction for our models.

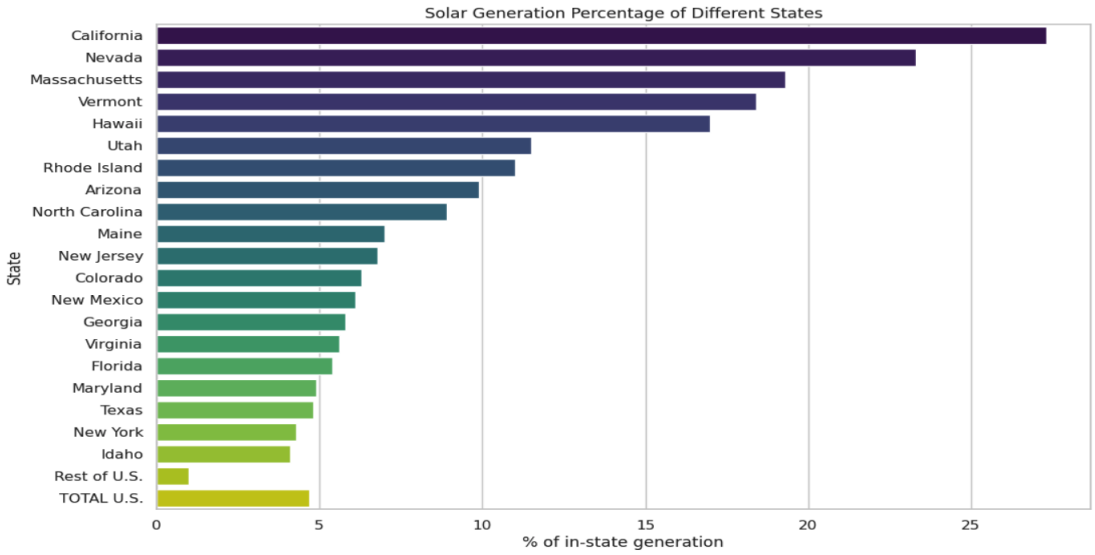


Figure 1: Solar Generation Percentage by State

Figure 2 shows all of the power plants in the United States.

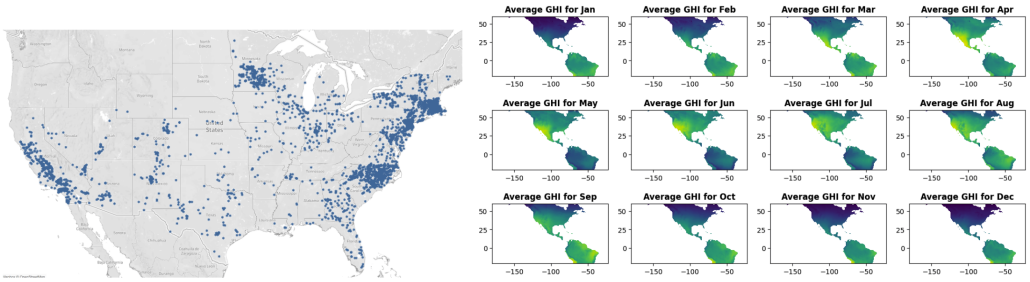


Figure 2: Map of Solar Power Plants in the US (left) and a map of the average GHI by month in the US (right)

4.2 Predictive Modeling Performance

To evaluate the performance of all of our regression models, we used an R Squared scoring method of each model individually as well as our ensemble model. The Random Forest model scored 0.797, the SVM model scored 0.019, the Gradient Boost model scored 0.881, and the Decision Tree scored 0.736. Looking at the SVM, it appears to have done very poorly because of the high proportion of counties with no recorded solar capacity drastically skewing all of it's predictions. After averaging all of the predictions for the models and calculating the R Squared score, the ensemble model scored a 0.818. After plotting our model's prediction, its performance looks fairly similar to the true capacity data.

