Data Analytics Report and Executive Summary

Research Question

A. Summarize the original real-data research question you identified in task 1. Your summary should include justification for the research question you identified in task 1, a description of the context in which the research question exists, and a discussion of your hypothesis.

The research question asks how predictive analytics can be used to forecast energy demand and supply mismatches in Puerto Rico's power grid under stress conditions. Accurate forecasting is essential for ensuring a sufficient and reliable supply of energy. This question exists in the context of Puerto Rico's energy transition campaign to transition away from all non-renewable energy technologies by 2050. The hypothesis is that predictive analytics can improve forecasting accuracy by using historical data and grid performance metrics to better prepare and respond to potential mismatches between energy supply and demand during high-stress events.

Data Collection

B. Report on your data-collection process by describing the relevant data you collected, discussing one advantage and one disadvantage of the data-gathering methodology you used, and discussing how you overcame any challenges you encountered during the process of collecting your data.

For this analysis, I collected two datasets: "3MS 2025," which includes predicted hourly energy generation data in MWh for various technologies from July 1, 2024, to June 30, 2025, generated by the Sienna model (National Renewable Energy Laboratory [NREL], 2024) as part of the PR100 Report (National Renewable Energy Laboratory, n.d.), and "energy_cap_3MS," which provides nameplate generation capacities in MW for different scenarios and years, modeled using the Engage tool (National Renewable Energy Laboratory, n.d.).

An advantage of this data-gathering methodology is that the datasets are publicly accessible and come from a reputable source: the U.S. Department of Energy. One disadvantage to this dataset though, was the amount of potential data available to choose from. Hourly energy usage had been predicted for 7 different years from 2025 to 2050. On top of that, there were 12 different scenarios to choose from that included variations on land use and electric load. This left me with 84 options to choose from for this analysis.

I decided to use data for the year 2025 under the 3MS scenario. The year 2025 for timely relevance and the 3MS scenario because it involved more land use and more stress on electric load. It was the scenario that could be considered the "worst case" and thus, most urgent and relevant for analyzing energy and supply mismatches.

Data Extraction and Preparation

C. Describe your data-extraction and -preparation process and provide screenshots to illustrate *each* step. Explain the tools and techniques you used for data extraction and data preparation, including how these tools and techniques were used on the data. Justify why you used these particular tools and techniques,

including one advantage and one disadvantage when they are used with your data-extraction and -preparation methods.

1. Data Extraction:

The Pandas library was used for data extraction and manipulation, and loading CSV files into DataFrames. The os module's `os.path.join` was used to ensure that file paths could be used cross-platform if needed. Pandas is effective in managing large datasets with minimal code but for very large datasets Pandas can be memory-intensive.

```
import itertools
import matplotlib.pyplot as plt
import os
import pandas as pd
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.statespace.sarimax import SARIMAX
import warnings
```

```
# Read in csv files and preprocess data
path = r"C:\Users\prest\WGU\Capstone Project\IntegratedCapacityInvestment\3MS Working Copies"
three_ms_path = os.path.join(path, "3MS_2025.csv")
cap_path = os.path.join(path, "energy_cap_3MS.csv")
```

2. Data Preparation:

For data preparation, I created a `load_and_process` function that accepts two dataframes, one for generated energy and one for energy caps. The function first drops unnecessary columns that won't be used in the analysis: 'Unserved Energy' and 'Over Generation'. Next, the 'ts' column is converted to datetime format. The energy cap dataframe is then filtered to only include data for the year 2025, column names for wind energy are standardized between the two dataframes, and unused columns are dropped.

Next, a 'relevant_technologies' list was created to include only technologies that exist in both the 'generated energy' DataFrame and the 'energy caps' DataFrame. A dictionary, 'cap_mapping', is then made where each technology is a key, and its associated energy capacity is the value, allowing for quick lookup of capacities by technology type.

