The Renewable Energy Project

FINAL REPORT

Written By

Daianne Starr

daiannestarr@gmail.com

github.com/daistarr

www.linkedin.com/in/dfstarr

Which type of renewable energy source is projected to witness the most substantial growth in the next decade in the United States?

PROBLEM STATEMENT

This case study presents an exploration of time series analysis using small datasets, focusing on the global paradigm shift towards renewable energy sources including hydropower, wind, solar, biofuel, and geothermal energy. It underscores the critical role this shift plays in environmental sustainability and strategic planning for policymakers, energy corporations, and investors, with a particular emphasis on the United States, a nation experiencing a relentless surge in energy demands. Through this analysis, we aim to demonstrate how even limited data can yield significant insights into the trends and dynamics of renewable energy adoption and its implications on future energy strategies.

OBJECTIVES

- To accurately forecast the growth trajectories of various renewable energy sources in the United States for the next decade using time series analysis.
- To identify the renewable energy source poised for the most significant expansion.

STAKEHOLDERS

- Policymakers require this data to craft informed policies and infrastructural plans.
- Energy Companies seek this information to align their investment and operational strategies with the anticipated growth.
- Investors utilize this information to optimize their investment portfolios in alignment with the forthcoming trends.

DATASET

- Historical record ranging from 1965 to 2022 from the "Renewable Energy World Wide" dataset from Kaggle.
- It was sourced from Kaggle and encompasses 17 distinct CSV files, each addressing various aspects of renewable energy sources on a global scale: production and consumption by source, such as hydropower, wind, solar, biofuels, and others.

RAW DATASET

	count	mean	std	min	25%	50%	75%	max
Year	57.0	1993.000000	16.598193	1965.000000	1979.000000	1993.000000	2007.000000	2021.000000
Geo Biomass Other - TWh	57.0	51.640315	24.831284	13.332232	25.756037	64.815414	73.159860	84.070230
Solar Generation - TWh	57.0	13.143548	34.702751	0.000000	0.000000	0.478253	1.095411	165.356570
Wind Generation - TWh	57.0	52.347070	98.781362	0.000000	0.000000	3.036189	34.797905	383.603270
Hydro Generation - TWh	57.0	274.981598	33.478110	198.974090	256.028530	274.030030	289.822450	355.973100
Electricity from solar (TWh)	57.0	25.374476	68.942213	0.000000	0.000000	0.956506	1.220000	328.840000
Solar Capacity	57.0	7.415070	19.288595	0.000000	0.000000	0.000000	0.974000	93.713016
Geothermal Capacity	57.0	995.899298	1231.805291	0.000000	0.000000	0.000000	2382.000000	3170.960000
Solar (% electricity)	57.0	0.308808	0.836247	0.000000	0.000000	0.012888	0.014777	3.959809
Wind (% electricity)	57.0	1.267605	2.375313	0.000000	0.000000	0.089173	0.831609	9.108380
Hydro (% equivalent primary energy)	57.0	3.557790	0.696467	2.383536	2.967953	3.602766	4.084431	5.202278
Solar (% equivalent primary energy)	57.0	0.135164	0.354995	0.000000	0.000000	0.005903	0.011445	1.675419
Renewables (% equivalent primary energy)	57.0	5.303081	1.705532	3.421371	4.327570	4.614475	5.334361	10.655991
Wind (% equivalent primary energy)	57.0	0.545342	1.016217	0.000000	0.000000	0.038384	0.363568	3.886730
Electricity from hydro (TWh)	57.0	547.594989	67.588075	397.948180	509.400000	544.508000	579.644900	711.946200
Biofuels Production - TWh - Total	57.0	103.151536	151.322025	0.000000	0.000000	22.044724	162.899440	424.440060
Electricity from wind (TWh)	57.0	103.603974	195.393345	0.000000	0.000000	6.072378	68.900000	756.400000
Other renewables including bioenergy (TWh)	57.0	51.265390	24.452388	13.332232	25.756037	64.640000	72.240005	83.070000
Wind Capacity	57.0	19.024343	34.748126	0.000000	0.000000	0.000000	16.515000	132.737600
Hydro (% electricity)	57.0	4.866521	3.775441	0.000000	0.000000	6.506543	7.653214	11.097251
Renewables (% electricity)	57.0	7.606756	6.241404	0.000000	0.000000	9.291262	11.360623	20.749863

DATA WRANGLING

During data preprocessing, I merged all these datasets based on the year and country name. In case the observation was split in more than one row for a given year and country, the rows were cumulatively summed and the parameters coalesced. This strategy facilitated the formulation of a consolidated dataset. Subsequently the data was filtered to exclusively encompass the United States, which is the focus of the report. Rows with absent values were removed to assure data integrity.

The final dataset contains 57 rows, representing data from 1965 to 2022, and 21 columns containing the year feature as well as renewable power sources types: geo biomass, wind, solar, hydro electricity, biofuels electricity generation (TWh) as well as percentage of electricity produced from the same sources.

EXPLORATORY DATA ANALYSIS

HISTOGRAMS

In this step, the objective is to explore historical trends and patterns among various renewable energy sources. Initial exploration involved generating histograms for numerical variables to evaluate data distribution, revealing predominantly skewed distributions.

Also, the majority of variables originating from the same energy source exhibit similar, if not virtually identical, data distributions.

Using all of them in subsequent modeling stages post-Exploratory Data Analysis (EDA) could yield circular results. Therefore, only one of these variables will be selected for use in the subsequent steps.

For example, as illustrated in **Figure A**, the variables - solar generation, solar capacity, percentage of equivalent solar primary energy, and percentage of electricity sourced from solar - present a significant similarity. As a result, only one among them will be applied in the phases post-EDA.