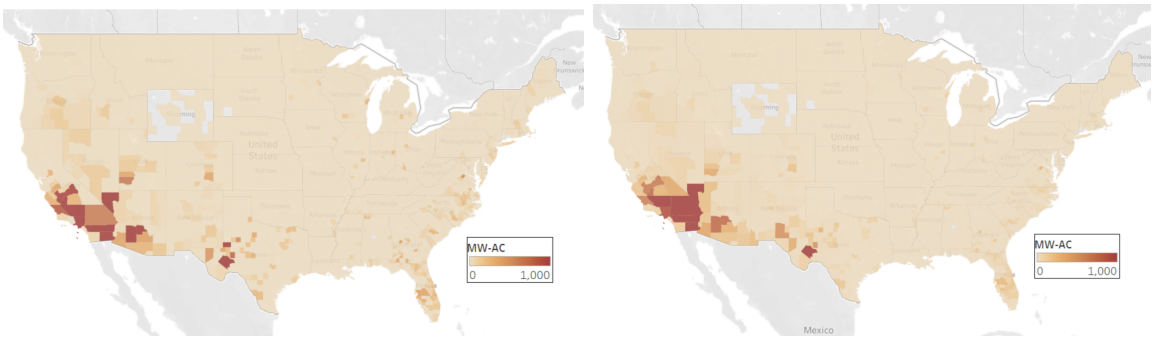


Figure 3: The true (left) and predicted (right) values of Utility Scale Solar Capacity

The darker counties have higher capacities and are very closely aligned between the two maps. The model generally smoothed out high-capacity areas like in Southern California and Western Texas which shows us that the model is able to understand that surrounding counties with similar conditions have similar potential capacities. One side effect of this is that more isolated counties in the North-Western and Central parts of the country seem to be drowned out by their surrounding counties that have little-to-no solar capacity. The practical goal of our investigation is to identify the counties in which our model's predicted capacity was higher than the true capacity, which can be modeled by simply calculating the difference between the predicted and true capacities.

Orange counties are counties where the model under-estimated the solar capacity and blue counties are where the model over-estimated the capacity, indicating a county which meets our criteria. The sporadic orange counties throughout the country further show that the model tends to distribute capacity more evenly around counties with higher capacity. We believe that this is because the uneven distribution of solar capacity across the U.S. allowed our model to learn that less solar capacity is more normal than a higher capacity, as can be seen in the large

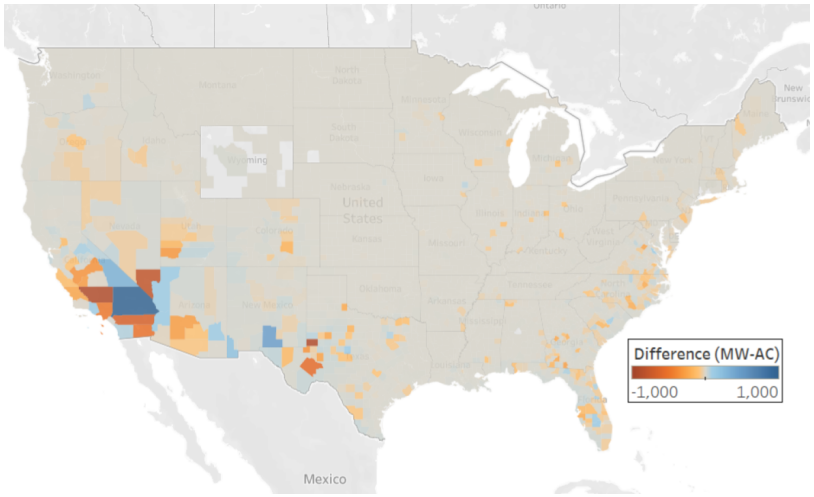


Figure 4: The difference in predicted and true values of Solar Capacity for each county, blue being over-predicting and orange, under-predicting

amount of empty space in the true and predicted capacity visualizations. In this visualization, we can also see that the areas with more generation also have stronger predictions through their darker coloring.

