

Solar Power Generation Forecasting

Brayden Mi
New York University
Courant Institute of Mathematical Sciences

Contents

1 Summary **4**

2 Data Selection **6**
2.1 Solar Generation Time Series 6
2.2 Exogenous Features 7

3 Simple Regression Modeling **10**
3.1 Sinusoidal Exponential Midline Regression 10
3.2 Sinusoidal Logistic Midline Regression 11
3.3 Limitations of Simplistic Regression Models 12

4 SARIMAX Modeling **14**
4.1 Parameter Selection for SARIMAX Model 14
4.2 Model Preliminary 16
4.3 Model Training and Prediction Setup 17
4.3.1 Data Preparation 17
4.3.2 Model Configuration 17
4.4 Results and Discussion 18
4.4.1 Model Summary: Utility-Scale 18
4.4.2 Parameter Estimates 18
4.4.3 Model Diagnostics 19
4.4.4 Discussion 19
4.4.5 Model Summary: Small-Scale 20
4.4.6 Parameter Estimates 20
4.4.7 Model Diagnostics 21
4.4.8 Discussion 21

5 Artificial Neural Network (ANN) Modeling **23**
5.1 Model Inputs and Outputs 23
5.2 Model Architecture 23
5.3 Training and Evaluation 23
5.4 Performance Metrics 24
5.5 Attempts to Improve Model Performance 25
5.6 Discussion and Justification for Exclusion 26

6 Time-Lag Neural Network Modeling **27**
6.1 Initial Performance 27
6.2 Issues with Full Dataset Prediction 28

7 Seasonal Artificial Neural Network (SANN) Modeling **29**
7.1 Preliminary Testing Results 29
7.2 Results 29
7.3 Discussion and Observations 30
7.4 Conclusion 31

8 Discussion and Conclusion **32**
8.1 Comparison of Model Performance 32
8.2 Model Recommendation 33

9 Assumptions, Limitations, and Challenges **34**
9.1 Assumptions 34
9.2 Limitations 34
9.3 Challenges 34
9.4 Future Steps 34

1 SUMMARY

Forecasting solar generation is a complex task that requires accurate modeling techniques to capture both seasonal and time-dependent patterns. In this study, solar generation data was collected from the U.S. Energy Information Administration (EIA) for the period of January 2022 to October 2024. This dataset was supplemented with a derived growth rate based on a regression analysis of solar energy capacity and monthly Total Solar Irradiance (TSI) data. Multiple models were evaluated, including sinusoidal regression approaches, traditional statistical methods, and neural networks, to determine their efficacy in predicting solar generation.

The Seasonal Artificial Neural Network (SANN) emerged as the most accurate model, with a normalized root mean square error (NRMSE) of 0.0499. The SANN demonstrated exceptional performance in capturing the 12-month seasonality inherent in solar generation data while maintaining computational efficiency and robustness across training sessions, making it the optimal choice for high-precision forecasting tasks.

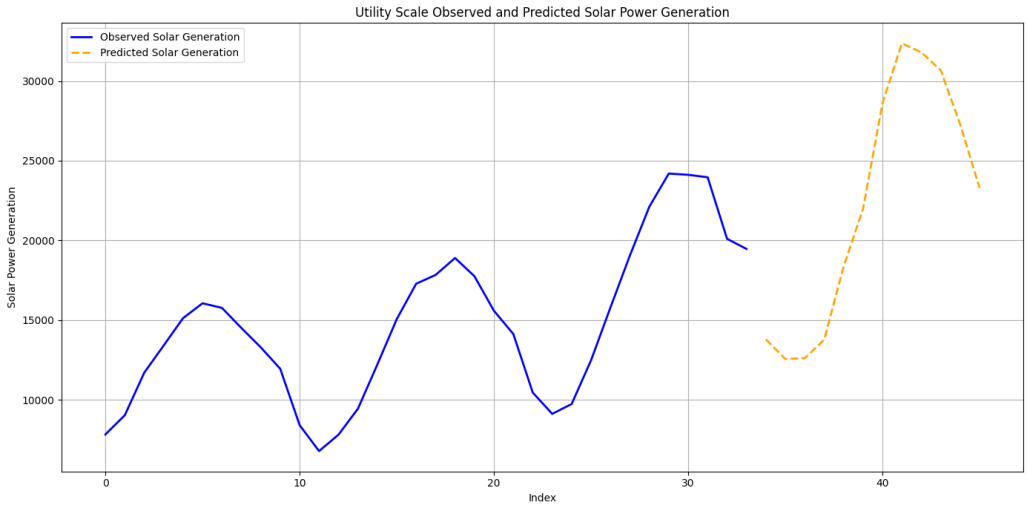


FIGURE 1: UTILITY SCALE SANN PREDICTION

For applications where simplicity and computational efficiency are required, the Sinusoidal Exponential Midline Regression model presented a viable alternative. With an NRMSE of 0.0395, this model provided high accuracy while maintaining a straightforward implementation. However, its inability to incorporate exogenous variables limits its use to simpler forecasting scenarios.

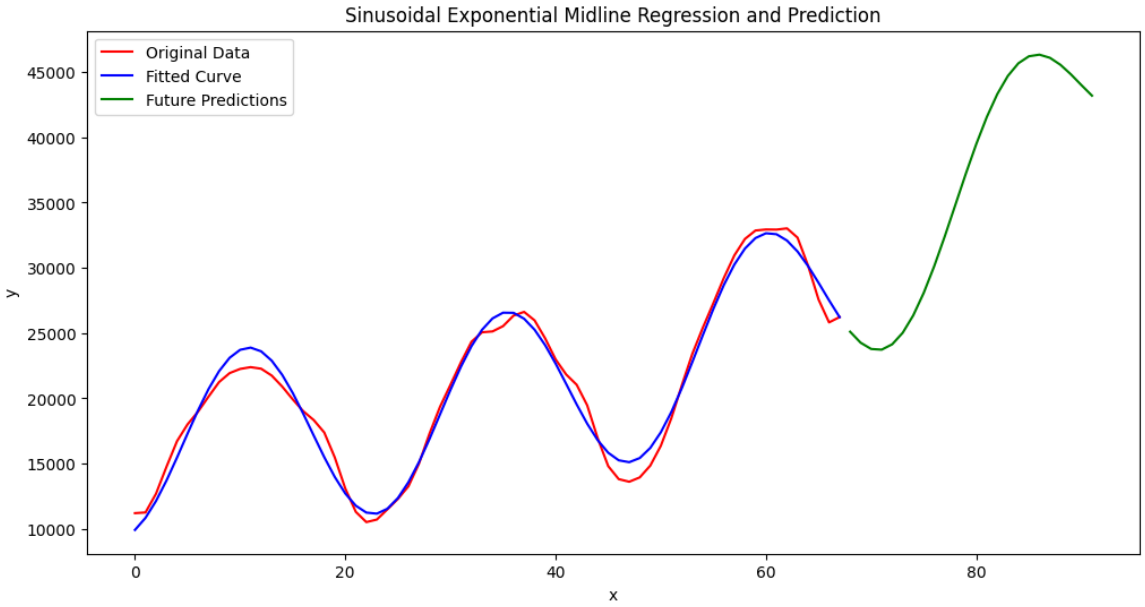


FIGURE 2: *SINUSOIDAL EXPONENTIAL MIDLINE REGRESSION*

The performance of all models is summarized in the table below:

TABLE 1: PERFORMANCE METRICS OF DIFFERENT MODELS

Model	RMSE	NRMSE	MAE
Sinusoidal Exponential Midline Regression	889.7744	0.0395	748.6832
Sinusoidal Logistic Regression	3354.0380	0.1490	2628.7147
SARIMAX	1021.9122	0.1989	880.1877
Artificial Neural Network	3294.3956	0.1900	2828.0303
Time Lagged Neural Network	1360.2551	0.1209	1141.2519
Seasonal Artificial Neural Network	817.1864	0.0499	637.7981

2 DATA SELECTION

2.1 SOLAR GENERATION TIME SERIES

The primary time series analyzed in this study represents monthly solar energy generation in the United States from January 2022 to October 2024, obtained from the Energy Information Administration (EIA). This dataset provides a detailed view of the solar energy production trends over a nearly three-year period, with measurements reflecting the aggregated contributions of solar installations across both small and utility scale.

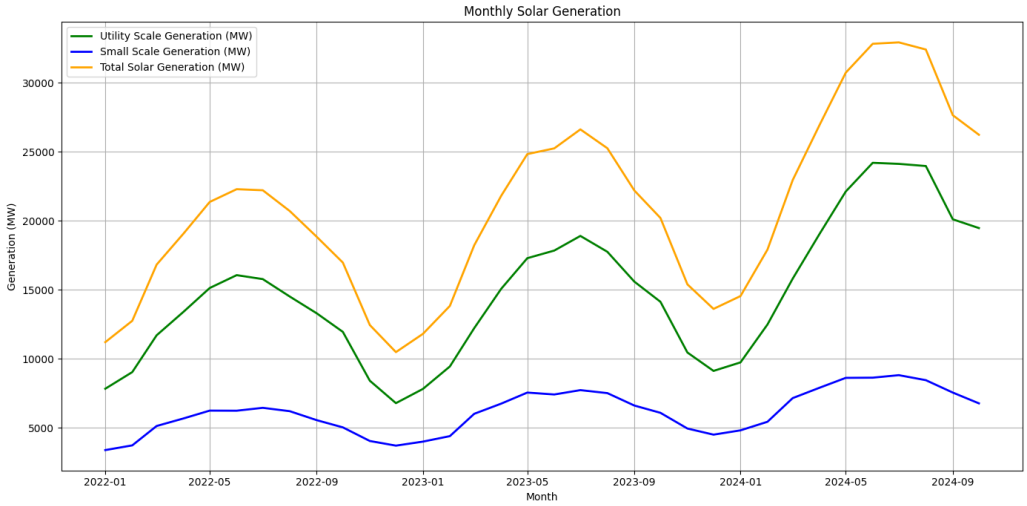


FIGURE 3: *MONTHLY SOLAR GENERATION (UNITED STATES)*

A prominent characteristic of the dataset is its clear sinusoidal pattern, indicative of strong seasonal variability. Solar energy generation peaks during the summer months due to increased solar irradiance and longer daylight hours, whereas the winter months exhibit lower generation levels as a result of reduced sunlight availability. This repeating pattern aligns with the annual cycle of solar irradiance, driven by the Earth's tilt and orbital motion.