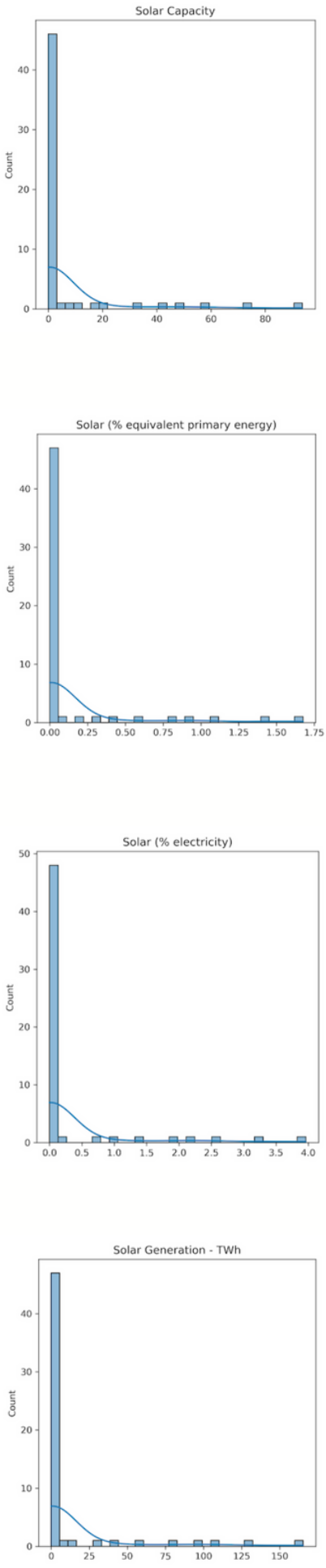
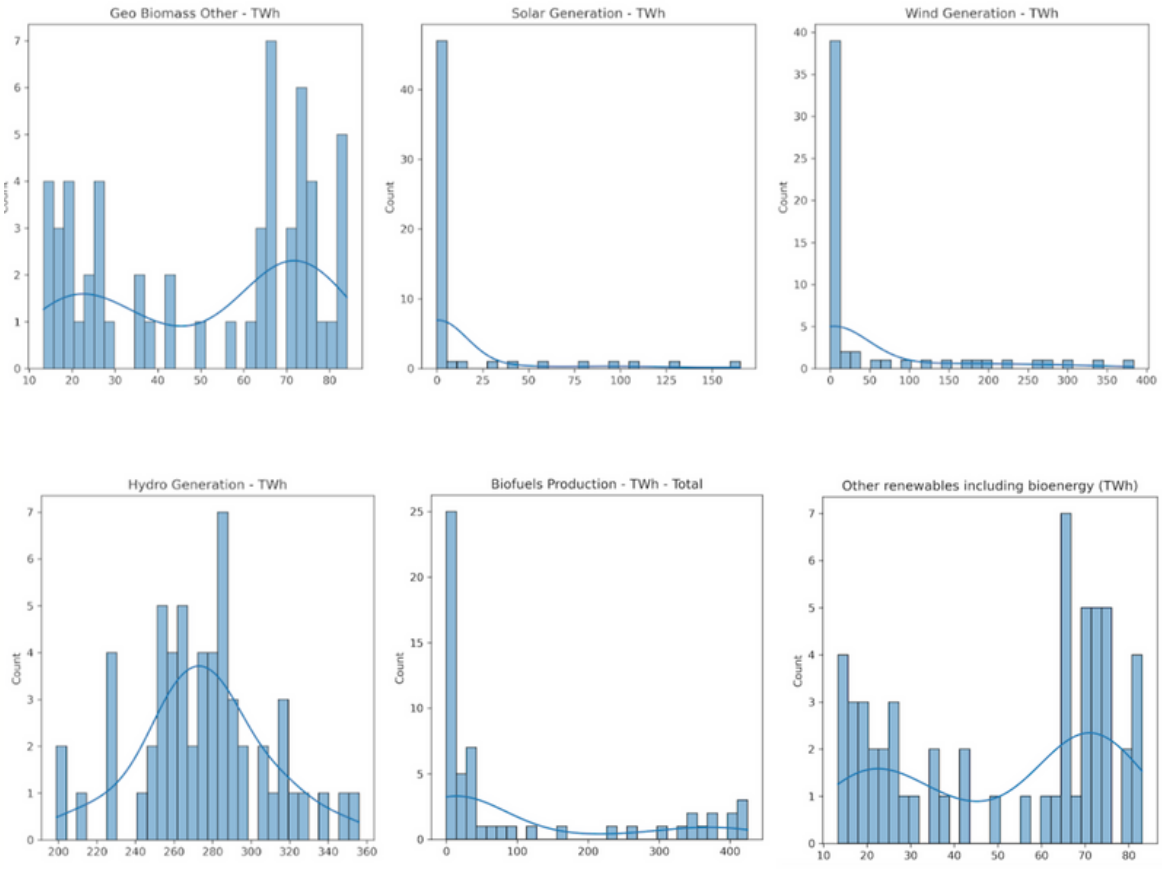


Figure A: Distributions of solar-related variables showcase notable similarities, potentially leading to circular data interpretations in subsequent post-EDA phases.

Figure B: Distributions of geo biomass, solar, wind, hydro, biofuels and other sources of renewable energy.



Overall, these histograms suggest that shifts in the energy landscape have occurred over time, with certain renewable sources emerging more prominently in recent years as seen in **Figure B**. Geo biomass energy is predominantly clustered towards the lower range, interspersed with a few years exhibiting elevated values. Solar, wind and geothermal energy demonstrated a notable left skew.

This suggests an initial substantial contribution that has been gaining momentum in recent years. Hydro and biofuel sources displayed a more balanced distribution, signaling a steady energy source throughout the years. Other renewable sources exhibited a mild left skew yet maintained a wider distribution, indicating a gradually increasing contribution over time.