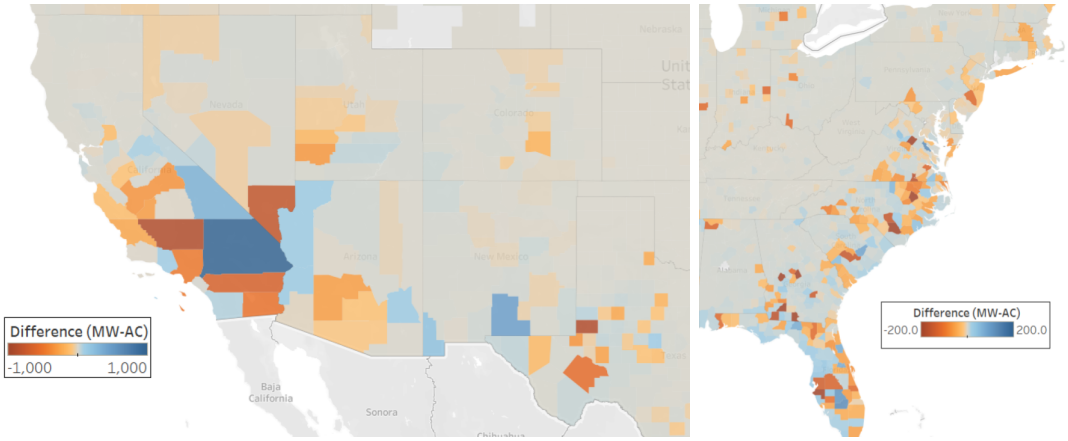


Figure 5: The Southwest region (left) and East coast region (right) difference between Predicted and True Solar Capacity. These two are scaled differently to highlight each particular area

The Southwest region had the lowest under-predictions with California counties like Kern county (897 MW below true), Riverside county (656 MW), Imperial County (523 MW), and Los Angeles County (474 MW). It also had some of the highest over-predictions with San Bernardino, California (974 MW over true), Otero, New Mexico (451 MW), Inyo, California (323 MW), and Hidalgo, New Mexico (203 MW). The East coast the highest over-predictions in Virginia with Essex county 120 MW over true and in Florida with Glades and Highlands county with 123 MW and 110 MW, respectively.

4.3 SARIMA Forecasting Results

Once we had the data prepared, we focused on forecasting the Global Horizontal Irradiance (GHI). The initial analysis included a Fourier Transform, the results of which are shown in

Figure 6, to detect any dominant trends or seasonality. This analysis revealed significant periodic components in the GHI data, indicative of its cyclical nature. To assess the stationarity of the time series, an Augmented Dickey-Fuller (ADF) test was performed. The results indicated non-stationarity, as the test statistic did not surpass the critical values, and the p-value exceeded the standard threshold of 0.05, suggesting that the series had a unit root, which implies that its statistical properties, such as mean and variance, could change over time. To address non-stationarity, we employed differencing, a technique crucial for stabilizing the mean of the time series. The ACF and PACF plots were examined to determine the appropriate model parameters. The ACF plot showed a gradual decline in correlation, while the PACF plot identified a significant spike at the first lag and others, suggesting the presence of a higher-order autoregressive process and potential seasonality at lag 18.