In addition to seasonal fluctuations, the dataset exhibits a consistent upward trend, indicative of some sort of continued growth of solar energy installation and increasing adoption of solar technologies. The combination of seasonality and long-term growth trends makes this time series particularly suitable for advanced forecasting models like SARIMAX, which can capture both periodic and nonperiodic components in the data.

Statistical analysis of the time series reveals the presence of positive autocorrelations at seasonal lags, further confirming the significance of seasonality in solar generation patterns. The growth dynamics are evident in the upward shift of the mean and amplitude of the sinusoidal cycles over time, underscoring the compounded effect of capacity expansion and in-

creased efficiency of solar technologies.

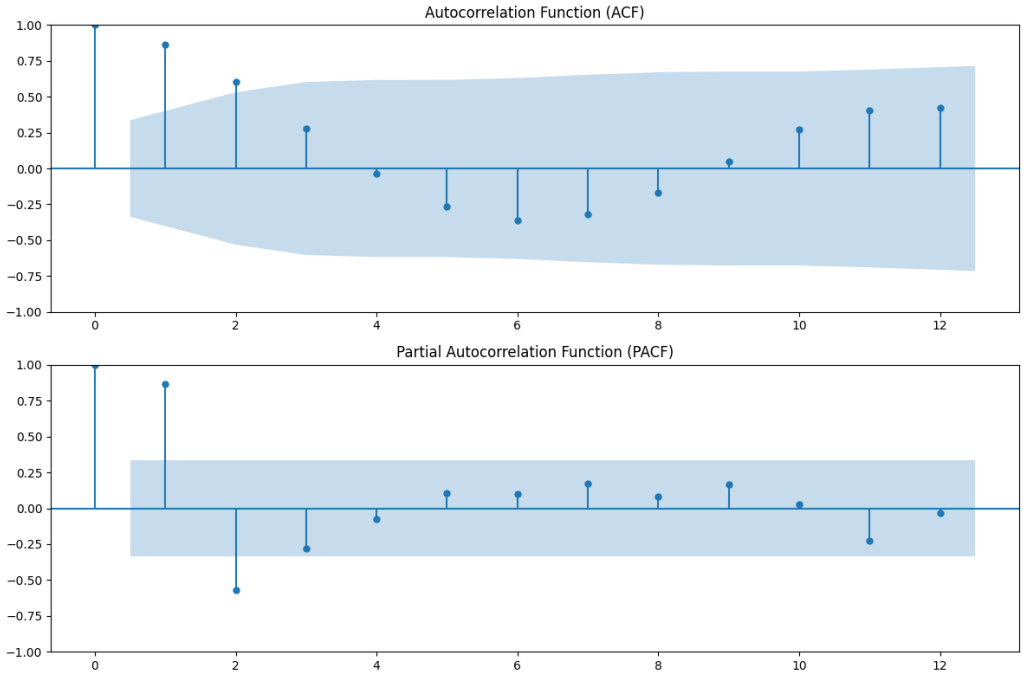


FIGURE 4: *ACF AND PACF OF TOTAL SOLAR GENERATION*

2.2 EXOGENOUS FEATURES

To supplement the solar generation data, two main exogenous features were chosen:

SOLAR CAPACITY GROWTH RATE:

The solar capacity growth rate is a critical indicator of the expansion and scalability of solar energy generation infrastructure. It reflects the compounded growth of installed solar energy capacity over time, driven by technological advancements, declining costs of solar technology, and increased funding and adoption across various sectors.

For this study, the solar capacity data was taken from the Berkeley Lab's comprehensive report, titled *Utility Scale Solar*. The dataset provided gives insights into not only utility-scale installations, complemented by information on residential and commercial solar production. To achieve a holistic view of total solar capacity across small-scale and utility-scale solar, the residential and commercial solar production data were aggregated into the small-scale solar category.

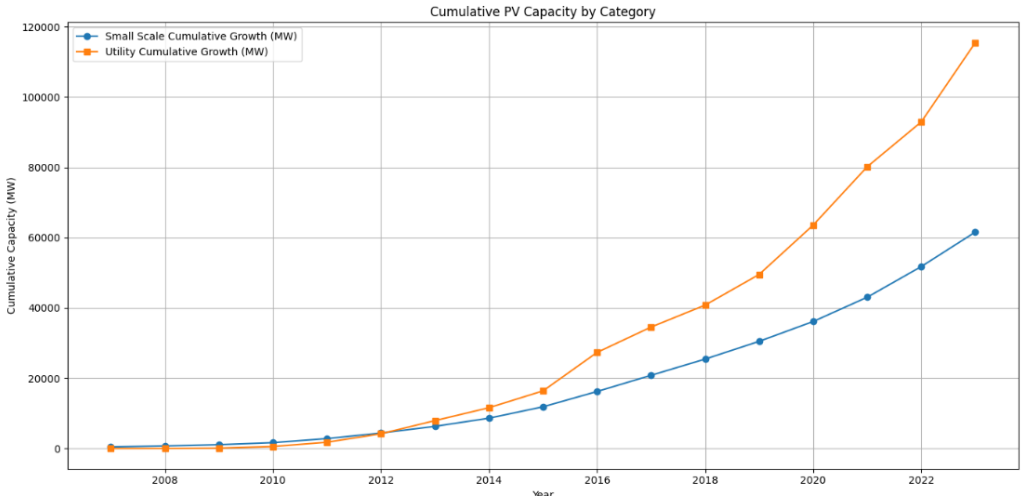


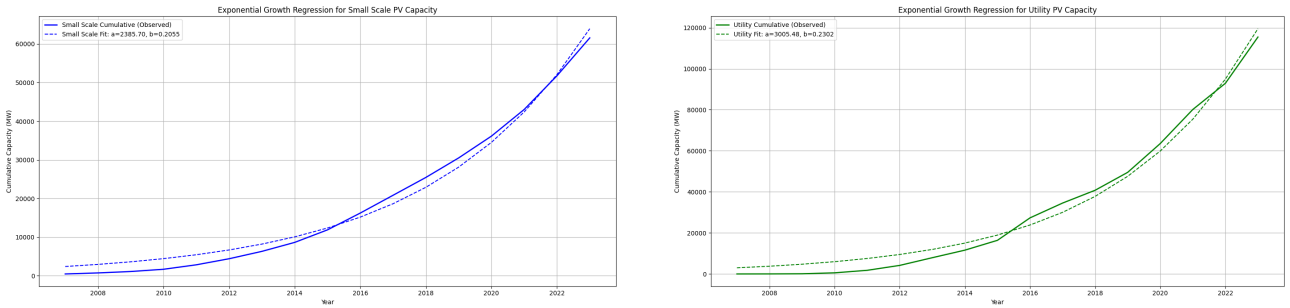
FIGURE 5: *SOLAR PHOTOVOLTAIC CAPACITY GROWTH BY SCALE*

Due to the data's exponential nature, a regression analysis was conducted to determine the respective growth rates of small-scale and utility-scale solar capacities. Exponential regression was chosen to accurately capture the compounding growth trends observed in the solar energy sector. The regression equation used is defined as:

$$C(t) = C_0 \cdot e^{rt}$$

where:

- $C(t)$: Solar capacity at time t,
- C_0: Initial solar capacity,
- r: Growth rate (constant to be estimated),
- t: Time in years.



(a) SMALL SCALE ($R^2 = 0.9889$) **(b)** UTILITY SCALE ($R^2 = 0.9873$)

FIGURE 6: *SMALL SCALE AND UTILITY SCALE REGRESSIONS*

The exponential regression analysis conducted on small-scale and utility-scale solar capacity data yielded high coefficients of determination (R^2), indicating strong fits for both mod-

els. The R^2 values for small-scale and utility-scale regressions were 0.9889 and 0.9873, respectively, demonstrating that the exponential model accurately captures the growth trends in both categories. From this, the growth rate was extracted with a rate of 0.2055 for small-scale and 0.2302 for utility scale.

In terms of significance, the solar capacity growth rate provides a quantitative measure of the rapid expansion in solar infrastructure, driven by the increasing adoption in the residential, commercial, and utility sectors. By incorporating these growth rates into forecasting models, the study captures the compounded effects of technological advancements, policy incentives, and market dynamics, allowing accurate predictions of future expansion of solar capacity and energy generation potential.

TOTAL SOLAR IRRADIANCE (TSI):

Total Solar Irradiance (TSI) is a fundamental variable in understanding the variability in solar energy availability. TSI refers to the solar power per unit area received at the Earth's upper atmosphere and is measured in Watts per square meter (W/m^2). It accounts for fluctuations in solar radiation caused by factors such as solar cycles, sunspots, and long-term trends in solar output.

Variability in TSI impacts the amount of solar radiation available for energy conversion. These variations are closely linked to the approximately 11-year solar cycle, characterized by changes in the number of sunspots and overall solar activity. While TSI generally shows minor annual variations (on the order of approx. 0.1%) [1], over decades, changes in solar irradiance due to shifts in solar dynamics can lead to significant effects on energy generation potential.

