Forecasting Wind Power Generation: A Time Series Analysis of Meteorological Impacts

Anton Yang

2024-11-26

Abstract

Many industry experts, researchers, and data scientists are exploring renewable energy, with a particular focus on wind power. The goal is to understand the dynamic relationship between meteorological variables and their impact on wind power generation. This study considers time series models, including the Seasonal Autoregressive Model, to analyze these relationships. Since many variables may or may not contribute to the model's construction, our objective is to identify the optimal model and evaluate whether additional variables improve the prediction of future observations. This approach is supported by examining autocorrelation and partial autocorrelation graphs and validating the model by minimizing mean squared errors. As a result, we found that the ARIMA(0,1,1) with regression coefficient of wind speed and dew point is the most optimal model for forecast the wind power generation.

Introduction

Rising oil costs emphasize the need for sustainable energy solutions, with wind power standing out as one of the most promising renewable energy technologies due to its high efficiency and minimal environmental impact. Wind power is a clean and renewable energy source that harnesses the wind's energy using turbines. These turbines convert mechanical power into electricity without burning fuel or polluting the air. As an abundant and inexhaustible resource, wind energy helps reduce carbon emissions significantly, avoiding approximately 336 million metric tons of carbon dioxide annually in the United States—equivalent to the emissions from 73 million cars [1]. The forecast of wind vitality assumes a vital part in the portion of balance control.

However, wind energy production is inherently variable, influenced by complex meteorological factors such as temperature, humidity, wind speed, wind direction, and gusts. The increasing reliance on renewable energy highlights the importance of optimizing wind power generation through predictive modeling. By understanding how meteorological conditions impact wind energy output, we can enhance forecasting accuracy and operational efficiency.

For this project, we will analyze the Wind Power Generation Dataset from Kaggle [2]. This dataset contains 4 location site, and for this project, we'll analyze location 1. This dataset was created to explore the dynamic relationship between meteorological variables and their effect on wind power generation. It is a comprehensive compilation of field-based meteorological observations and wind turbine output data, offering valuable insights into the interplay between weather conditions and renewable energy production. The dataset spans detailed hourly records starting from January 2, 2017. To facilitate easier analysis and observe broader trends, the hourly data will be aggregated into weekly averages across all variables. This transformation will enable a more meaningful exploration of the patterns and factors influencing wind power generation. The goal of this project is to develop an optimal forecasting model and identify the key variables that contribute to its predictive accuracy. By understanding the factors that significantly influence wind power generation, we aim to enhance the model's forecasting capability and provide actionable insights for improving energy efficiency and planning. The detail of the explanatory variables are shown in Table 1.

Variables	Descriptions
Date	Weekly date when readings occured.
temperature_2m	Temperature in degrees Farenheit at 2 meters above the surface.
relativehumidity_2m	Relative humidity (as a percentage) at 2 meters above the surface.
dewpoint_2m	Dew point in degrees Farenheit at 2 meters above the surface.
windspeed_10m	Wind speed in meters per second at 10 meters above the surface.
windspeed_100m	Wind speed in meters per second at 100 meters above the surface
winddirection_10m	Wind direction in degrees (0-360) at 10 meters above the surface
winddirection_100m	Wind direction in degrees (0-360) at 100 meters above the surface.
windgusts_10m	Wind gusts in meters per second at 100 meters above the surface.
Power	Turbine output, normalized to be between 0 and 1.

Table 1: Explanatory Variables

Methods

To analyze the wind power generation data, we propose constructing a SARIMA (Seasonal Autoregressive Integrated Moving Average) model. The first step involves examining the weekly average power generation from January 2017 to December 2020. The dataset is divided into two parts: a training set (2017–2021) for model fitting and a test set (2021-2022) for validating model performance.

Before building the model, we will explore the time series data to identify any seasonal patterns, such as periods when the wind turbines generate the most power and when they produce the least. To proceed, the time series must satisfy the conditions for (weak) stationarity:

- The mean value function, μ_t, remains constant over time.