The mismatch between energy generation and capacity is then calculated in a for loop with new columns created for each technology type. The data was then resampled to a daily frequency, which reduces the dataset from almost 9000 rows to 365 rows.

```python
# Function to load and preprocess data
def load_and_preprocess_data(three_ms_path, cap_path):
    df_3ms = pd.read_csv(three_ms_path).drop(columns=['Unserved Energy', 'Over Generation'])
    df_3ms['ts'] = pd.to_datetime(df_3ms['ts'])

    df_cap_2025 = pd.read_csv(cap_path).query('Year == 2025').replace({'Land-based Wind':
        'Onshore Wind'}).drop(columns=['Unnamed: 0', 'Year'])

    relevant_technologies = df_3ms.columns.intersection(df_cap_2025['Technology'])
    cap_mapping = df_cap_2025.set_index('Technology')['Energy Capacity (MW)'].to_dict()

    # Calculate mismatch between generation and capacity
    for tech in relevant_technologies:
        df_3ms[f'{tech} Capacity (MW)'] = cap_mapping.get(tech, 0)
        df_3ms[f'{tech} Mismatch'] = df_3ms[tech] - df_3ms[f'{tech} Capacity (MW)']

    df_3ms['Total Mismatch'] = df_3ms.filter(like='Mismatch').sum(axis=1)
    df_3ms.set_index('ts', inplace=True)
    return df_3ms.resample('D').mean().reset_index()
```

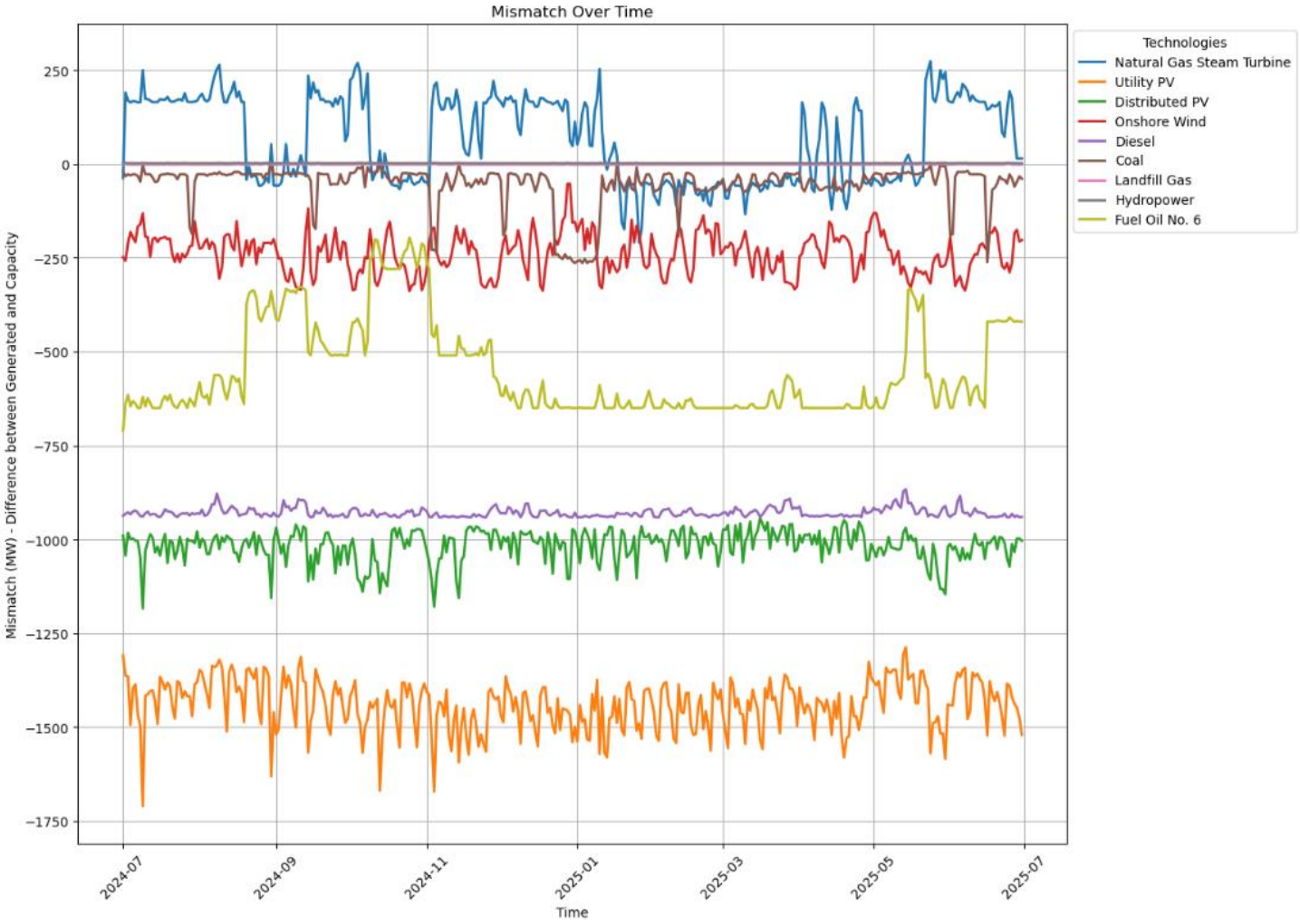
```python
df_3ms_daily = load_and_preprocess_data(three_ms_path, cap_path)
df_3ms_daily.set_index('ts', inplace=True)
```

Analysis

D. Report on your data-analysis process by describing the analysis technique(s) you used to appropriately analyze the data. Include the calculations you performed and their outputs. Justify how you selected the analysis technique(s) you used, including one advantage and one disadvantage of these technique(s).