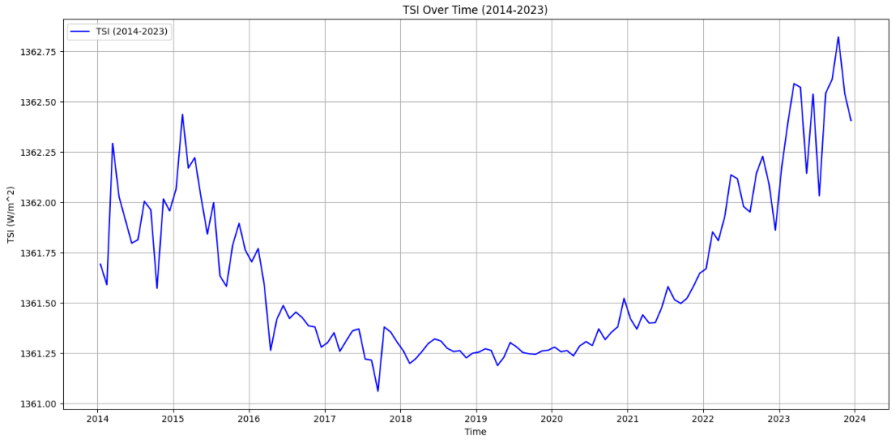


FIGURE 7: *VARIATIONS IN TOTAL SOLAR IRRADIANCE (TSI)*

[1] Willson, Hudson. (1991). *The Sun's luminosity over a complete solar cycle.*

3 SIMPLE REGRESSION MODELING

3.1 SINUSOIDAL EXPONENTIAL MIDLINE REGRESSION

To explore the seasonal and growth dynamics in the solar generation time series, a regression model was implemented using a sinusoidal function combined with an exponential midline. The total between small-scale and utility-scale generation was used for the sake of testing. The mathematical form of the model was defined as:

$$y_t = A \cdot \sin(B \cdot t + C) + D \cdot e^{E \cdot t} + F$$

where:

- y_t: Solar generation at time t,
- A: Amplitude of the sinusoidal component, representing seasonal variation,
- B: Frequency of the sinusoidal component, with periodicity corresponding to a yearly cycle,
- C: Phase shift, aligning the sinusoidal curve with observed seasonality,
- D: Initial scaling coefficient for the exponential midline,
- E: Exponential growth rate, capturing long-term growth in solar generation,
- F: Constant offset, representing the baseline level of solar generation.

The regression was performed using non-linear least squares optimization. The results demonstrated an excellent fit, with a coefficient of determination (R^2) of 0.9803. This high R^2 value indicated that the model effectively captured both the periodic nature of the solar generation data and the underlying growth trend. The fitted sinusoidal curve closely matched the observed data points, particularly during peak and trough periods, further validating the model's ability to represent the seasonal dynamics.

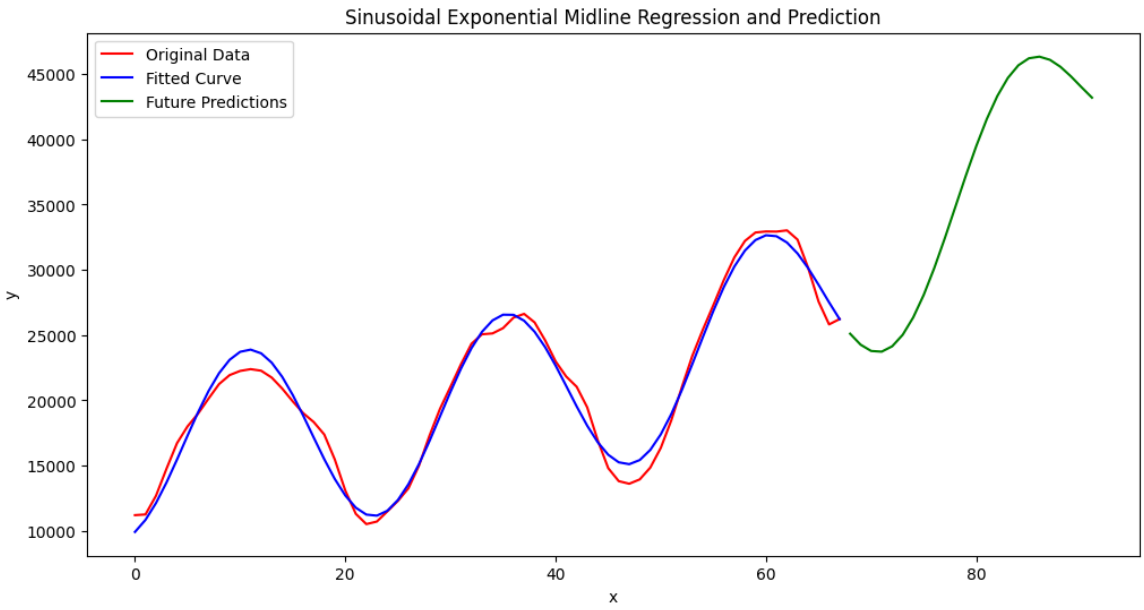


FIGURE 8: *SINUSOIDAL EXPONENTIAL MIDLINE REGRESSION*

3.2 SINUSOIDAL LOGISTIC MIDLINE REGRESSION

To explore alternative approaches for capturing the seasonal and growth dynamics in solar generation data, a sinusoidal model with a logistic midline was implemented. This model aimed to account for potential saturation in solar energy production due to limitations such as physical infrastructure, technological constraints, or market saturation. The chosen functional form was:

$$y_t = A \cdot \sin(B \cdot t + C) + \frac{D}{1 + \exp(-E \cdot (t - F))} + G$$

where:

- A: Amplitude of the sinusoidal component, representing seasonal variation,
- B: Frequency of the sinusoidal component, corresponding to annual periodicity,
- C: Phase shift aligning the sinusoidal curve with observed seasonality,
- D: Scaling factor for the logistic midline, representing the potential maximum capacity,
- E: Growth rate of the logistic curve,
- F: Midpoint of the logistic curve, marking the inflection point,
- G: Constant offset, representing baseline solar generation.

The logistic midline was chosen to model a scenario where the growth in solar generation approaches an upper limit, such as a potential maximum capacity of the solar industry. This choice reflects the theoretical constraints that might arise due to saturation effects, such as

limited available land for installations, technological plateaus, or market equilibrium.

However, when fitted to the solar generation dataset using non-linear least squares optimization, the sinusoidal logistic model performed poorly compared to the sinusoidal exponential model. The coefficient of determination (R^2) was 0.7197, indicating a suboptimal fit. While the sinusoidal component adequately captured the seasonal patterns early on, the logistic midline stretched the sinusoidal graph, which caused it to lose seasonality.

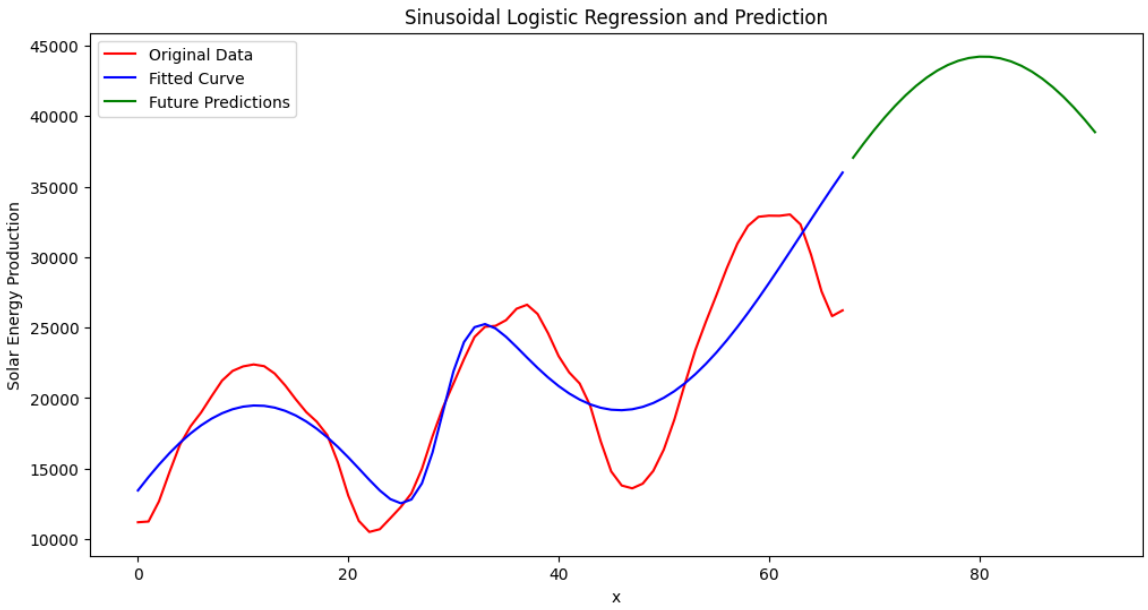


FIGURE 9: *SINUSOIDAL LOGISTIC MIDLINE REGRESSION*

The poorer fit is most likely due to the mismatch between the logistic midline's assumption of approaching an upper limit and the observed data's continued upward trajectory. The inflection point and asymptotic behavior of the logistic curve did not align with the nearly linear exponential growth in solar capacity, leading to significant deviations between the model and the actual data.

While the sinusoidal logistic model was a valuable experiment to explore the potential effects of saturation, its limitations in capturing the current growth trends reinforce the suitability of models like the sinusoidal exponential midline for this dataset.

3.3 LIMITATIONS OF SIMPLISTIC REGRESSION MODELS

Despite the promising performance of the sinusoidal model with an exponential midline and the exploratory attempt with the logistic midline, neither model was ultimately pursued for further analysis. Although these models provided initial insights into the seasonal and growth dynamics of the solar generation data, their underlying structures were deemed too simplistic

to fully capture the complexities of the system.

Both models also lacked the capacity to incorporate external drivers of solar generation, such as Total Solar Irradiance (TSI) and solar capacity growth rates, which are critical for understanding and forecasting solar energy trends. These external factors introduce variability that cannot be captured by models that rely solely on intrinsic data patterns.

Ultimately, the decision to not proceed with either model was driven by the need for a more sophisticated approach capable of handling the interaction between seasonal patterns, long-term trends, and external influences. This led to the adoption of the SARIMAX model, which integrates exogenous variables and offers greater flexibility in modeling the underlying dynamics of solar energy generation. Moving beyond these initial simplistic models, the analysis aims to provide a more comprehensive and actionable framework for forecasting solar power trends.