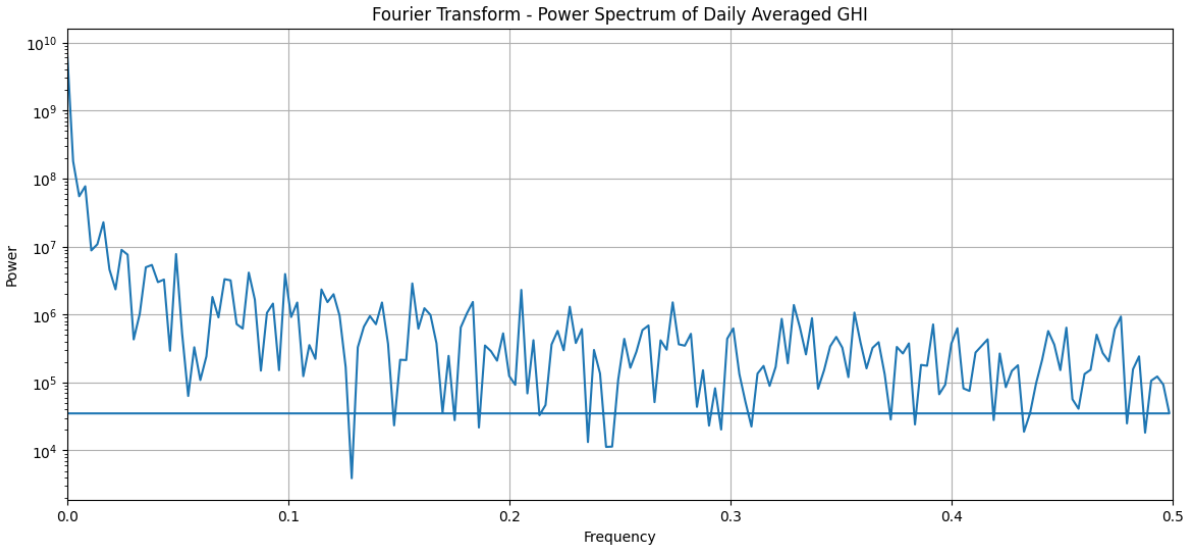


Figure 6: Fourier Analysis

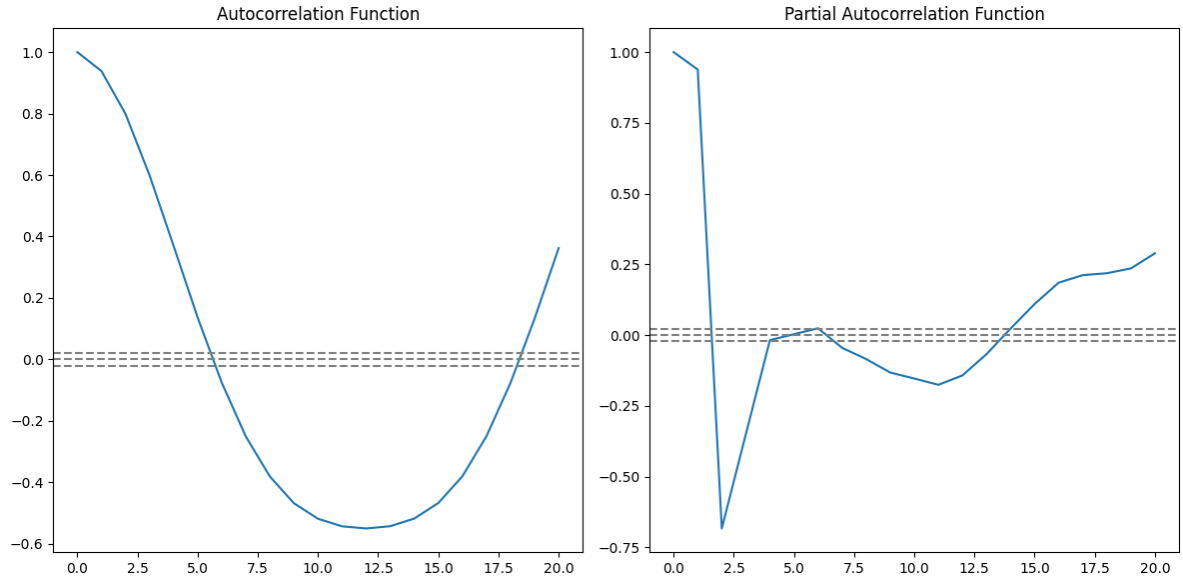


Figure 7: ACF and PACF tests

```
{'Test Statistic': -2.435701578693906,
 'p-value': 0.13191234463069257,
 'Number of Lags Used': 5,
 'Number of Observations Used': 359,
 'Critical Values': {'1%': -3.4486972813047574,
  '5%': -2.8696246923288418,
  '10%': -2.571077032068342}}
```

Figure 8: ADF Test

SARIMAX Results

Dep. Variable:	GHI	**No. Observations:**	8271
Model:	SARIMAX(1, 1, 1)x(0, 1, [1, 2], 7)	**Log Likelihood**	-1807.459
Date:	Tue, 30 Apr 2024	**AIC**	3624.918
Time:	03:33:06	**BIC**	3660.006
Sample:	02-01-1999 - 09-23-2021	**HQIC**	3636.910