HEATMAP

The heatmap provides a visual representation of the Pearson correlation coefficients between the selected features as seen in **Figure C**. A strong positive correlation exists between solar and wind energy generation. The synchronous growth in the solar and wind energy generation could be attributed to simultaneous advancements or investments in these renewable energy technologies. Also, geo biomass and wind show a positive correlation, although not as strong as the correlation between solar and wind generation.

This implies that higher biomass energy generation is associated with higher wind energy generation to some extent. Further analysis might explore if certain factors, like governmental policies or global energy trends, are simultaneously influencing these two energy sectors. Hydro demonstrates weak to moderate positive correlations with wind and solar, which might be influenced by factors such as geographical conditions or policies promoting renewable energy sources. Biofuels do not exhibit strong correlations with the other variables, which suggests that biofuels production might be influenced by different factors compared to the other forms of energy generation, or it might operate independently from them in the context of development, investment, and production.

Figure C: The heatmap reveals all the correlations among these variables are mostly positive, suggesting that increases in one type of energy generation tend to coincide with increases in others, although to varying degrees. This might be reflective of overall trends in energy production over time, such as a general shift towards or development in renewable energy sources.

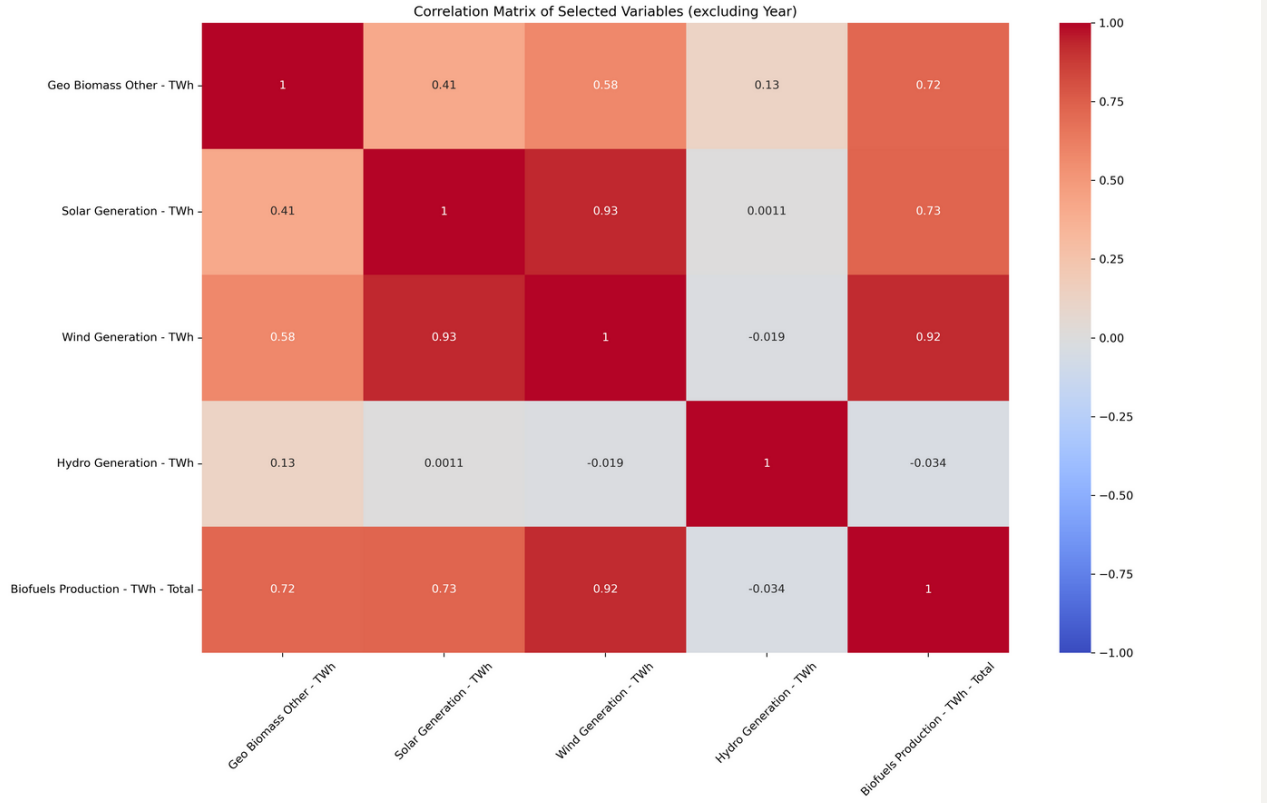
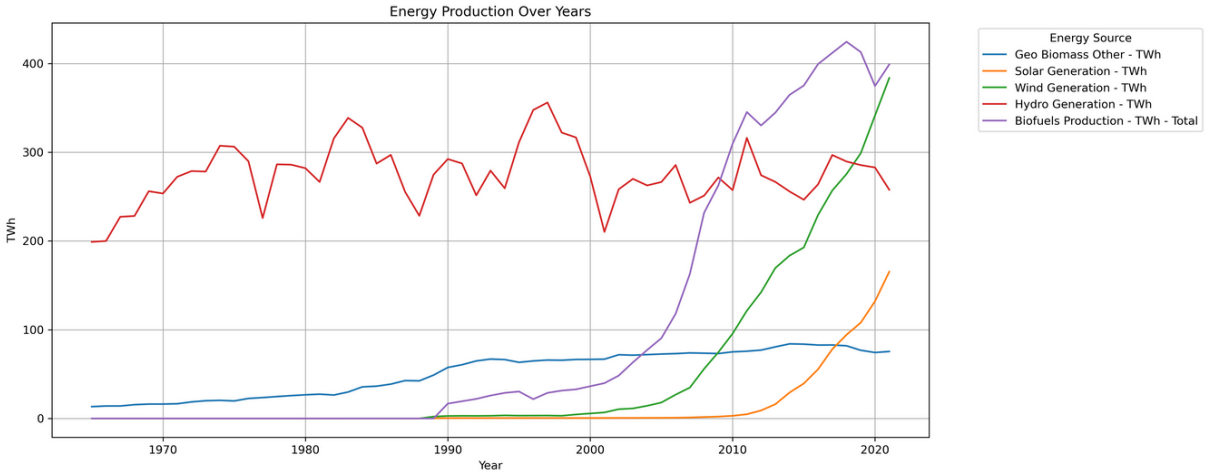


Figure D: The line plots depict the varied evolution of energy sources, with solar and wind seeing recent sharp rises, hydro and geothermal remaining consistent, and an overall increase in renewable contributions.