TABLE 2: PERFORMANCE METRICS OF REGRESSIONS

Model	RMSE	NRMSE	MAE
Sinusoidal Exponential Midline Regression	889.7744	0.0395	748.6832
Sinusoidal Logistic Regression	3354.0380	0.1490	2628.7147

4 SARIMAX MODELING

4.1 PARAMETER SELECTION FOR SARIMAX MODEL

Selecting appropriate parameters (p, d, q) is a critical step in constructing an effective SARIMAX model. Due to both the small-scale generation and utility-scale generation exhibiting the same upwards trending sinusoidal behavior, the total generation was analyzed to select parameters for both models.

DETERMINING P AND Q USING ACF AND PACF

The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are key tools in selecting the autoregressive (p) and moving average (q) orders in the SARIMAX model.

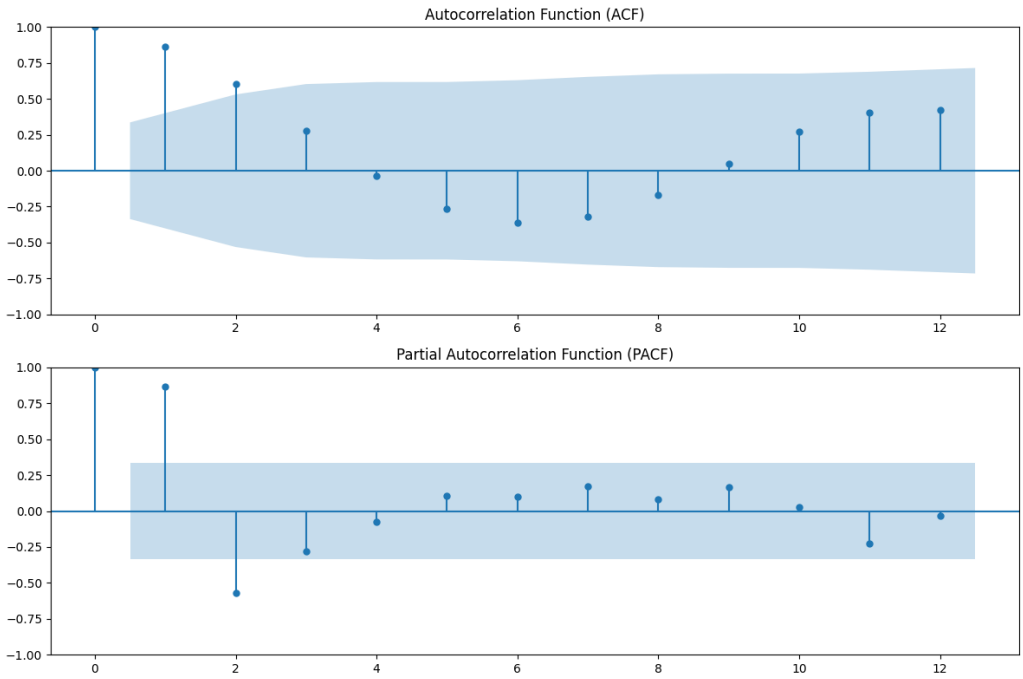


FIGURE 10: ACF AND PACF PLOT

Autoregressive Order (p):

The PACF plot highlights the partial correlations of the time series with its own lagged values. Significant spikes in the PACF suggest potential autoregressive terms. Here, the first lag in the PACF is significantly out of the confidence limits, and the second lag is also slightly outside the limits, albeit to a lesser extent. Based on this observation, we select $p = 1$.

Moving Average Order (q):

The ACF plot, which captures correlations between the series and its lagged values, is analyzed

to identify moving average terms. A sharp drop-off in the ACF suggests the order of the moving average process. In this case, the ACF shows no clear sharp drop-offs, but smaller q values are plausible. Based on these observations, $q = 1$ is selected.

DETERMINING D USING THE AUGMENTED DICKEY-FULLER TEST AND DIFFERENCING

The augmented Dickey-Fuller (ADF) test is employed to determine whether the time series is stationary. The null hypothesis of the ADF test is that the time series is non-stationary. If the p-value exceeds 0.05, the null hypothesis cannot be rejected, indicating non-stationarity.

```
ADF Statistic: 3.2519566477447523
p-value: 1.0
Critical Values:
   1%: -3.7377092158564813
   5%: -2.9922162731481485
   10%: -2.635746736111111
```

The ADF test applied to the original time series yields a p-value significantly greater than 0.05, confirming that the data is not stationary.

Differencing:

To achieve stationarity, we apply first-order differencing ($d = 1$). The ADF test for the first-order differenced series is:

```
ADF Statistic: -3.4136934311896043
p-value: 0.01050082735150278
Critical Values:
   1%: -3.7377092158564813
   5%: -2.9922162731481485
   10%: -2.635746736111111
```

After differencing, the ADF test produces a p-value below 0.05, indicating the series is now stationary. Thus, $d = 1$ is selected.

SUMMARY OF PARAMETERS

Based on the analyses:

- Autoregressive order (p): 1 (based on PACF plot),
- Differencing order (d): 1 (based on ADF test and differencing),
- Moving average order (q): 1 (based on ACF plot).

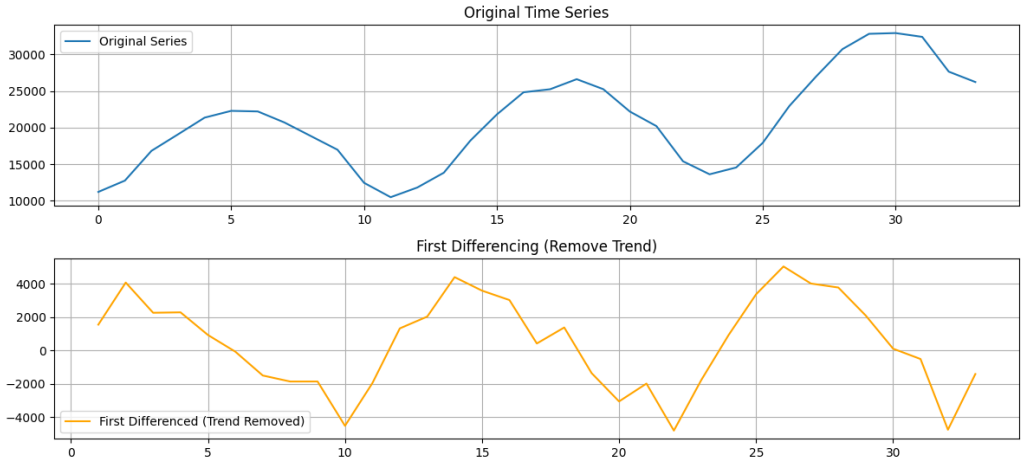


FIGURE 11: TIME SERIES AFTER FIRST-ORDER DIFFERENCING ($D = 1$)

4.2 MODEL PRELIMINARY

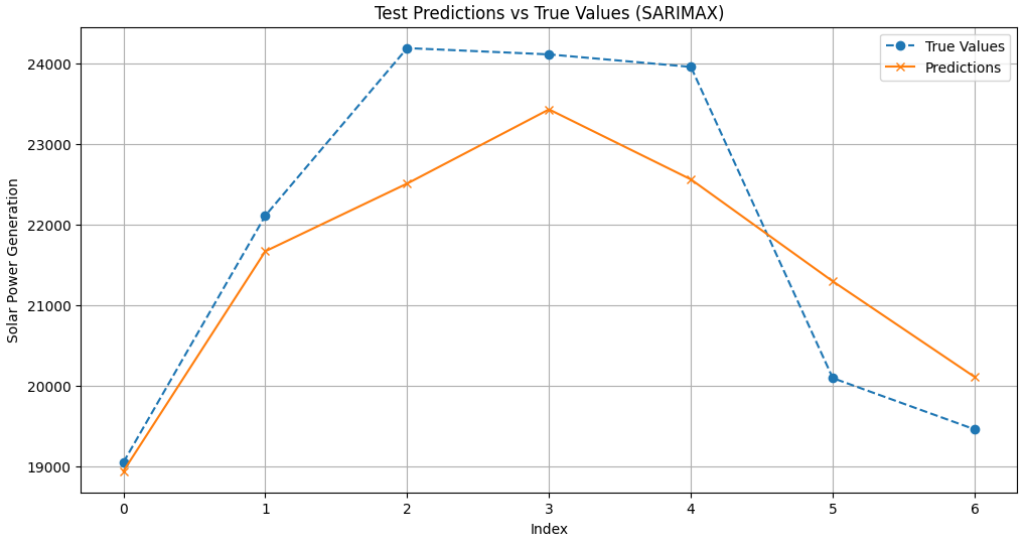


FIGURE 12: PRELIMINARY ON THE SARIMAX MODEL

Preliminary testing of the SARIMAX model was conducted using an 80/20 train-test split to evaluate its performance on solar generation data. The model exhibited consistent and reasonably accurate behavior, suggesting its potential for capturing seasonal trends in the data. The key performance metrics obtained during testing were as follows: root mean square error (RMSE) of 1021.9122, normalized root mean square error (NRMSE) of 0.1989, and mean absolute error (MAE) of 880.1877.

These results indicate that SARIMAX is capable of handling the seasonality inherent in solar generation data while providing interpretable outputs. Thus, work on the model was continued to predict on the full dataset.

4.3 MODEL TRAINING AND PREDICTION SETUP

The training and prediction process was implemented using the `SARIMAX` model from the `statsmodels` library in Python.

4.3.1 DATA PREPARATION

To incorporate both the target variable and exogenous features, the data was prepared by adding the TSI data and growth factor data into a features vector (pandas dataframe), which aligned with the indices of the main monthly solar generation time series.