Covariance Type: opg

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.2338	0.125	-1.867	0.062	-0.479	0.012
ma.L1	-0.0167	0.132	-0.127	0.899	-0.275	0.241
ma.S.L7	-0.7019	0.045	-15.520	0.000	-0.791	-0.613
ma.S.L14	-0.3076	0.044	-7.002	0.000	-0.394	-0.221
sigma2	1757.9505	85.664	20.521	0.000	1590.052	1925.849

Ljung-Box (L1) (Q):	8.95	**Jarque-Bera (JB):**	6379288.81
Prob(Q):	0.00	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.60	**Skew:**	-2.37
Prob(H) (two-sided):	0.00	**Kurtosis:**	139.17

Figure 9: SARIMAX Model for San Bernardino County, California

Considering these insights, a Seasonal ARIMA with eXogenous variables (SARIMAX) model was deemed suitable for capturing the seasonal effects apparent in the data, specifically applied to San Bernardino County. SARIMAX extends the SARIMA model by incorporating external factors that could impact the forecasting, allowing for a more detailed and adjusted analysis. We selected a SARIMAX configuration with a non-seasonal MA component, as indicated by the ACF's slow decay, and a seasonal AR component, as suggested by the significant spikes in the PACF, especially at lag 18. The SARIMAX model's parameters were fine-tuned, incorporating differencing to achieve stationarity, and seasonal elements to account for periodic trends observed in the Fourier analysis. Once we built the model, we created a custom function that would run the whole process based on the given county dataset and provide a 7-day forecast along with all relevant information about the dataset and our SARIMAX model.

In our SARIMAX model summary, tailored to forecast Global Horizontal Irradiance (GHI) for San Bernardino County, we note significant coefficients, particularly for AR.L1 at -0.2338 and MA.L1 at -0.0167, both with p-values suggesting significant effects on the model (p-value for AR.L1 at 0.062 and for MA.L1 at 0.899). We employed one non-seasonal difference and a seasonal difference with a seven-day period to account for weekly patterns in the data. The model's fit is quantitatively assessed by AIC (3624.918), BIC (3660.006), and HQIC (3636.910) metrics; lower values generally denote a more optimal fit. Our standard errors are notably low, exemplifying the stability of our coefficient estimates. We must acknowledge the Ljung-Box Q statistic (8.95) and the Jarque-Bera statistic (6379288.81), both resulting in p-values of 0.00, indicating potential autocorrelation in the residuals and non-normality, respectively. The kurtosis value at 139.17 suggests a heavy-tailed distribution for the residuals. With a convergence warning in mind, we managed to produce a precise 7-day GHI forecast. This model, despite the statistical cautions, offers us a valuable predictive tool for renewable energy analysis and informs our broader energy strategy.

4.4 Final Results

In our analysis, we employ two distinct predictive models that leverage similar datasets but are utilized in different contexts to maximize their utility. The first model, a regression model, predicts which counties could have higher solar capacity than currently exploited. This is visually represented in the first image, a color-coded map, which highlights the 15 counties with the largest discrepancy between their predicted solar capacity and their actual solar production. These counties, marked in varying shades, indicate significant untapped solar energy resources. This model helps identify regions where there is a substantial opportunity for increased solar investment and infrastructure development.

The second model, a Seasonal ARIMA (SARIMA) model, is utilized for making precise short-term predictions of Global Horizontal Irradiance (GHI) within specific areas. The second image showcases a detailed 7-day GHI forecast for the top 15 counties that have a high potential solar capacity. It is important to note that while the 7-day forecasts are for the upcoming week, each of the top 15 counties highlighted has different starting dates for their respective forecasts, reflecting the specific timing and data collection nuances for each region. This table provides valuable predictive insight for short-term solar power planning, with forecasts indicating variable solar irradiance levels over the week. This predictive data is instrumental for stakeholders in the solar energy sector, from policymakers to renewable energy developers, in making informed decisions to enhance solar energy production, aligning with sustainability goals, and supporting economic growth within these high-potential regions. This dual-layered approach of classification and prediction exemplifies the strategic use of data analytics to drive renewable energy advancements.