LINEPLOT

The line plots give us a temporal perspective on the evolution of various energy sources as seen in Figure D. Solar generation has witnessed a significant uptick, especially in the latter years, indicating a growing adoption and utilization of solar energy. This could be due to advancements in solar technology, decreases in the cost of solar installations, and possibly due to policy initiatives promoting renewable energy sources. Wind generation has experienced a consistent and significant upward trend. Similar to solar energy, advancements in technology and supportive policies might have contributed to this upward trajectory. The growing global emphasis on renewable energy sources to mitigate climate change could also be a key factor. Geo Biomass generation shows an overall increase but with noticeable fluctuations.

This might suggest that the production is influenced by various factors, such as biomass availability, technological applications, or policy changes that might impact its adoption. Biofuels show a consistent increase over the years. This might be associated with a rise in the adoption of biofuels as an alternative to fossil fuels, particularly in sectors like transportation. The development of biofuel technology and possibly incentives for cleaner fuel alternatives might drive this growth. Hydro generation seems to be stable and doesn't exhibit the same growth trend as wind and solar energy. This could be due to geographical or environmental limitations, as hydroelectric power generation often requires substantial water bodies and specific geographical conditions. Also, existing hydroelectric facilities might have reached their production capacities.

MODELING

The objective of this stage involves employing Time Series Forecasting to formulate predictive models, aimed at forecasting the growth of various renewable energy sources over the subsequent decade. A variety of approaches and models were explored for a methodological comparison, being compared based on the **Mean Absolute Error (MAE)**, **Root Mean Square Error (RMSE)**, the latest recorded value, the predicted value, growth percentage, **Compound Annual Growth Rate (CAGR)**, and **Sum of Squared Residuals (SSR)**. Also, to validate and evaluate the model, data were analyzed utilizing machine learning, allocating 80% of the dataset to the training set and the remaining 20% to the testing set. Furthermore, all model plots feature solid lines representing the actual observed data, dashed lines indicating the forecasted values for the test set period (model validation period), and dotted lines projecting forecasted values for the next decade (2022-2031).

LINEAR REGRESSION

Linear regression selected as the preliminary model due to its simplicity, as illustrated in **Figure E**. Predominantly, the model anticipates significant growth in solar generation, while wind, hydro, geo biomass, and other energy sources, including bioenergy, are also expected to observe substantial development. The results indicate that hydro generation could potentially witness the steepest predicted growth rate (288%) and CAGR (14.52%) over the ensuing decade, positioning it as a compelling candidate for the most rapidly expanding renewable energy source. Nonetheless, it's pivotal to note that the prevalent appearance of a large Sum of Squared Residuals (SSR) across all energy sources signals that the linear regression model may lack the robustness required for reliable forecasting.

ARIMA

NO-STATIONARY DATA

For the second model, I chose ARIMA (AutoRegressive Integrated Moving Average), a prominent model known for its proficiency in analyzing and forecasting time series data, particularly due to its capability to handle data with a trend. For this second approach, I decided not to check the stationarity of the data because I wanted to preserve the original structure and trends within the dataset, which I believed could hold valuable information for my predictions. In this approach, for some energy sources, such as Solar Generation, the model appears to capture the trend but underestimates later-year values as seen in **Figure F**. Wind Generation shows visible deviations from the rapid increase in later years. Hydro Generation aligns well with the observed data, indicating decent predictive performance.

Figure E: Linear regression model predicting notable growth in energy sources, especially hydro generation over the decade, albeit with potential forecasting limitations signaled by substantial SSR.