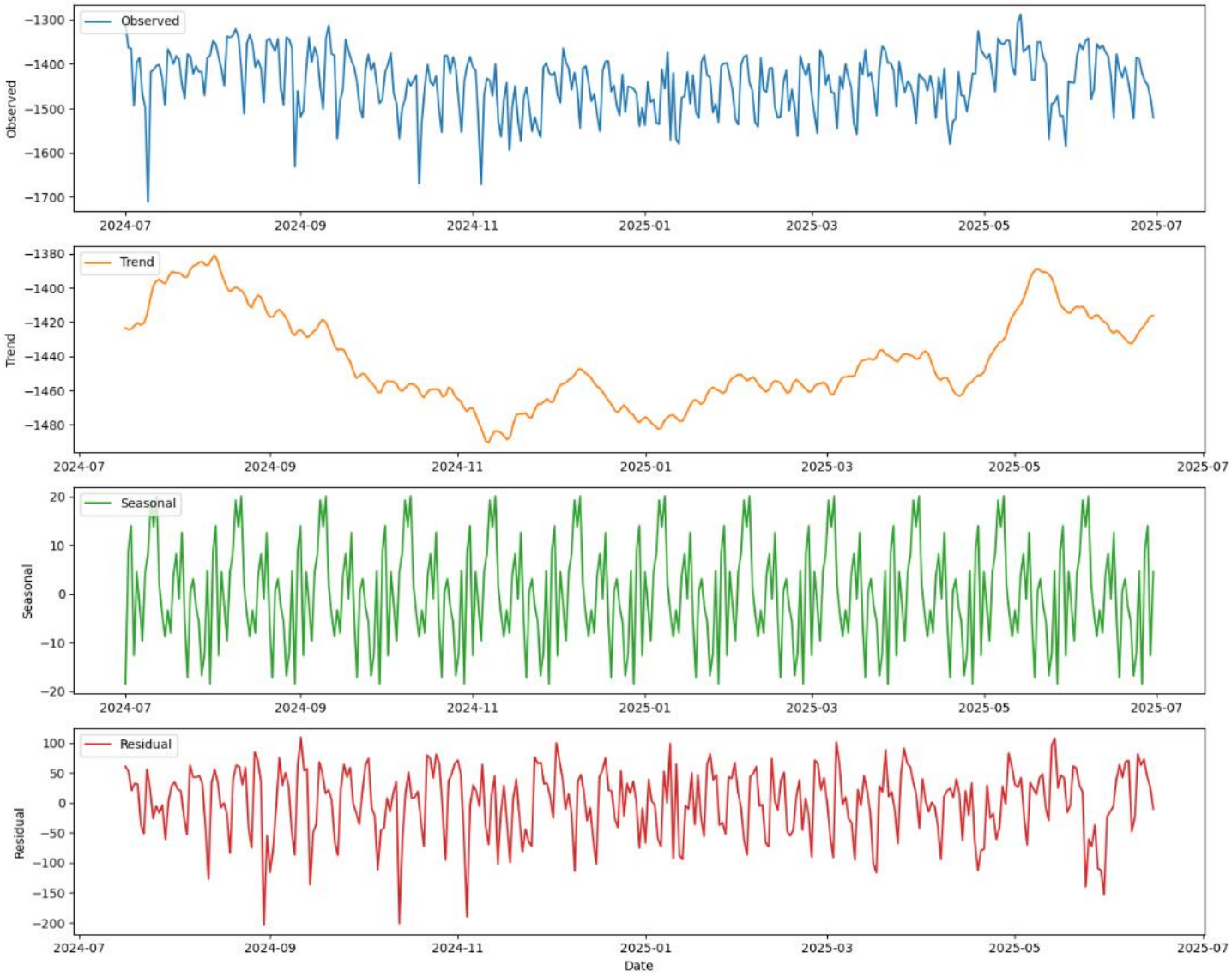
1. Mismatch Plotting:

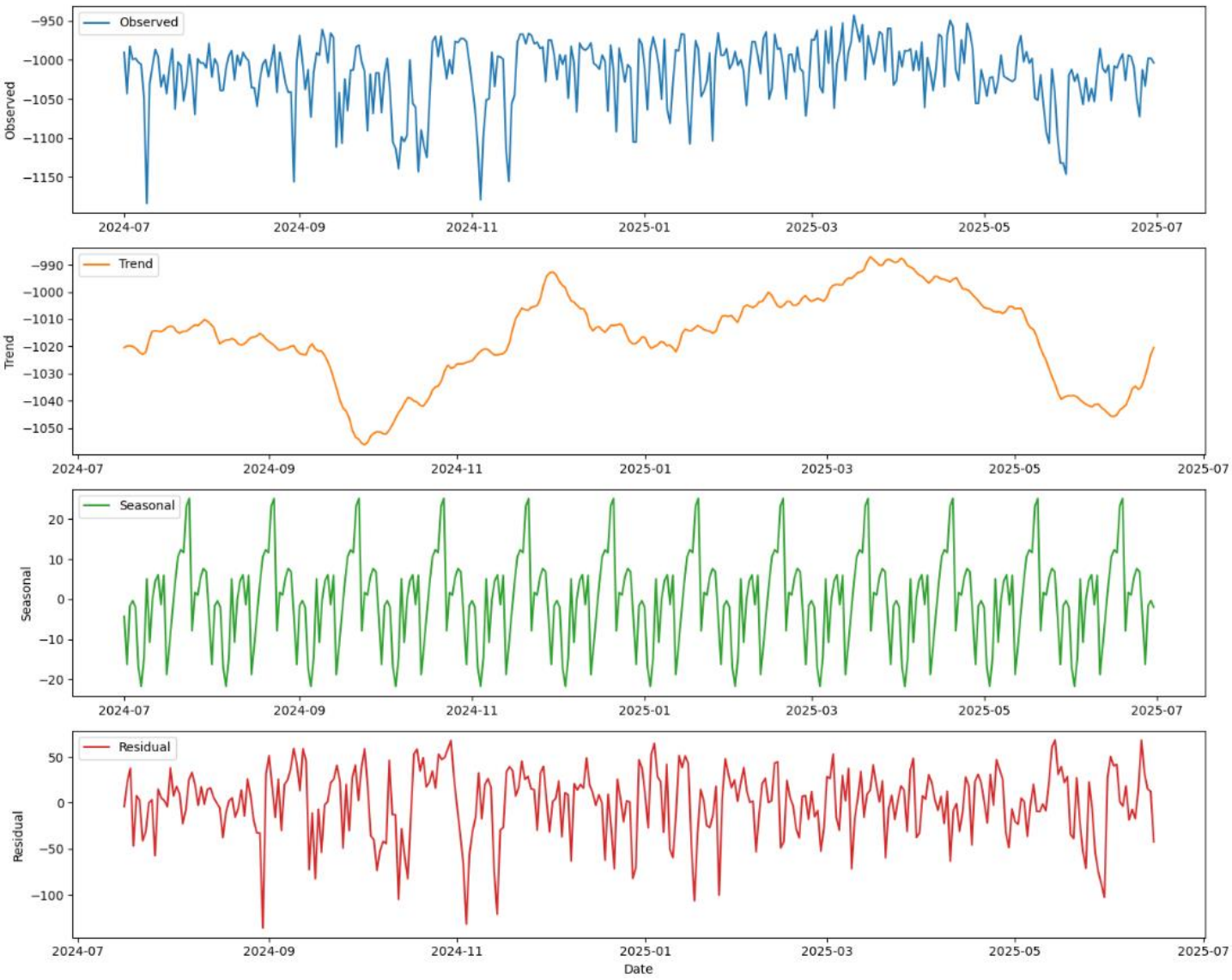
I used a time series plot to visualize the mismatch between energy generation and capacity for each technology. This plot provided a clear view of how the mismatches changed over time and which energy technologies had a greater mismatch than others.