- The autocovariance function, $\gamma(s,t)$, depends only on the distance $|s-t|$ between points, not on their specific location in time.

We will also examine the relationship between the target variable, Power, and other variables to determine which factors might contribute significantly to the model. Lastly, analyzing the autocorrelation and partial autocorrelation functions will help decide whether to use an ARIMA or SARIMA model and determine the appropriate parameters for the model.

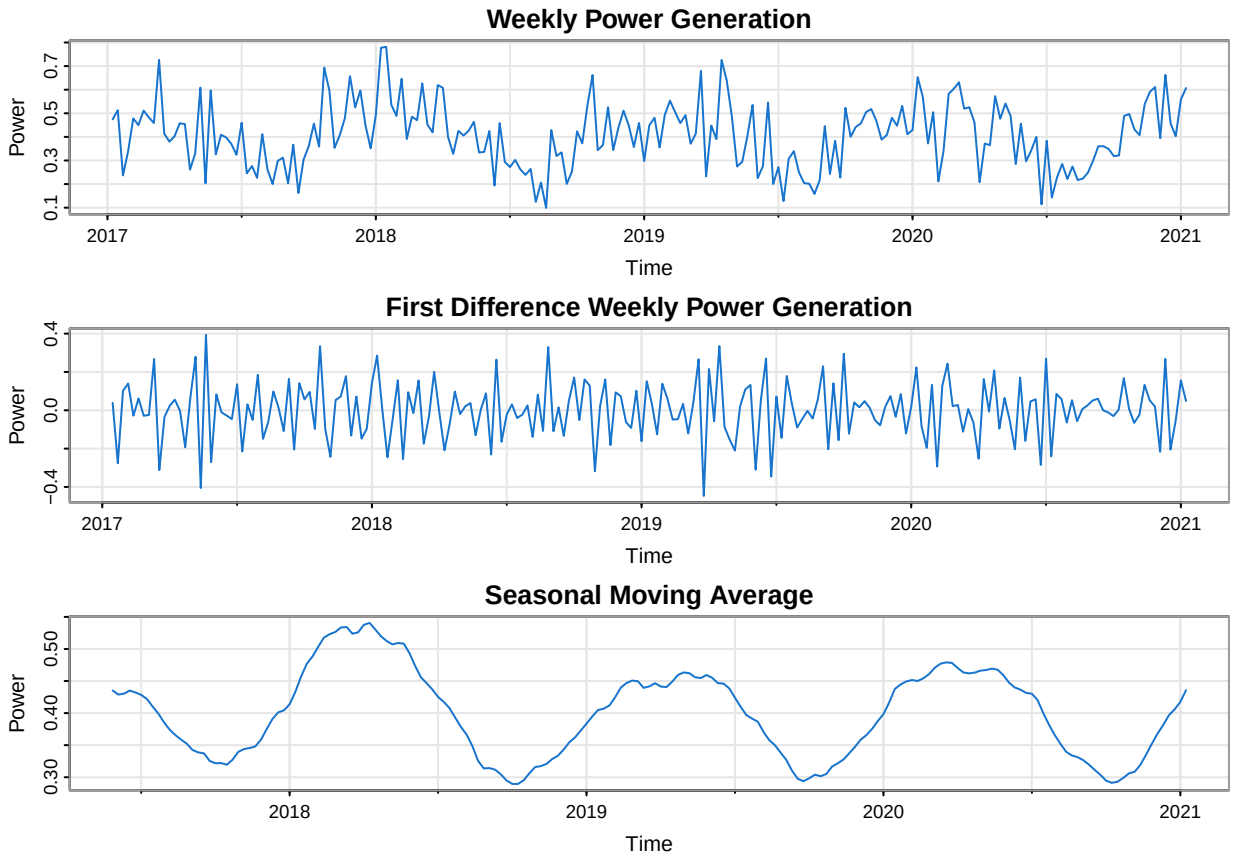

Figure 1: Analysis of Weekly Power Generation from 2017-2021 (Weekly, First Difference, Seasonal Moving Average)

According to Figure 1, the Weekly Power Generation chart reveals a distinct seasonal pattern, with peak power generation occurring in the early part of the year and the lowest levels observed mid-year. This suggests that wind turbines generate the most electricity during winter and the least during summer, aligning with the natural tendency for stronger winds in winter months, which boost turbine efficiency. Additionally, the Weekly Power Generation data is not homogeneous (non-stationary), prompting us to take the first difference to stabilize the data. After differencing, the data appears more homogeneous and stationary as seen in Figure 1.