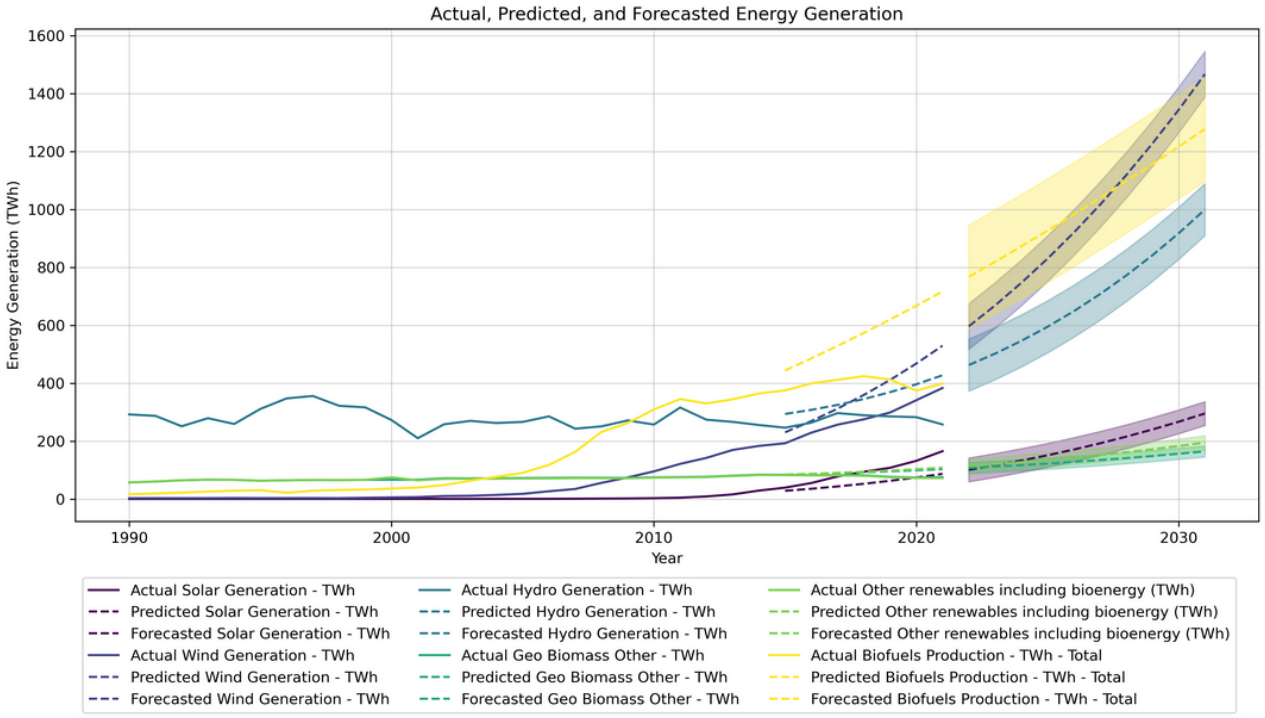
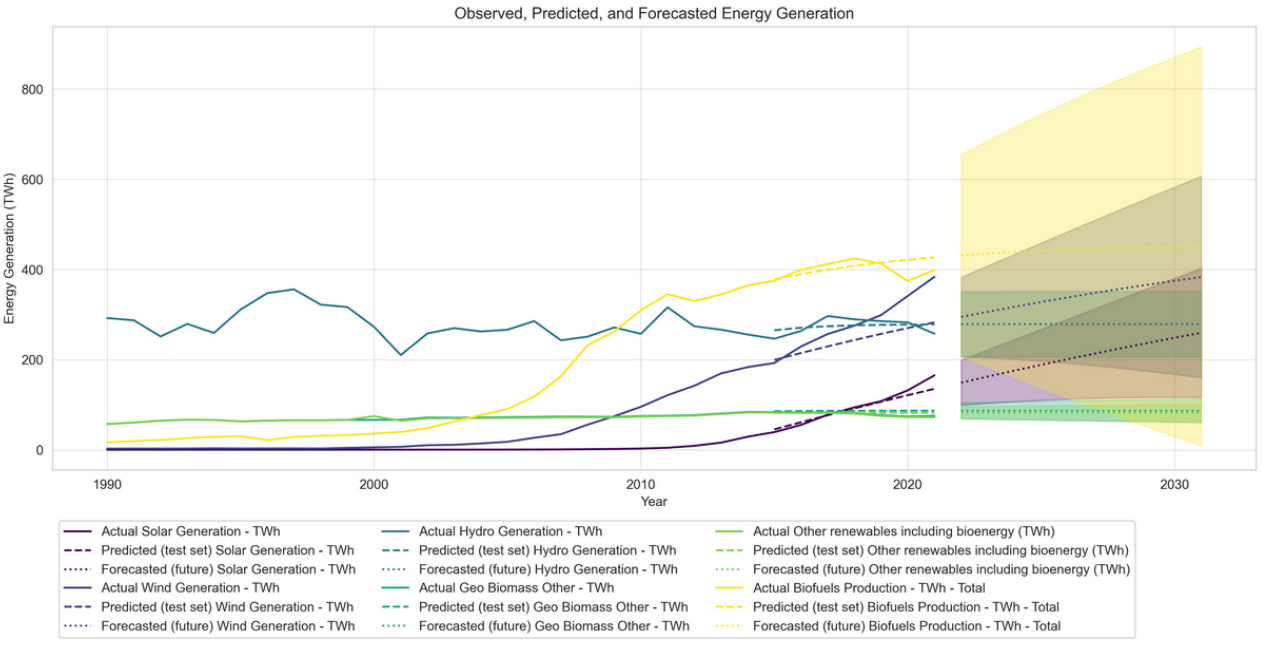


Figure F: In the first ARIMA model, Solar generation appears to have the highest predicted growth rate (57%) and CAGR (4%) over the next decade, making it a strong candidate for being the renewable energy source that will grow the most.