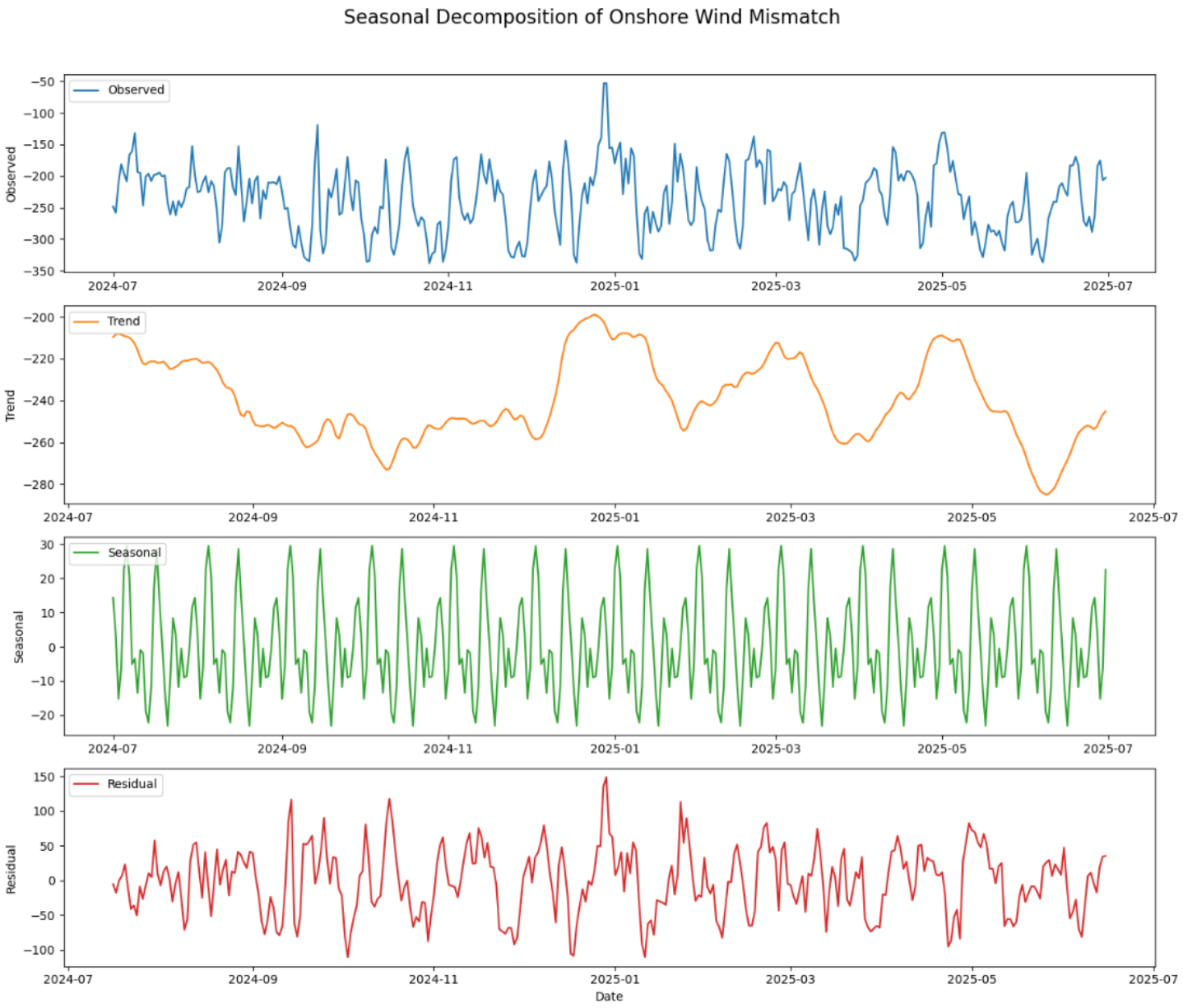


2. Seasonal Decomposition:

I then used the `seasonal_decompose` method to break down the time series data into observed, trend, seasonal, and residual components. This decomposition was applied to the three renewable energy technologies with significant mismatch values: `Utility PV', 'Distributed PV', and 'Onshore Wind'. These plots highlight the influence of seasonal effects and long-term trends on the mismatches.