4.3.2 MODEL CONFIGURATION

The `SARIMAX` model was initialized with the following parameters:

- **Endogenous Variable:** Monthly time series of either small-scale or utility-scale data.
- **Exogenous Variables:** The prepared dataframe of features (TSI and growth rate of respective scale).
- **Order:** $(1, 1, 1)$, specifying the non-seasonal autoregressive order (p), differencing order (d), and moving average order (q).
- **Seasonal Order:** $(1, 1, 1, 12)$, indicating the seasonal autoregressive order (P), seasonal differencing order (D), seasonal moving average order (Q), and seasonal period ($s = 12$).

4.4 RESULTS AND DISCUSSION

4.4.1 MODEL SUMMARY: UTILITY-SCALE

The model results are summarized in Table 3.

TABLE 3: SARIMAX MODEL RESULTS FOR UTILITY-SCALE GENERATION

Metric	Value
Dependent Variable	Utility Scale Generation
Number of Observations	34
Log Likelihood	-171.266
Akaike Information Criterion (AIC)	356.532
Bayesian Information Criterion (BIC)	363.844
Hannan-Quinn Information Criterion (HQIC)	358.119

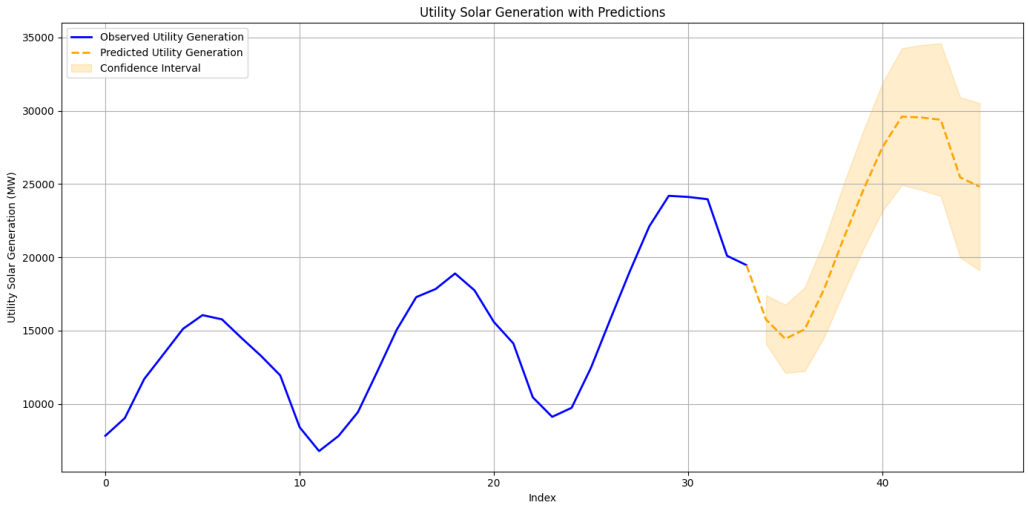


FIGURE 13: SARIMAX PREDICTION ON UTILITY SCALE SOLAR

4.4.2 PARAMETER ESTIMATES

Table 4 displays the parameter estimates for the model. Key observations include:

- The exogenous variable coefficients for TSI (-352.9116) and Growth Factor (2.669×10^{-8}) were statistically insignificant ($p > 0.05$).
- Autoregressive (ar.L1) and moving average (ma.L1) terms also lacked statistical significance ($p > 0.05$).
- The seasonal components (ar.S.L12 and ma.S.L12) showed no significant effect ($p > 0.05$).

- The error variance (σ^2) was estimated at 7.092×10^5, with statistical significance ($p = 0.033$).

TABLE 4: PARAMETER ESTIMATES FOR SARIMAX MODEL

Parameter	Estimate	Std. Error	z-value	p-value
TSI	-352.9116	745.989	-0.473	0.636
Growth Factor	2.669×10^{-8}	10.338	2.58×10^{-9}	1.000
ar.L1	-0.4011	5.091	-0.079	0.937
ma.L1	0.3985	5.146	0.077	0.938
ar.S.L12	0.0013	183.000	6.94×10^{-6}	1.000
ma.S.L12	0.0011	182.894	6.21×10^{-6}	1.000
σ^2	7.092×10^5	3.32×10^5	2.136	0.033

4.4.3 MODEL DIAGNOSTICS

Diagnostics were performed to assess model fit and residual behavior:

- **Ljung-Box Test:** The Ljung-Box statistic for lag 1 was 2.75 ($p = 0.10$), indicating no significant autocorrelation in the residuals.
- **Jarque-Bera Test:** The Jarque-Bera statistic was 2.46 ($p = 0.29$), suggesting that the residuals follow a normal distribution.
- **Heteroskedasticity:** The heteroskedasticity test statistic was 2.89 ($p = 0.18$), showing no evidence of significant heteroskedasticity.

4.4.4 DISCUSSION

While the model successfully captured seasonality and generated forecasts, the results highlight several challenges:

- The exogenous variables, TSI and Growth Factor, did not exhibit significant influence on utility-scale generation. This may indicate a weak relationship or the need for additional predictors.
- The insignificance of the autoregressive and moving average terms suggests that simpler models may be sufficient for this dataset.
- The statistically significant variance parameter (σ^2) indicates the presence of inherent variability in the generation process, possibly due to unobserved factors.

Overall, the model diagnostics suggest that the SARIMAX model provides a reasonable fit, with residuals meeting key assumptions. However, further refinement or alternative modeling approaches may be needed to improve predictive performance.

4.4.5 MODEL SUMMARY: SMALL-SCALE

The SARIMAX model was configured with the same parameters as the utility-scale model.

TABLE 5: SARIMAX MODEL RESULTS FOR SMALL-SCALE GENERATION

Metric	Value
Dependent Variable	Small Scale Generation
Number of Observations	34
Log Likelihood	-131.824
Akaike Information Criterion (AIC)	277.648
Bayesian Information Criterion (BIC)	284.960
Hannan-Quinn Information Criterion (HQIC)	279.235

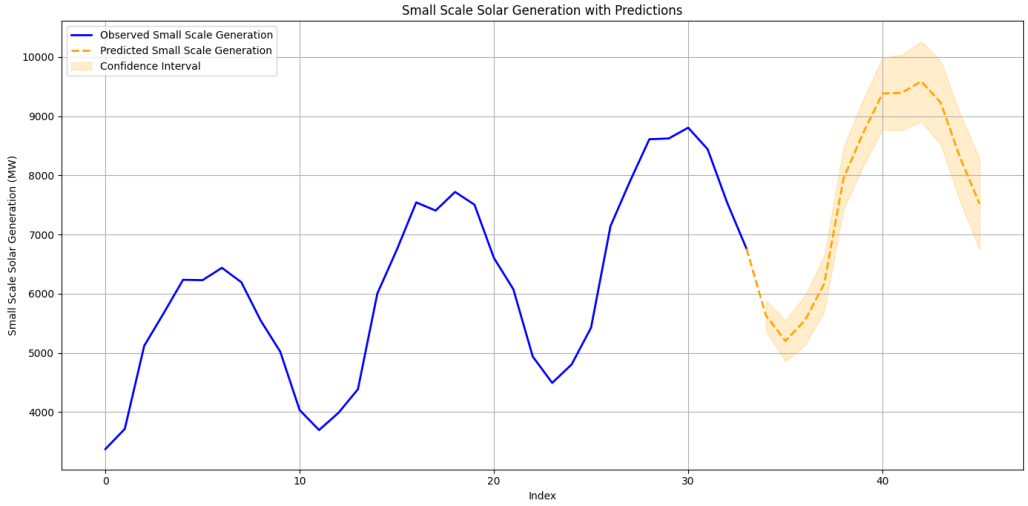


FIGURE 14: SARIMAX PREDICTION ON SMALL-SCALE SOLAR

4.4.6 PARAMETER ESTIMATES

Table 6 summarizes the parameter estimates. Key observations include:

- The coefficient for TSI (-177.1171) was not statistically significant ($p = 0.167$).
- The Growth Factor coefficient (-1.669×10^{-7}) was also statistically insignificant ($p = 1.000$).
- The autoregressive term (ar.L1) was significant ($p = 0.021$), with an estimate of 0.6989.

- The moving average term (ma.L1) was significant ($p = 0.008$), with an estimate of -0.7462.
- Seasonal parameters (ar.S.L12 and ma.S.L12) and the error variance (σ^2) were not statistically significant.

TABLE 6: PARAMETER ESTIMATES FOR SARIMAX MODEL (SMALL-SCALE GENERATION)

Parameter	Estimate	Std. Error	z-value	p-value
TSI	-177.1171	128.051	-1.383	0.167
Growth Factor	-1.669×10^{-7}	1.251	-1.33×10^{-7}	1.000
ar.L1	0.6989	0.302	2.316	0.021
ma.L1	-0.7462	0.281	-2.656	0.008
ar.S.L12	0.0060	111.467	5.38×10^{-5}	1.000
ma.S.L12	-0.0148	111.590	-0.000	1.000
σ^2	1.631×10^4	1.02×10^4	1.593	0.111

4.4.7 MODEL DIAGNOSTICS

Diagnostics were conducted to evaluate the model's residuals and goodness of fit:

- **Ljung-Box Test:** The Ljung-Box statistic for lag 1 was 0.87 ($p = 0.35$), indicating no significant autocorrelation in the residuals.
- **Jarque-Bera Test:** The Jarque-Bera statistic was 0.97 ($p = 0.62$), suggesting that the residuals are approximately normally distributed.
- **Heteroskedasticity:** The heteroskedasticity test statistic was 1.87 ($p = 0.43$), showing no evidence of significant heteroskedasticity.

4.4.8 DISCUSSION

The SARIMAX model for small-scale generation demonstrates some key differences from the utility-scale model:

- Unlike the utility-scale results, the autoregressive (ar.L1) and moving average (ma.L1) terms were statistically significant, suggesting that short-term dependencies are important for small-scale generation.
- Similar to the utility-scale model, TSI and Growth Factor were not significant predictors, indicating that these variables may not strongly influence small-scale generation.
- The seasonal components (ar.S.L12 and ma.S.L12) did not show significance, suggesting that seasonal patterns are less pronounced or not well-captured in this dataset.