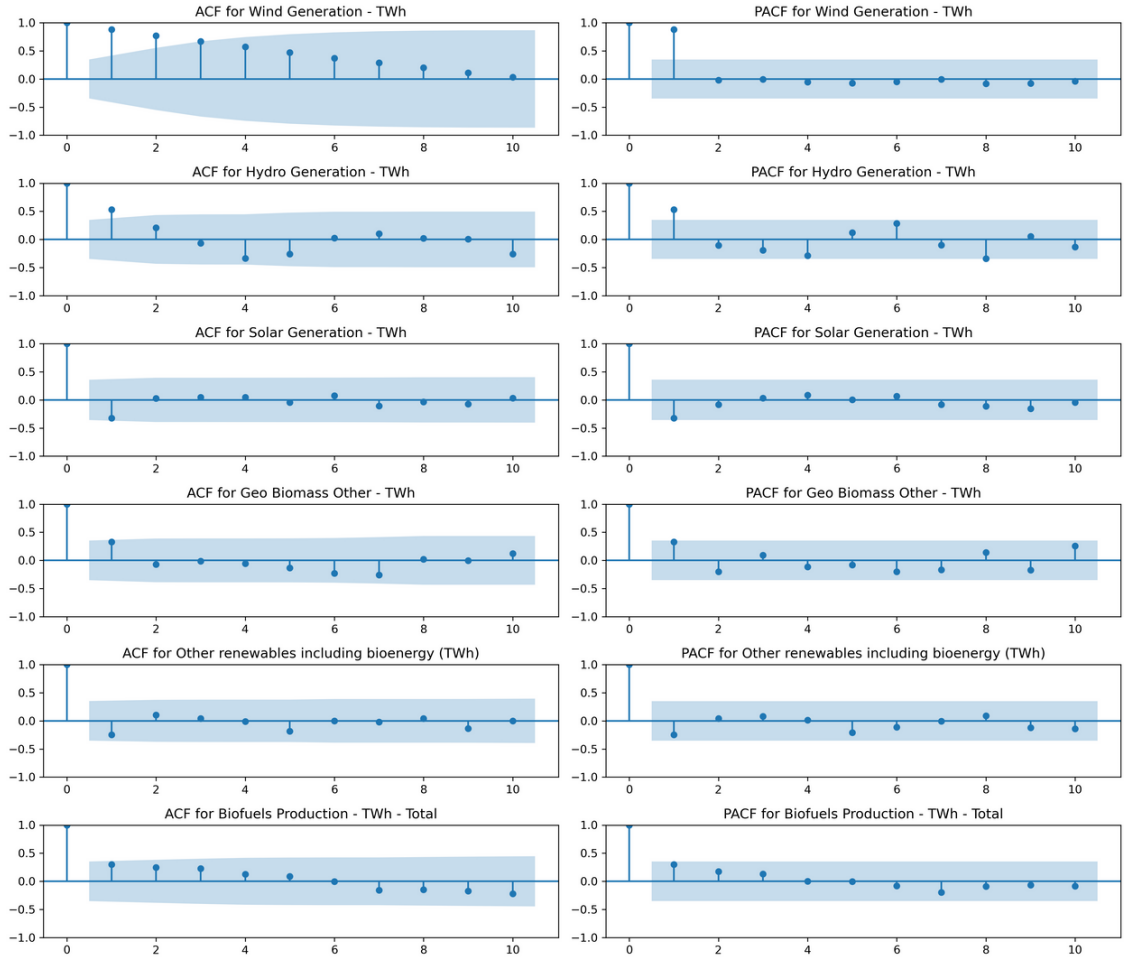
Other sources exhibit varied performance and might benefit from further refinement. Relatively high MAE and RMSE for Solar and Wind values suggest potential inaccuracies and underperformance. Conversely, Hydro and Geo Biomass generation demonstrates lower MAE and RMSE, suggesting better short-term forecast accuracy. Biofuels Production displays moderate MAE and RMSE, indicating reasonable performance with room for improvement. However, similarly to the last model, it's worth noting that the widespread occurrence of a large SSR for the Solar source suggests that this model may not be robust enough for reliable forecasting.

STATIONARY DATA

The second approach entailed evaluating the stationarity of the time series data utilizing the Augmented Dickey-Fuller (ADF) test.

Engaging in an assessment of stationarity facilitates suitable data transformation, ensuring that fundamental statistical properties, such as mean and variance, are preserved consistently over time. Contrarily, in the preceding approach, models were initially formulated utilizing a general ARIMA(1, 0, 1) order, a strategy potentially suboptimal for all individual series within the analysis, thereby possibly compromising the accuracy of the forecasts. If the data was deemed non-stationary, it underwent first and second order differentiation followed by log transformation. By analyzing Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for each energy source, I obtained the we can discern potential model orders, as seen in the **Figure G**:

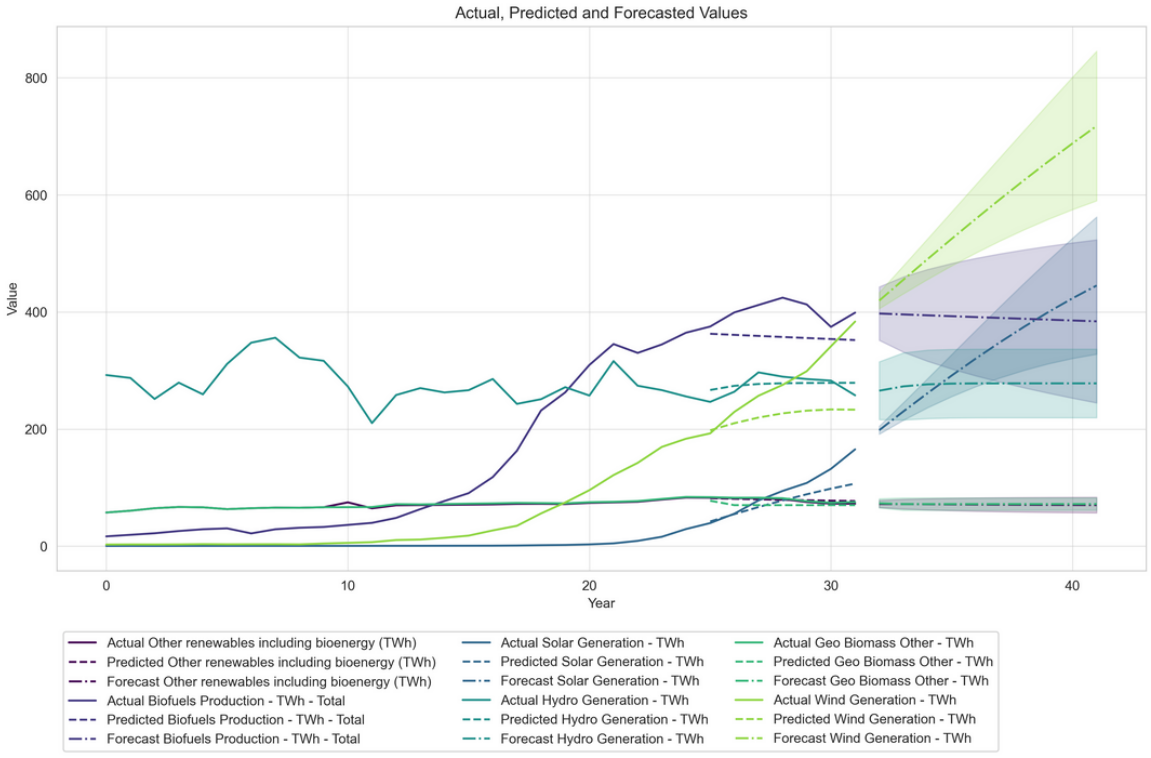
Figure G: The ACF and PACF tests identify the most fitting ARIMA model for each renewable energy type.