Further analysis of the seasonal moving average highlights a wave-like pattern, reinforcing the need to incorporate seasonal components in our SARIMA model. These observations

will inform the development of a robust and efficient model tailored to accurately forecast wind power generation. This approach ensures the model captures both seasonal trends and short-term fluctuations, providing actionable insights into renewable energy efficiency.

Next, we aim to analyze the relationship between the target variable, Power, and the other variables. According to Figure 2, there is a strong positive correlation between Power and wind speed at both 10 and 100 meters above the surface, as well as wind gusts at 10 meters above the surface. However, wind speed at 10 meters and 100 meters are likely to be strongly correlated with each other, and including both variables in the model could introduce collinearity. Similarly, wind gusts are also related to wind speed. Therefore, from these three variables, we will likely include only one to effectively summarize their collective impact.

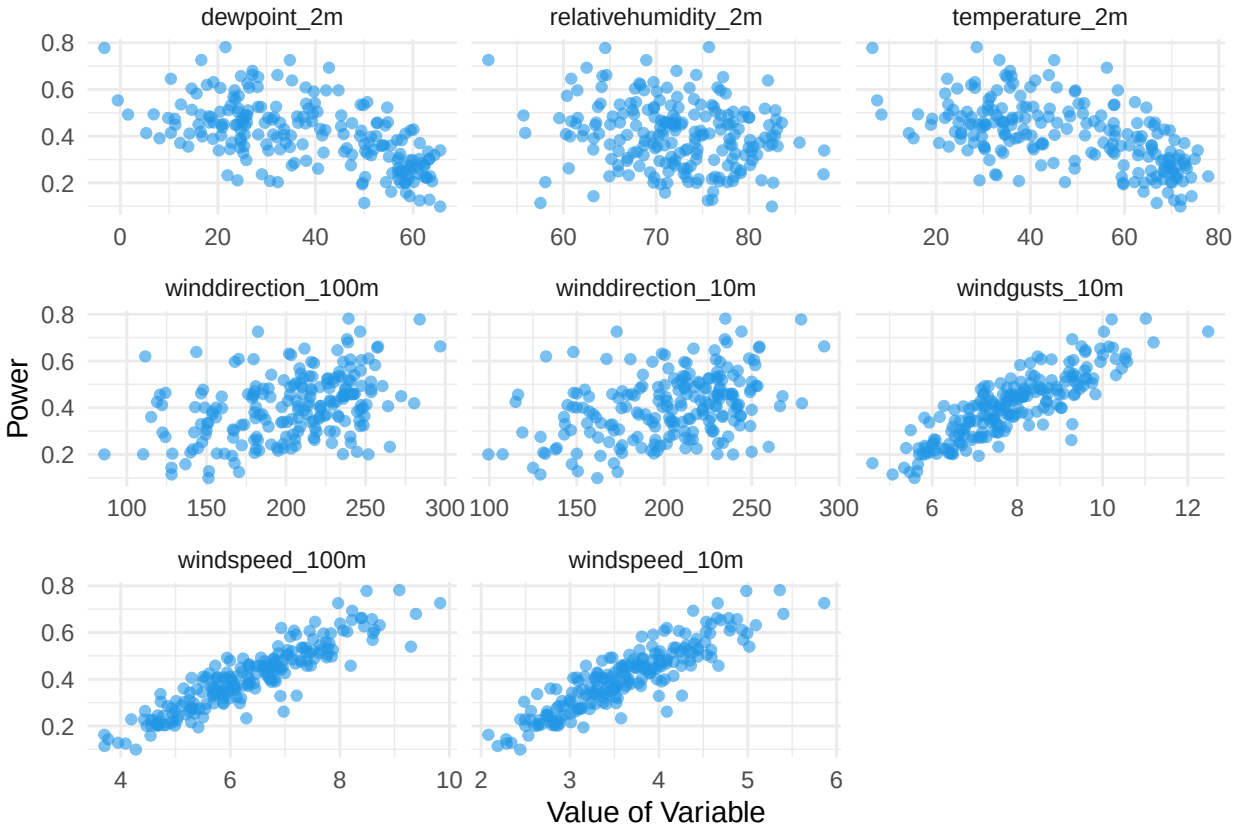

Figure 2: Relationship between Power and Other Variables

Based on Figure 3, we can see that the ACF is cutoff at lag 1 and PACF is tailing off. This suggests that we build ARIMA(0,1,1) model. From Figure 1, we can see that there are repeating patterns every year (52 weeks), which suggests that incorporating seasonal term could be beneficial to the model.

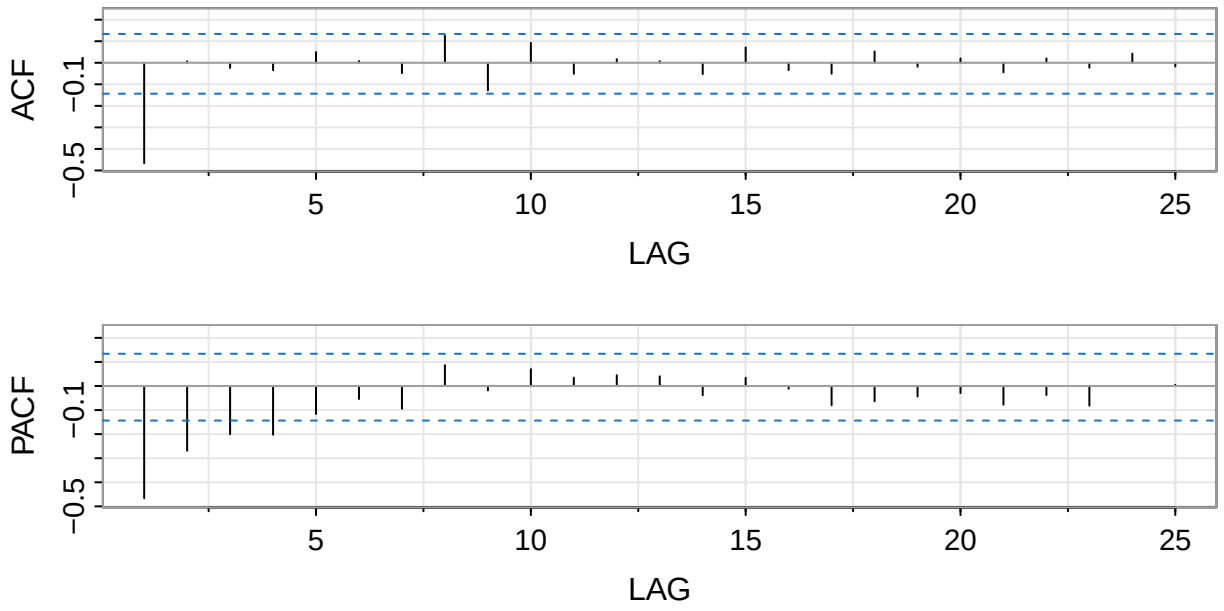

Figure 3: ACF and PACF of First Difference Weekly Power Generation

We have a test set containing data from 2021-2022, and we can see that in Figure 4 shows a similar trend to our training set, with the highest power generated in the winter and the lowest in the summer. Therefore, we seek a model that captures and aligns with this seasonal trend observed in the training set to provide accurate forecasts for the test period. We will consider four models shown in Table 2.