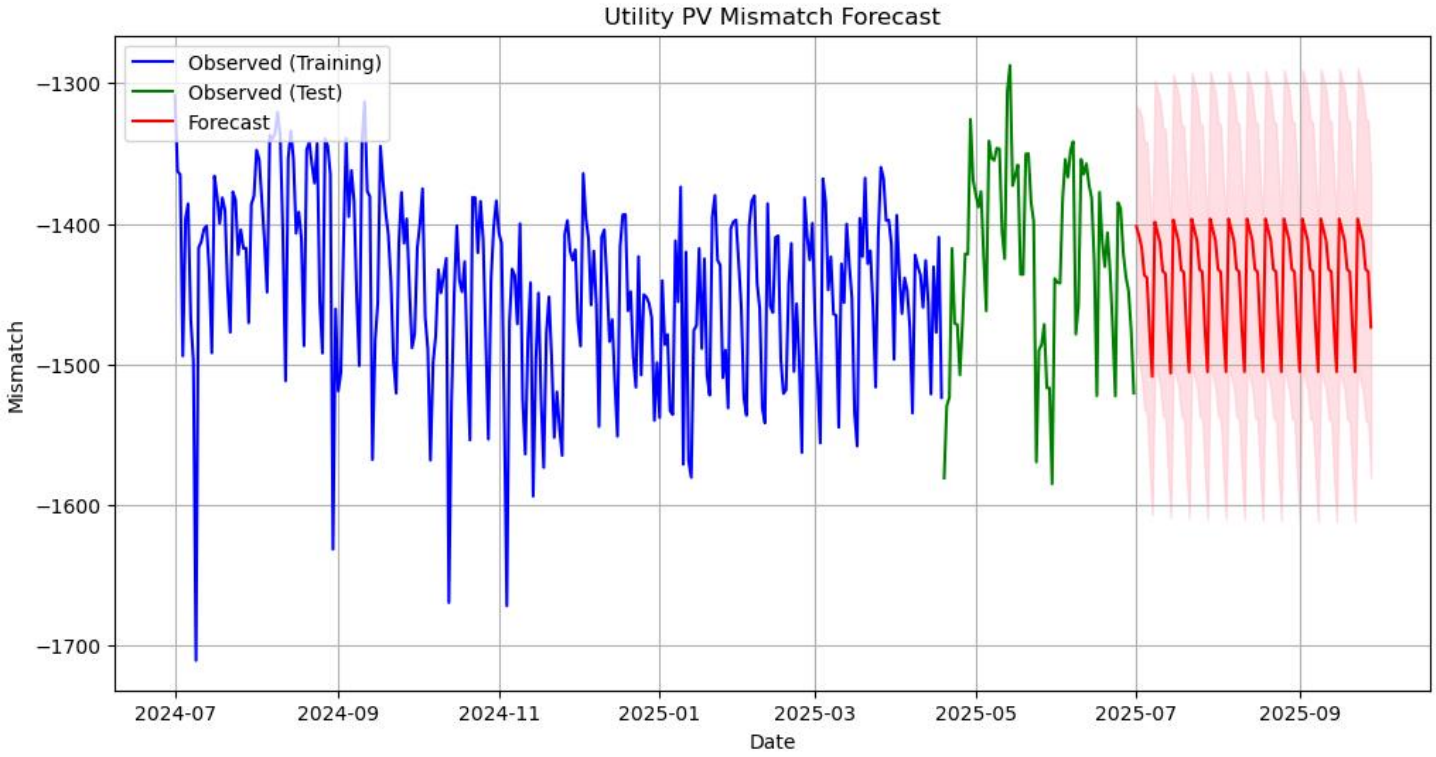

Seasonal Decomposition of Utility PV Mismatch
Observed
Trend
Seasonal
Residual
Date
2024-07
2024-09
2024-11
2025-01
2025-03
2025-05
2025-07