Overall, the model provides a reasonable fit, with residuals meeting diagnostic criteria. However, the lack of significant exogenous predictors suggests that other variables or model refinements may improve performance. Future research could explore additional factors influencing small-scale generation.



5 ARTIFICIAL NEURAL NETWORK (ANN) MODELING

5.1 MODEL INPUTS AND OUTPUTS

The input features (X), similar to the SARIMAX model, consisted of the Total Solar Irradiance time series and the exponential growth rate of the solar capacity.

The target variable (y) was the main time series to be predicted, either utility-scale or small-scale generation.

5.2 MODEL ARCHITECTURE

The ANN consisted of three fully connected (dense) layers. The architecture and parameters of the model are summarized in Table 7.

TABLE 7: ARCHITECTURE OF THE ARTIFICIAL NEURAL NETWORK (ANN)

Layer (Type)	Output Shape	Parameters (#)
Dense (32 units, ReLU activation)	(None, 32)	96
Dense (64 units, ReLU activation)	(None, 64)	2,112
Dense (1 unit, Linear activation)	(None, 1)	65
Total Trainable Parameters		2,273
Non-trainable Parameters		0

The input features (X) were passed through the network layers, transforming them into predictions of the target time series (y).

5.3 TRAINING AND EVALUATION

The ANN model was trained using the following configuration:

- **Batch Size:** 32,
- **Epochs:** 500,
- **Loss Function:** Mean Squared Error (MSE),
- **Optimizer:** Adam.

The model was trained on the training dataset ($X_{\text{train}}, y_{\text{train}}$) and validated on the testing dataset ($X_{\text{test}}, y_{\text{test}}$). Training and validation errors were monitored throughout the 500 epochs.

5.4 PERFORMANCE METRICS

The performance of the ANN was evaluated using the Root Mean Squared Error (RMSE) and R^2 metrics. Table 8 presents the results on the training and testing datasets.

TABLE 8: PERFORMANCE METRICS OF THE ANN MODEL

Metric	Training	Testing
RMSE (Normalized Scale)	0.6376	0.6845
RMSE (Original Scale)	-	3,290.34
R^2 (Testing Data)	-	0.5374
R^2 (Overall)	-	0.5822

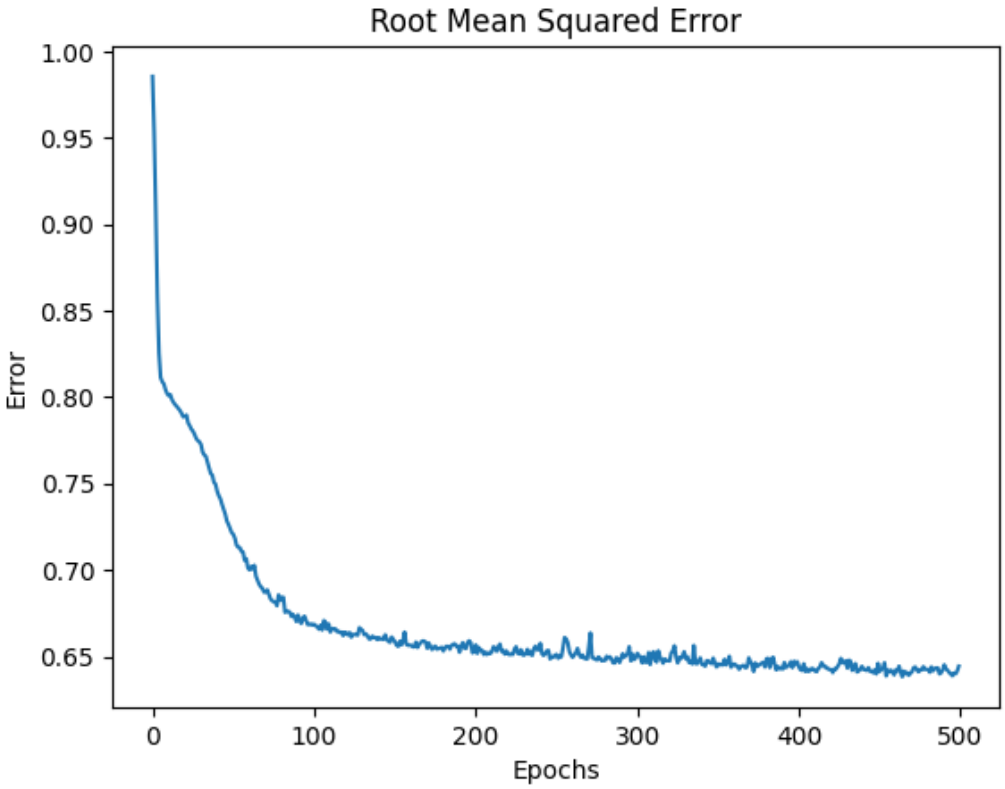


FIGURE 15: RMSE GRAPH OF THE ANN

The model demonstrated a normalized RMSE of 0.6376 on the training set and 0.6845 on the testing set. When converted to the original scale, the testing RMSE was approximately 3,294.395, with a NRMSE of 0.1900 and MAE of 2828.03.

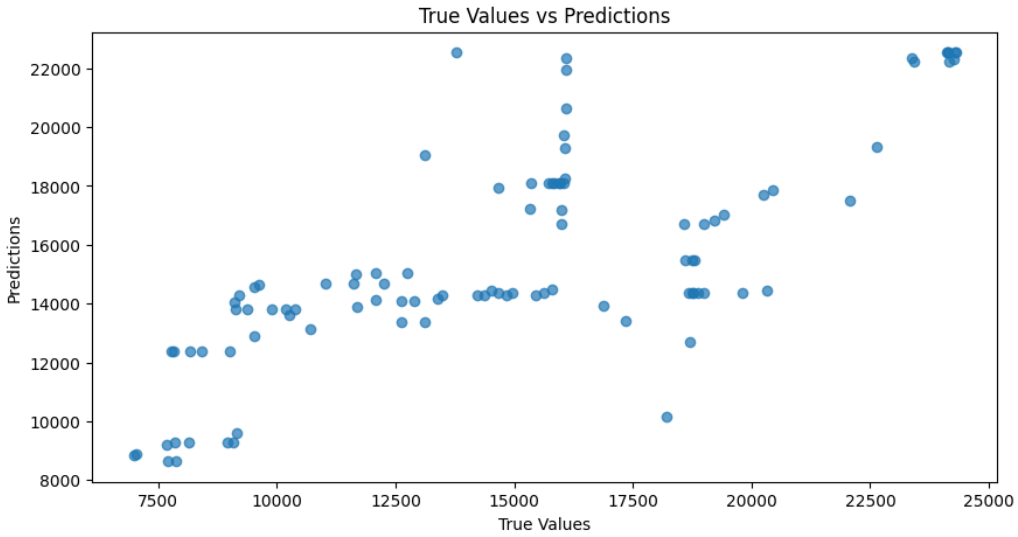


FIGURE 16: TESTING OF THE ANN

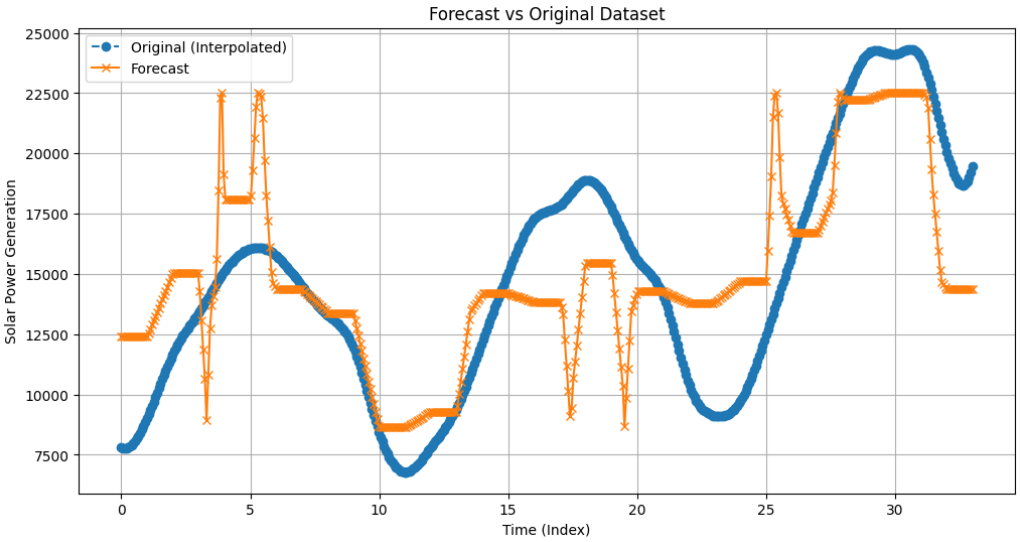


FIGURE 17: BACKTESTING FROM ANN

5.5 ATTEMPTS TO IMPROVE MODEL PERFORMANCE

In an effort to improve the performance of the ANN model, various hyperparameter tuning and architectural adjustments were tested. These included tuning epochs and batch size, creating a learning rate scheduler, and adding dropout regularization.

While these adjustments were expected to enhance the model's predictive capabilities, they introduced unexpected artifacts in the prediction and testing phases. Predictions displayed

sudden, unnatural fluctuations that did not align with the underlying patterns in the data. Testing errors became inconsistent, with significant variability across repeated trials. The introduction of dropouts often caused large deviations in predicted values, particularly for testing data.

5.6 DISCUSSION AND JUSTIFICATION FOR EXCLUSION

Despite its ability to capture non-linear relationships, the ANN model demonstrated limited predictive accuracy, as evidenced by the high RMSE and moderate R^2 values. During preliminary testing, the model produced inaccurate forecasts, making it less suitable for practical implementation compared to other approaches.