- Solar Generation: The PACF has a sharp cut-off at lag 2, and the ACF decays gradually, suggesting an AR(2) model might be suitable.
- Geo Biomass: The ACF has a sharp cut-off at lag 1, indicating an MA(1) model might be suitable.
- Other Renewables including Bioenergy: The PACF has a sharp cut-off at lag 1, suggesting an AR(1) model might be appropriate.
- Biofuels Production: The PACF has a sharp cut-off at lag 1, while the ACF decays gradually, suggesting an AR(1) model might be suitable.
- Wind Generation: The ACF and PACF do not show a clear cut-off, which might indicate that an ARMA model could be suitable.
- Hydro Generation: The ACF shows a gradual decay, while PACF has a cut-off at lag 2, suggesting an AR(2) model might be appropriate.

Therefore, the second approach for ARIMA considering the different behavior of the stationary data for renewable energy sources shows relatively high MAE and RMSE for Solar, Wind, and Biofuels values might indicate that the model does not fit the data as well as for other energy types. However, the low MAE and RMSE for Other renewables, Hydro, Geo Biomass suggest that the model fits the data well, and thus, the forecast might be reliable. Based on these results, Solar generation appears to have the highest predicted growth rate (44 %) and CAGR (3.74%) over the next decade, making it a strong candidate for being the renewable energy source that will grow the most, as shown in **Figure H**. However, similarly to the last model, it's worth noting that the widespread occurrence of a large SSR for the Solar source suggests that this model may not be robust enough for reliable forecasting.

Figure H: Solar, Hydro and Wind predict a notable increase in generation over the next decade; Other renewables, Biofuels Production, and Geo Biomass Other are predicted to decrease slightly or remain relatively stable.



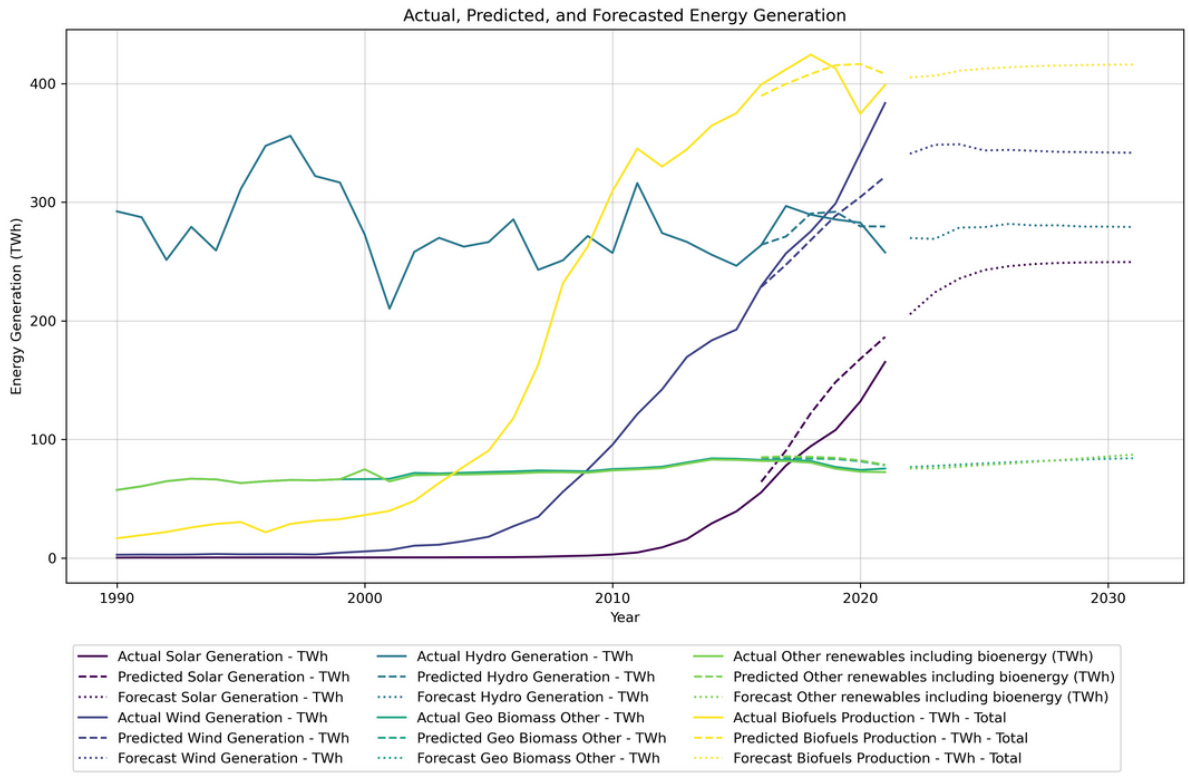
LSTM

The last model is LSTM , which stands for Long Short-Term Memory and is a specialized type of Recurrent Neural Network (RNN) adept at learning patterns in sequences of data. This model is t particularly beneficial for this time-series forecasting because it is able to predict the growth of various renewable energy sources by effectively capturing temporal dependencies and managing possible vanishing gradient issues.

In this model, Solar generation, though exhibiting a consistent upward trend, appears to be underestimated by the model as shown in **Figure I**. Wind generation, conversely, shows a slight overestimation in the latter years, while Hydro generation's predicted values align relatively well with actual data, indicating commendable model performance.

Other energy types demonstrate varying performances and may benefit from further model optimization. The model performance metrics are diverse across the different energy sources. Solar Generation displays relatively high MAE and RMSE values, suggesting potential underestimation in the forecasts. Wind Generation also shows considerable MAE and RMSE, indicating potential model overfitting. In contrast, Hydro Generation demonstrates lower MAE and RMSE values, hinting at more accurate short-term forecast performance. Geo Biomass and Other Renewables also exhibit relatively lower MAE and RMSE, which is a promising sign, whereas Biofuels Production portrays moderate MAE and RMSE, indicating a reasonable but improvable performance.