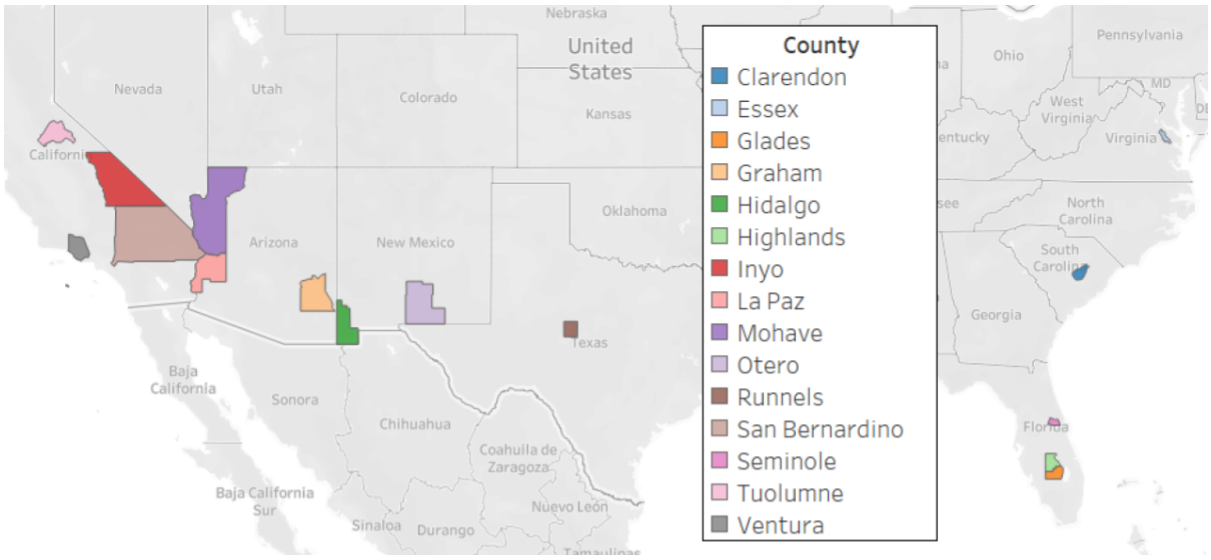
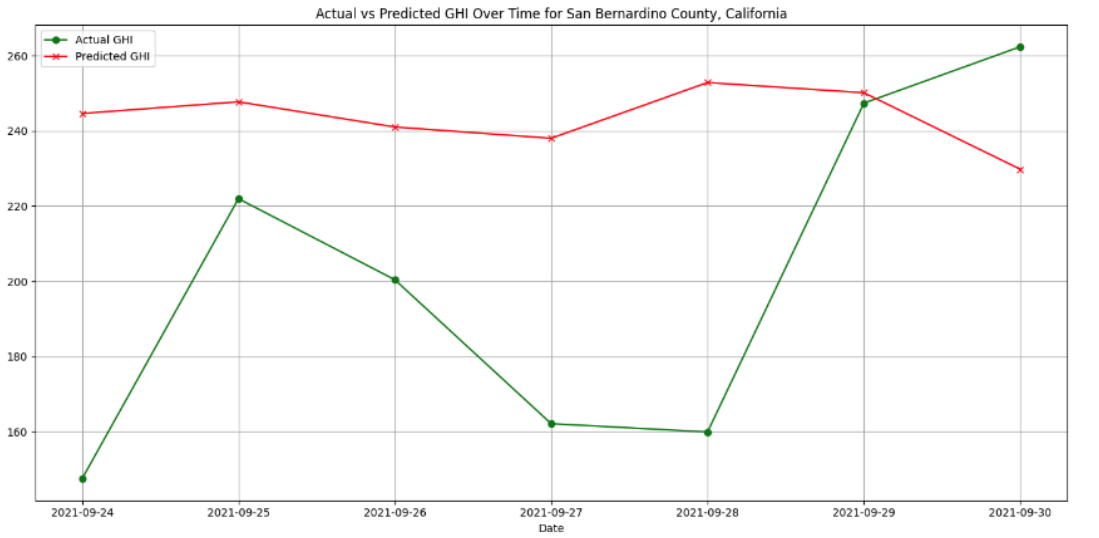