The following factors may have contributed to the model's underperformance:

- Insufficient complexity of the model architecture to fully capture the underlying patterns in the data,
- Overfitting or underfitting due to the selected hyperparameters and training configuration,
- Limited explanatory power of the input features (TSI and Growth Factor).

Given these limitations and the availability of alternative methods with superior performance, the ANN model was excluded from further testing and implementation. Future improvements may involve incorporating additional features, optimizing the network architecture, or employing advanced deep learning techniques to enhance accuracy.

6 TIME-LAG NEURAL NETWORK MODELING

A Time-Lag Neural Network (TLNN) was implemented as an alternative approach to capture temporal dependencies in the time series data. The TLNN used the same parameters as the Artificial Neural Network (ANN) model, but included a lagged structure to account for sequential patterns. While initial results were promising, further evaluation highlighted limitations in its ability to generalize.

6.1 INITIAL PERFORMANCE

The TLNN demonstrated significantly better performance during initial testing compared to the ANN model. The model was able to achieve an RMSE of 1360.255, NRMSE of 0.1209, and MAE of 1141.252.

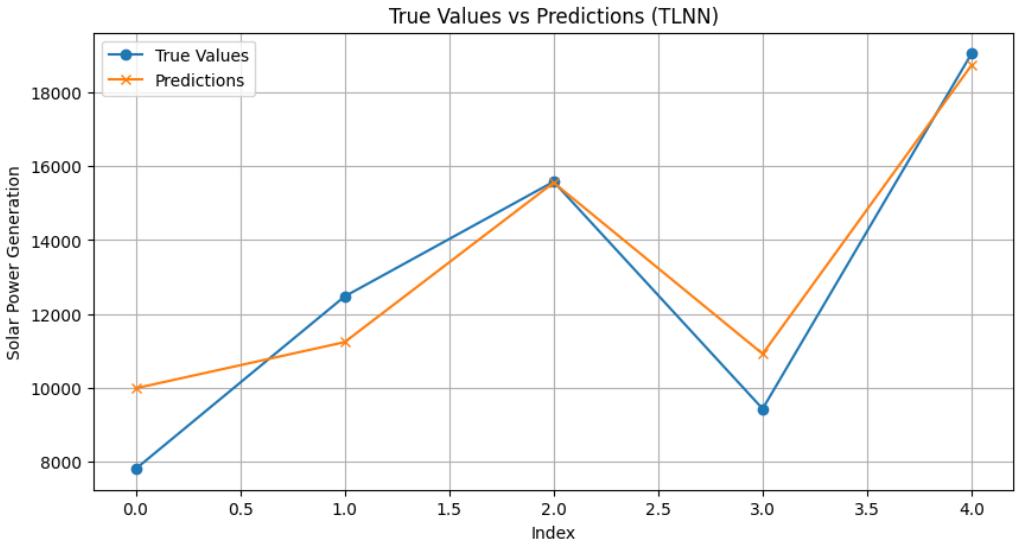


FIGURE 18: INITIAL FIT FOR TLNN

Furthermore, the predicted values closely matched the actual values, as shown in Figure 18, with the predictions capturing both the seasonal patterns and the overall trend.

6.2 ISSUES WITH FULL DATASET PREDICTION

When the TLNN model was trained on the entire dataset and used to forecast future values, the predictions exhibited a significant downward trend without the expected seasonal variations, as shown in Figure 19.

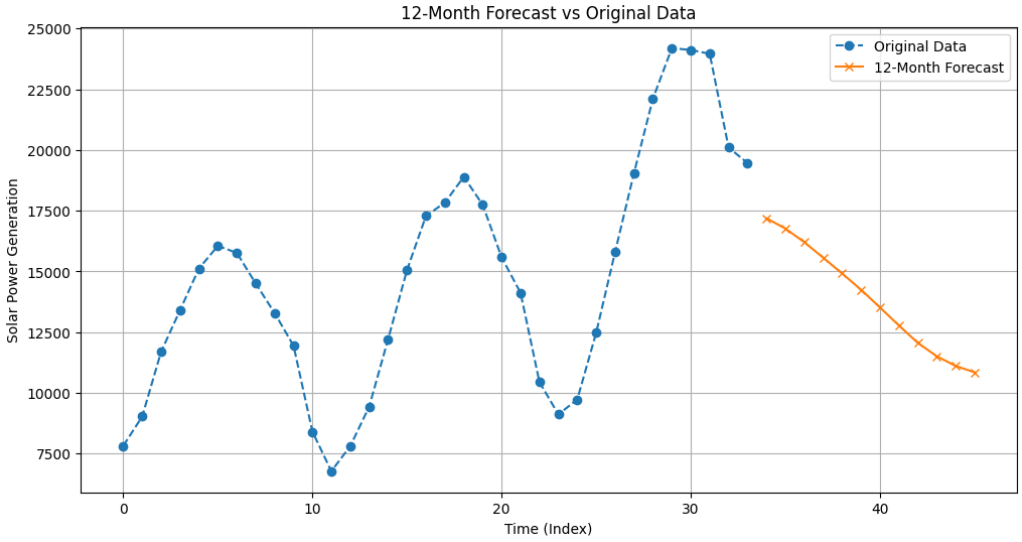


FIGURE 19: TLNN PREDICTIONS FOR FUTURE DATA

The downward trend and absence of seasonality in the predictions suggest that the TLNN failed to generalize effectively when applied to unseen data. Although the TLNN showed promise during initial testing, its inability to capture accurate seasonal patterns during future predictions rendered it unsuitable for the solar industry, where seasonality is a critical factor.

7 SEASONAL ARTIFICIAL NEURAL NETWORK (SANN) MODELING

7.1 PRELIMINARY TESTING RESULTS

During initial testing, the SANN demonstrated strong predictive performance on the test dataset, achieving an RMSE of 817.1864, NRMSE of 0.0499, and MAE of 647.7981.

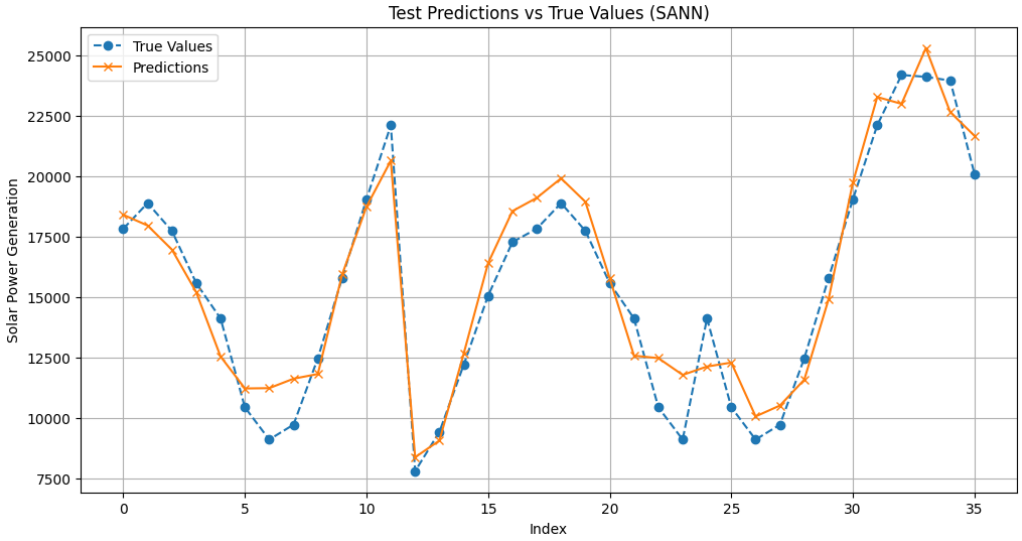


FIGURE 20: PRELIMINARY FITTING AND VALIDATION OF SANN

These results indicated the model's ability to accurately predict the target time series data. However, an artifact was observed when transforming the predictions back to the original scale, potentially due to non-linearities in the transformation or inconsistencies in the data preprocessing steps.

7.2 RESULTS

The SANN was trained on the full dataset and used to predict solar generation for the next 12 months. The predictions successfully captured seasonality and demonstrated reasonable growth trends for both utility-scale and small-scale solar generation.