Figure I: Based on these results, Solar generation appears to have the highest predicted growth rate 25.2%) and CAGR (2.2%) over the next decade, making it a strong candidate for being the renewable energy source that will grow the most.



MODEL SELECTION

In determining the optimal model for predicting the most rapidly growing renewable energy source over the next decade, I assessed various error metrics—Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Sum of Squared Residuals (SSR)—across several models, namely Linear Regression, Non-Stationary ARIMA, Stationary ARIMA, and LSTM, each employed to forecast energy generation for each source.

A prevalent approach to model selection entails opting for a model that minimizes error metrics such as MAE, RMSE, and SSR, where lower values typically suggest a better-fitting model, provided it does not overfit the data. Initially, I calculated the mean MAE, RMSE, and SSR for each model type to account for the different orders of magnitude of energy generation among various energy sources, thereby ensuring the mean reflects these variances comprehensively. Based on this approach, Non-Stationary ARIMA performed the best, with the lowest MAE and RMSE as well as the second lowest SSR, as seen in **Figure J**. Within this model, the renewable energy source with the highest percentage change is the Solar energy generation as seen in **Figure K**.

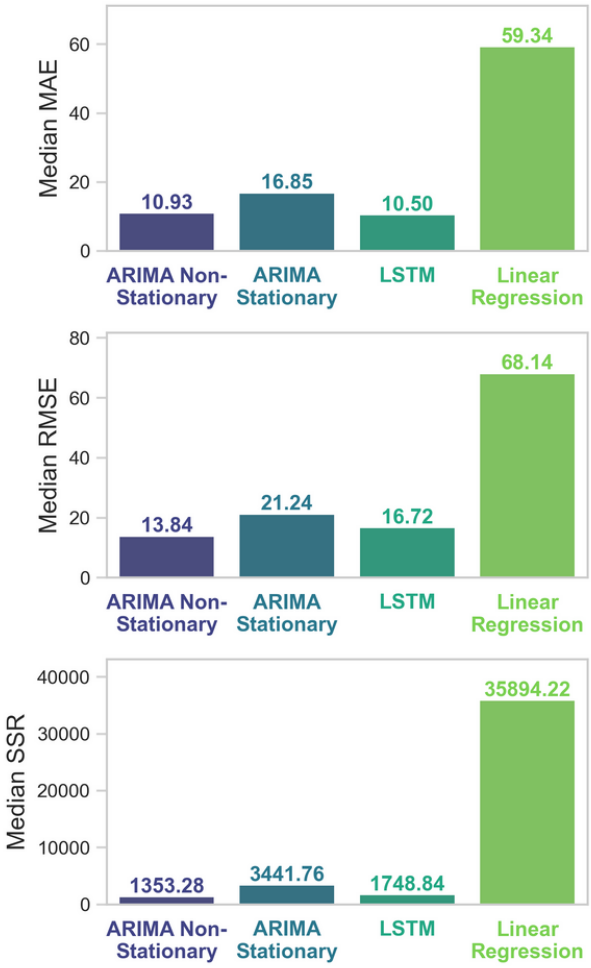
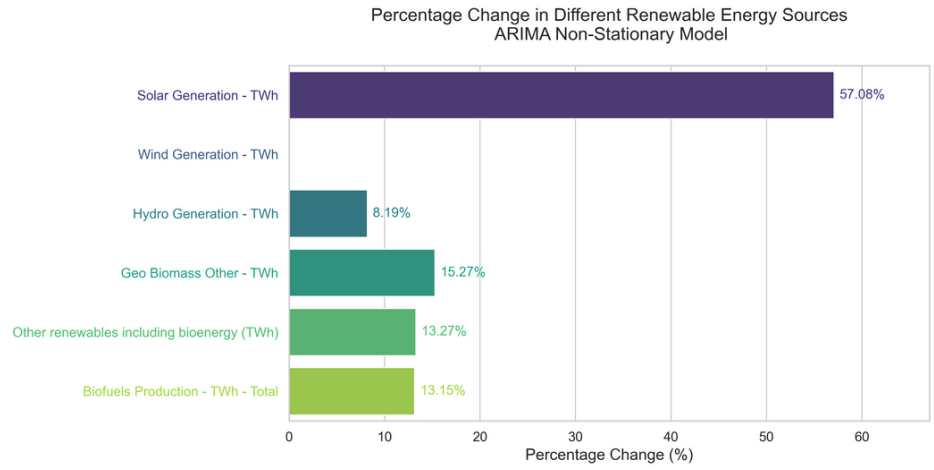


Figure J: Comparative Analysis of MAE, RMSE, and SSR for Solar Energy Generation, with Non-Stationary ARIMA Demonstrating Optimal Performance.

Figure K: Bar chart illustrating the projected percentage changes in various renewable energy sources based on the ARIMA Non-Stationary model. Notably, Solar generation exhibit the most expressive growth.



TAKEAWAYS

- **Insights & Limitations from Small Datasets:** While small datasets have provided valuable insights, it's crucial to recognize the limitations due to external factors. These include geopolitical tensions, policy changes, and financial incentives, which can all significantly sway renewable energy trends and outcomes.

- **Optimal Predictive Model:** The ARIMA Non-Stationary model stands out as the most suitable for this dataset.

- **Foremost in Future Growth**: Solar energy is forecasted to witness the most substantial growth in the next decade.

- **Navigating Towards Renewables**: Marked trends towards solar, wind, and biofuels suggest a shift to more sustainable energy, likely spurred by targeted policies and incentives.

- **Influence of Technological Progress**: Technological advancements and diminishing production costs, especially in the realms of solar and wind energy, are likely driving factors in their increased adoption and development.