Figure 10: Map of Counties

Location	GHI	GHI	GHI	GHI	GHI	GHI	GHI
San Bernardino	241.69	264.46	251.93	241.17	255.14	280.21	272.26
Otero	375.21	377.77	374.09	366.50	374.76	363.42	365.77
Inyo	234.75	224.79	238.49	230.95	231.54	236.07	235.12
Hidalgo	185.47	167.95	172.73	164.15	151.65	148.88	150.51
Glades	248.79	239.67	241.70	244.89	241.06	244.20	244.01
Essex	221.14	183.67	187.71	227.20	218.38	209.53	210.95
Mohave	201.62	200.19	203.65	199.89	188.70	189.64	193.70
La Paz	281.61	288.06	279.49	274.03	285.40	281.56	271.79
Graham	149.78	145.89	152.29	155.25	154.03	144.99	143.33
Highlands	246.34	212.76	216.01	213.65	220.37	266.21	214.77
Runnels	357.31	352.05	348.59	351.51	357.63	356.64	357.09
Ventura	179.19	229.12	280.26	270.45	228.38	148.95	97.81
Clarendon	298.54	294.36	283.85	289.40	293.06	293.28	296.95
Tuolumne	289.67	287.14	294.21	289.56	291.71	295.54	300.15
Seminole	312.40	299.59	319.95	327.61	333.65	307.37	321.54

Table 1: Prediction of GHI over the next 7-days



	2021-09-24	2021-09-25	2021-09-26	2021-09-27	2021-09-28	2021-09-29	2021-09-30
GHI	244.62	247.72	241.03	238.02	252.85	250.13	229.79

Figure 11: Actual vs Predicted GHI over time for San Bernardino County, California

Taking a closer look at our 7-day forecast where we compare the actual GHI data to the predicted values for San Bernardino County. As we observe the graph, you'll notice that there are points where the predicted values diverge significantly from the actual measurements. Several factors contribute to these discrepancies. First, the variability in solar radiation can be influenced by unexpected weather conditions which aren't always perfectly captured by historical data. Secondly, despite our efforts in data preprocessing, the dataset had significant missing values which were interpolated. Although we used interpolation to fill in the missing values, it still introduced a level of estimation that may not accurately reflect true data points. The SARIMAX model we used still did well under these circumstances, capturing the general trend and seasonal patterns effectively. The red line represents our model's prediction and as you can see, it follows the actual GHI trend represented by the green line closely. This demonstrates the effectiveness of our model and gives us valuable insights into our energy forecasting. Improvements like integrating real-time weather data, adjusting for local geographic and environmental conditions, and employing more sophisticated interpolation techniques could help reduce these discrepancies in future forecasts.

To summarize, our comprehensive approach combining regression and SARIMA models showcases the power of predictive analytics in identifying and exploiting solar energy potential across various counties, particularly those like San Bernardino. By accurately forecasting Global Horizontal Irradiance and recognizing regions with untapped capacity, we are equipping stakeholders—from policymakers to developers—with crucial data to drive strategic decisions. These efforts not only promise to optimize energy production but also contribute significantly to sustainability and economic advancement. As we continue to refine our methods and integrate more dynamic data sources, we anticipate even more robust models that will further enhance the precision and utility of our forecasts, ultimately fostering a more sustainable and efficient energy landscape.

5 Discussion

This study employed a range of advanced machine learning models, including Random Forest, SVM, Gradient Boosting, and Decision Trees, to predict solar capacity at the county level across the United States. Our ensemble approach yielded an impressive R-squared score of 0.818, indicating strong model performance in pinpointing underutilized solar resources. These findings reinforce the capability of machine learning tools to significantly improve the management and strategic deployment of solar energy resources, echoing findings by Abdulai et al. (2023)[2] who observed similar benefits in diverse geographical settings.

The application of these machine learning models in solar capacity prediction not only pushes forward our scientific capabilities in energy forecasting but also sets a new benchmark for future research. This is particularly important for integrating fluctuating solar energy into national power grids more efficiently. Our methods could serve as a model for future studies that aim to exploit large datasets and advanced analytics in renewable energy and environmental science.