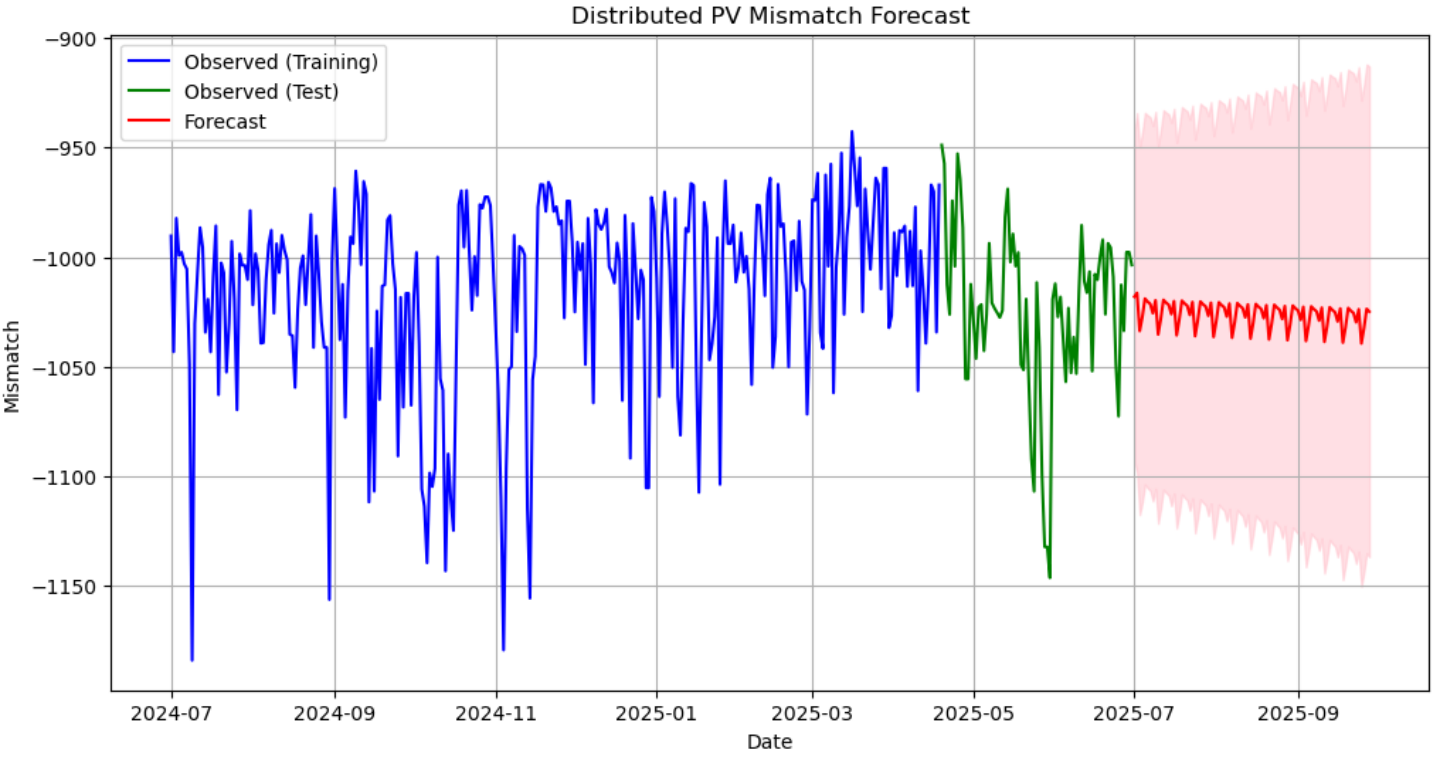

Seasonal Decomposition of Distributed PV Mismatch
Observed
Trend
Seasonal
Residual
Date
2024-07
2024-09
2024-11
2025-01
2025-03
2025-05
2025-07