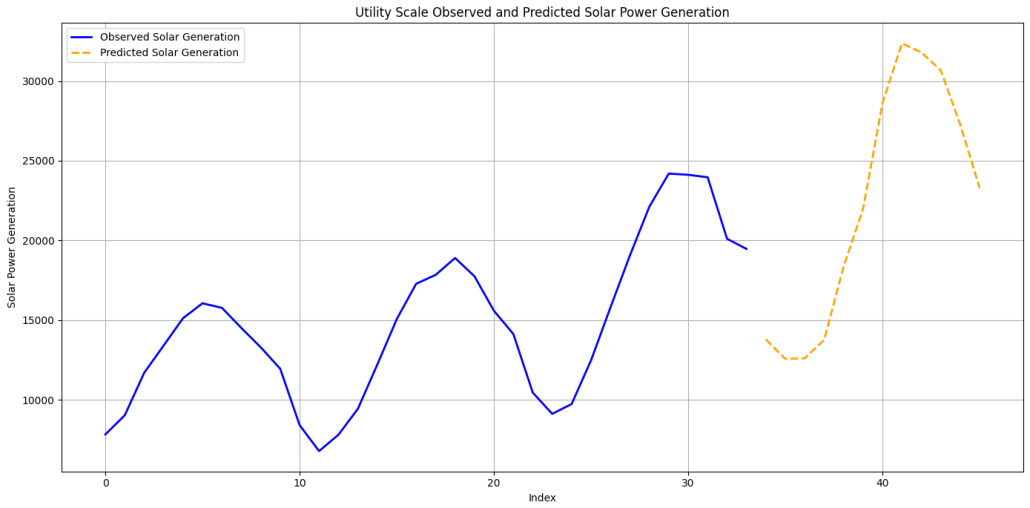


FIGURE 21: UTILITY SCALE SANN PREDICTION

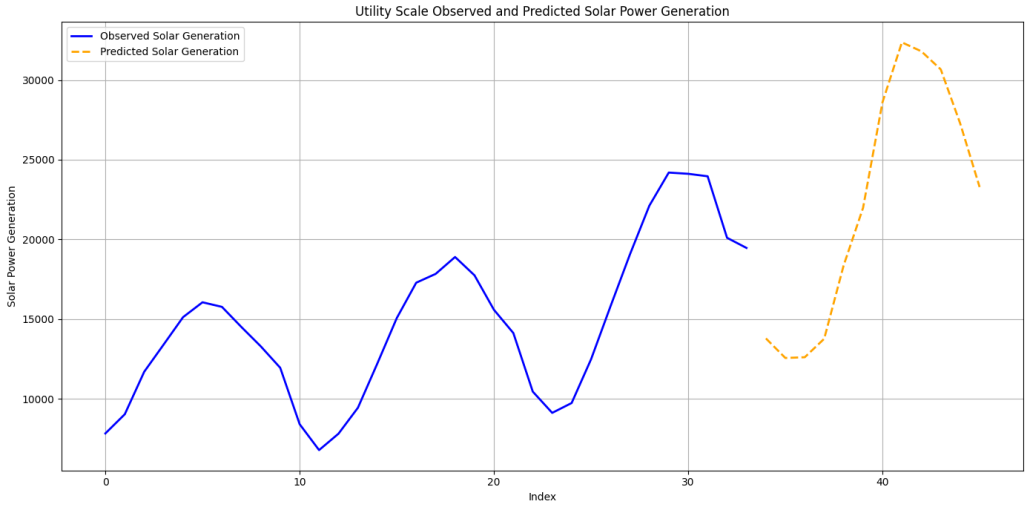


FIGURE 22: SMALL SCALE SANN PREDICTION

7.3 DISCUSSION AND OBSERVATIONS

The SANN model proved to be highly effective in capturing seasonal and growth trends for solar generation, particularly for small-scale applications. Key observations include:

- The SANN consistently reproduced seasonality, an essential characteristic of solar generation data, and modeled growth trends reasonably well.
- The higher error values for utility-scale predictions suggest potential areas for improvement, such as refining the feature set or exploring alternative transformations to mitigate artifacts in the original data.
- The artifact observed during the transformation back to the original scale warrants further investigation, as it could impact the accuracy of future forecasts.

7.4 CONCLUSION

The SANN model demonstrated significant potential for forecasting solar generation, with strong preliminary results and the ability to generalize seasonal patterns effectively. Despite some challenges related to artifacts and error magnitudes, the model represents a promising direction for future work. Refinements to the preprocessing steps, additional feature engineering, and further hyperparameter tuning could enhance its accuracy and reliability.

8 DISCUSSION AND CONCLUSION

8.1 COMPARISON OF MODEL PERFORMANCE

TABLE 9: PERFORMANCE METRICS OF DIFFERENT MODELS

Model	RMSE	NRMSE	MAE
Sinusoidal Exponential Midline Regression	889.7744	0.0395	748.6832
Sinusoidal Logistic Regression	3354.0380	0.1490	2628.7147
SARIMAX	1021.9122	0.1989	880.1877
Artificial Neural Network	3294.3956	0.1900	2828.0303
Time Lagged Neural Network	1360.2551	0.1209	1141.2519
Seasonal Artificial Neural Network	817.1864	0.0499	637.7981

SINUSOIDAL EXPONENTIAL MIDLINE REGRESSION

- **Pros:** Very accurate for given solar data.
- **Cons:** Too simplistic; cannot incorporate exogenous variables.

SINUSOIDAL LOGISTIC REGRESSION

- **Pros:** Higher accuracy; good for defined carrying capacity.
- **Cons:** Too simplistic; logistic graph stretches sinusoidal component.

SARIMAX

- **Pros:** Top contender for seasonal modeling; captures seasonality.
- **Cons:** Features TSI and Growth Factor had little significance.

ARTIFICIAL NEURAL NETWORK

- **Pros:** Captures some seasonality of solar industry.
- **Cons:** Many artifacts; inaccurate graph view; questionable jumps.

TIME LAGGED NEURAL NETWORK

- **Pros:** Captures time lag features of solar industry.
- **Cons:** Fails to capture seasonality; projects linear downtrend.



SEASONAL ARTIFICIAL NEURAL NETWORK

- **Pros:** Exceptional accuracy; captures 12-month seasonality.
- **Cons:** Slight variations across trainings; long-term stability unsure.

8.2 MODEL RECOMMENDATION

The evaluation of the models, based on the normalized root mean square error (**NRMSE**), highlights the **Seasonal Artificial Neural Network (SANN)** as the most accurate and reliable neural network model. With an NRMSE of **0.0499**, the SANN demonstrates exceptional performance, significantly outperforming all other NN models, and coming incredibly close to the sinusoidal exponential midline regression. Its ability to capture the 12-month seasonality inherent in solar generation data makes it ideal for high-precision forecasting. Additionally, SANN has exhibited consistent results even with minimal training (e.g., 150 epochs), showcasing its computational efficiency and robustness. For scenarios requiring high accuracy and reliability, SANN is the optimal choice.

For simpler applications where computational resources are limited or a straightforward implementation with no exogenous features is preferred, the **Sinusoidal Exponential Midline Regression** model is a practical alternative. With an NRMSE of **0.0395**, this model is not only accurate but also easy to apply, as it relies on a simple regression framework. Its simplicity makes it particularly suited for scenarios where the integration of exogenous variables or advanced modeling techniques is unnecessary. Despite its limitations in handling complex patterns or external factors, Sinusoidal Exponential Midline Regression remains a highly effective option for basic solar generation forecasting tasks.

In conclusion, the **Seasonal Artificial Neural Network (SANN)** is the recommended model for its superior accuracy and ability to capture seasonal trends in solar generation data. For cases where simplicity and ease of application are prioritized, the **Sinusoidal Exponential Midline Regression** provides a reliable and efficient alternative.

9 ASSUMPTIONS, LIMITATIONS, AND CHALLENGES

9.1 ASSUMPTIONS

Throughout the modeling process, several assumptions were made to guide the analysis and simplify the development of forecasting models. These assumptions include: - Exogenous variables, such as weather patterns, economic conditions, or policy changes, were excluded from the analysis under the assumption that solar generation trends could be accurately modeled based solely on historical patterns. - Error metrics such as RMSE, NRMSE, and MAE were assumed to be sufficient for evaluating model performance and drawing meaningful comparisons between different approaches. - It was assumed that there would not be a sudden crash in the solar industry or any major disruptions that could invalidate historical trends or significantly alter future generation patterns.

9.2 LIMITATIONS

Despite these assumptions, certain limitations impacted the project. A significant constraint was the limited availability of data; the U.S. Energy Information Administration (EIA) provided only monthly solar generation data from January 2022 to October 2024. For some of the more complex models, this data was insufficient to fully capture long-term trends or seasonality. Furthermore, the computational power available during the project was limited, as the models were trained on a standard i7 laptop. This restriction required reducing the number of features incorporated and limiting the number of epochs and folds during training.

9.3 CHALLENGES

In addition to these limitations, several challenges arose. As someone with limited expertise in the natural sciences and the solar industry, significant time and effort were spent researching how the solar industry operates, including the funding mechanisms and external factors that influence generation trends. This additional research was crucial for understanding the domain and informed decisions regarding model selection and data preprocessing.

9.4 FUTURE STEPS

Looking ahead, several steps can be taken to improve the forecasting models and address the limitations of the current approach. First, collecting or accessing more extensive datasets with longer time horizons and greater granularity (e.g., daily or hourly data) would enable the models to better capture both short-term and long-term trends. Second, incorporating additional exogenous features, such as weather data, economic indicators, or policy shifts, would likely

enhance model accuracy and relevance. Third, with access to more powerful computational resources, models could be trained with additional epochs and folds, allowing for more thorough exploration of hyperparameter spaces and improved generalization. Implementing k-fold validation during training would also help ensure that overfitting is identified and mitigated effectively.



References

[Willson & Hudson, 1991] Willson, R. C., & Hudson, H. S. (1991). The Sun's luminosity over a complete solar cycle. *Science*, *244*, 364–370.

[Pinto & Cavalieri, 1991] Pinto, R., & Cavalieri, S. (1991). Seasonal time series prediction with artificial neural networks and local measures. In *International Joint Conference on Neural Networks* (pp. 24–27).

[U.S. Energy Information Administration, 2023] U.S. Energy Information Administration. (2023). Electric Power Monthly. Retrieved from `https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=epmt_1_01_a`

[National Centers for Environmental Information, 2023] National Centers for Environmental Information. (2023). Total Solar Irradiance CDR. Retrieved from `https://www.ncei.noaa.gov/products/climate-data-records/total-solar-irradiance`

[Energy Markets & Policy Berkeley Lab, 2023] Energy Markets & Policy Berkeley Lab. (2023). Utility-Scale Solar. Retrieved from `https://emp.lbl.gov/utility-scale-solar`

[Islam et al., 2023] Islam, M. K., Hassan, N., Rasul, M., Emami, K., & Chowdhury, A. (2023). Forecasting of solar and wind resources for power generation. *Energies*. `https://doi.org/10.3390/en16176247`

[Lin & Pai, 2016] Lin, K.-P., & Pai, P.-F. (2016). Solar power output forecasting using evolutionary seasonal decomposition least-square support vector regression. *Journal of Cleaner Production*, *134*, 456–462. `https://doi.org/10.1016/J.JCLEPRO.2015.08.099`

[Abdullah et al., 2024] Abdullah, B. U. D., Khanday, S., Islam, N. U., Dhar, S. L. L., Fatima, H., & Nengroo, S. (2024). Comparative analysis using multiple regression models for forecasting photovoltaic power generation. *Energies*. `https://doi.org/10.3390/en17071564`