On an ethical level, the accurate prediction of solar energy capacity carries profound implications. By identifying areas with untapped solar potential, this research highlights opportunities to develop solar infrastructure in underserved communities, thereby promoting equitable energy access. This supports sustainable development goals and mitigates reliance on fossil fuels, aligning with our ethical duties towards environmental stewardship and equity across generations.

Echoing the comprehensive review by Gaboitaolelwe et al. (2023)[5], our study confirms that sophisticated predictive models can stabilize energy supply and enhance grid management, even across varied geographical landscapes. Their review underlines the importance of machine learning in improving solar PV power forecasting, a crucial aspect for the reliability of renewable energy systems. Like their findings, our research suggests that leveraging advanced machine learning techniques can address the intermittency challenges of solar energy and enhance the predictability of renewable energy sources.

As we look to the future, refining our modeling techniques and integrating more diverse datasets will be key. Collaborations across different disciplines will also be essential to tackle the multi-faceted challenges in renewable energy planning. By addressing these challenges proactively and leveraging the insights from current studies, we are well-placed to advance toward a more sustainable and equitable energy future.

6 Conclusions

In conclusion, this study underscores the critical importance of accurate solar capacity prediction in the realm of renewable energy planning. Through the application of ensemble regression models, including Random Forest, Support Vector Machines (SVM), Gradient Boosting, and Decision Tree, we have demonstrated the feasibility of predicting solar capacity at the county level. Despite encountering challenges such as data limitations and model complexity, our analysis has provided valuable insights into the distribution and utilization of solar resources across different regions.

The implications of our findings extend beyond scientific curiosity to encompass economic, environmental, and ethical considerations. Economically, accurate solar capacity predictions can inform investment decisions, leading to cost savings and enhanced competitiveness in the renewable energy sector. Environmentally, the transition towards renewable energy sources facilitated by precise capacity forecasting contributes significantly to mitigating climate change and reducing reliance on fossil fuels. Ethically, ensuring equitable access to clean energy technologies is paramount, particularly for marginalized communities disproportionately affected by environmental injustices.

Looking ahead, future research directions include refining modeling techniques, integrating additional datasets, and fostering interdisciplinary collaborations to address complex challenges in renewable energy planning. By tackling these challenges head-on and leveraging the insights gained from this study, we can advance towards a more sustainable and equitable energy future.

References

[1] Green power markets: Policies and regulations, Dec 2023.

[2] Dampaak Abdulai, Samuel Gyamfi, Felix Amankwah Diawuo, and Peter Acheampong. Data analytics for prediction of solar pv power generation and system performance: A real case of bui solar generating station, ghana, Sep 2023.

[3] National Solar Research Database. Data viewer.

[4] David Feldman and Robert Margolis. Solar industry update: (h2 2020) [slides]. 4 2021.

[5] Jwaone Gaboitaolelwe, Adamu Murtala Zungeru, Abid Yahya, Caspar K. Lebekwe, Dasari Naga Vinod, and Ayodeji Olalekan Salau. Machine learning based solar photovoltaic power forecasting: A review and comparison. *IEEE Access*, 11:40820–40845, 2023.

[6] M. M. Hasan, Shakhawat Hossain, M. Mofijur, Zobaidul Kabir, Irfan Anjum Badruddin, T. M. Yunus Khan, and Esam Jassim. Harnessing solar power: A review of photovoltaic innovations, solar thermal systems, and the dawn of energy storage solutions. *Energies*, 16(18):6456, Jan 2023.

[7] Lenssen, Schmidt, Hansen, Menne, Persin, Ruedy, and Zyss. Improvements in the gistemp uncertainty model. *J. Geophys. Res. Atmos*, 2019.

[8] Rishi Sankhe. Solar energy prediction and forecasting, Oct 2023.

[9] Seel, Joachim, Bolinger, Mark, Warner, Cody, Robson, and Dana. Utility-scale solar, 2022 edition: Analysis of empirical plant-level data from u.s. ground-mounted pv, pv+battery, and csp plants (exceeding 5 mwac). 09 2022.

[10] GISTEMP Team. Giss surface temperature analysis (gistemp), version 4. *NASA Goddard Institute for Space Studies*, 2024.

7 Code Appendix

GitHub Repository