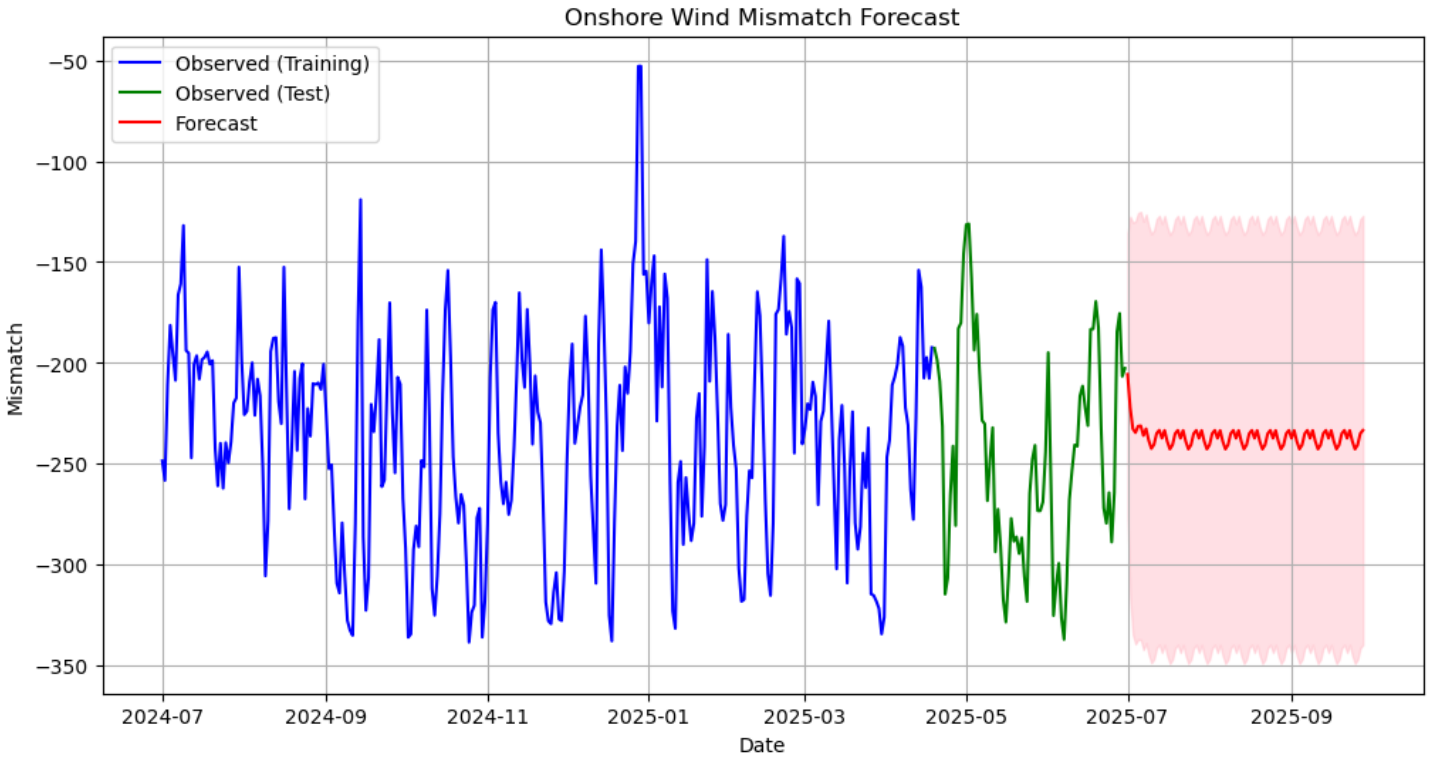


3. ARIMA Modeling and Forecasting:

Last of all, I used an ARIMA (AutoRegressive Integrated Moving Average) model to forecast future mismatch values. A grid search was used to find the best parameters based on the AIC (Akaike Information Criterion). The projections provided by this model can help anticipate potential issues in energy supply and demand, and in planning infrastructure development. ARIMA is a robust technique for time series forecasting, especially for data with trends and seasonality. A downside of the ARIMA model, though, is it can be computationally intensive and require tuning to avoid overfitting.


Utility PV Mismatch Forecast
Observed (Training)
Observed (Test)
Forecast
−1300
−1400
−1500
−1600
−1700
Mismatch
2024-07
2024-09
2024-11
2025-01
2025-03
2025-05
2025-07
2025-09
Date


Distributed PV Mismatch Forecast
Observed (Training)
Observed (Test)
Forecast
−900
−950
−1000
−1050
−1100
−1150
Mismatch
2024-07
2024-09
2024-11
2025-01
2025-03
2025-05
2025-07
2025-09
Date



Data Summary and Implications

E. Summarize the implications of your data analysis by discussing the results of your data analysis in the context of the research question, including one limitation of your analysis. Within the context of your research question, recommend a course of action based on your results. Then propose two directions or approaches for future study of the data set.

This analysis shows that ARIMA modeling can effectively forecast energy supply and capacity mismatches in Puerto Rico's energy grid. The potential future mismatches portrayed in the forecasts can help in planning and mitigating supply issues. The Mismatch Over Time plot also paints a clear picture of which energy technologies are being used efficiently and which are not. One limitation of this analysis is that the historical data used is for only one year and is based on a prediction model. Real-world historical data that span several years would certainly make for a more robust and accurate forecast.

Based on the results of this analysis, predictive analytics can provide valuable foresight into making proactive adjustments to energy distribution and generation. Integrating responsive energy resources, like battery storage, can be used to prepare for and smooth out the forecasted energy mismatches. Upgrades to existing infrastructure can lead to efficient energy distribution where energy is generated but not utilized.

This study used daily averages for energy generation. A future study of the data set could include hourly data for more precise and timely forecasting. Another approach could integrate the impacts of extreme weather events on energy supply in order to increase the resilience and flexibility of Puerto Rico's energy grid.

F. Acknowledge sources, using in-text citations and references, for content that is quoted.

1. National Renewable Energy Laboratory (NREL). (2024). *Sienna Model*. Retrieved from https://www.nrel.gov/analysis/sienna.html
2. National Renewable Energy Laboratory (NREL). (n.d.). *Engage Tool*. Retrieved from https://engage.nrel.gov/en/login/?next=/en/
3. National Renewable Energy Laboratory. (n.d.). PR100: Power sector modeling for Puerto Rico's grid modernization. Retrieved August 20, 2024, from https://pr100.gov/