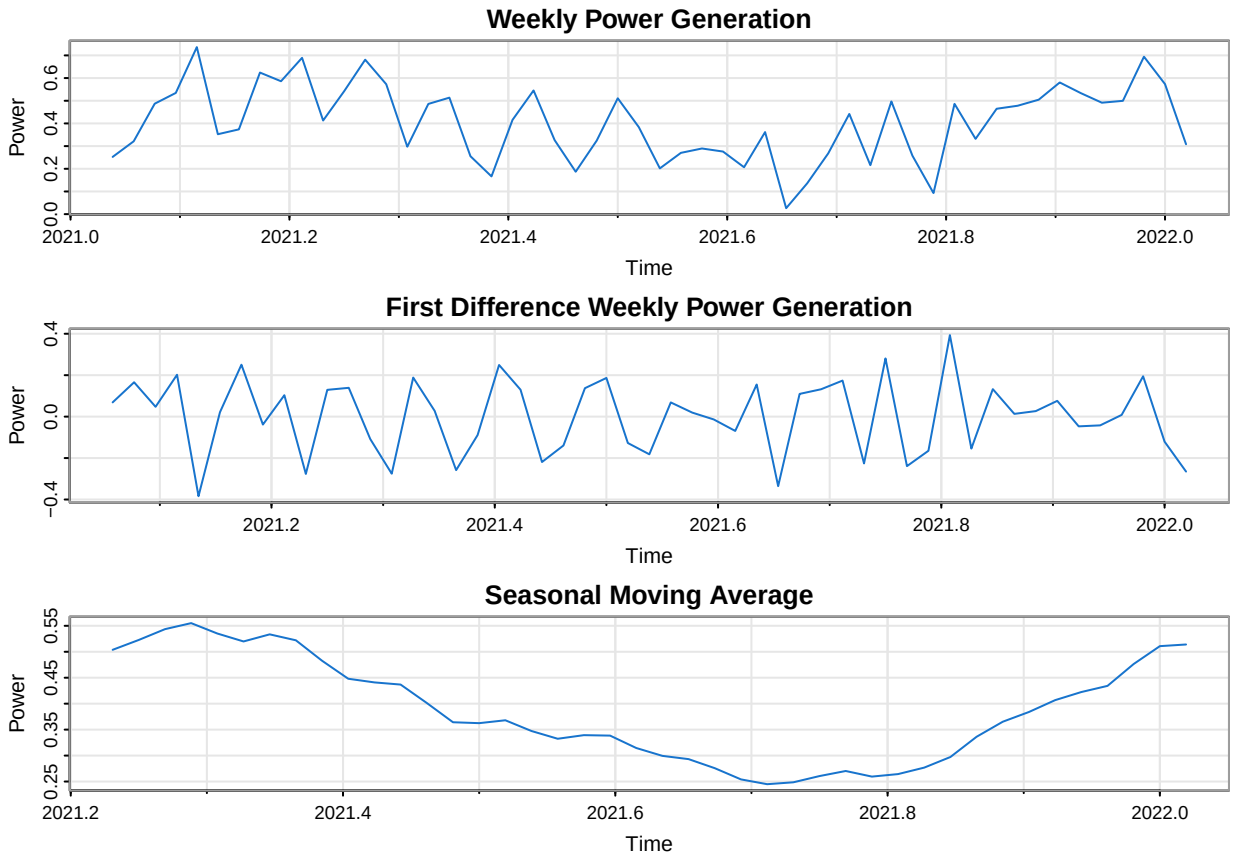

Figure 4: Analysis of Weekly Power Generation from 2021-2022 (Weekly, First Difference, Seasonal Moving Average)

Model	.
1	ARIMA(0,1,1)
2	ARIMA(0,1,1) with external regressors
3	SARIMA(0,1,1)x(0,1,1)_52
4	SARIMA(0,1,1)x(0,1,1)_52 with external regressors

Table 2: Models

We aim to conduct spectral analysis to investigate whether the spectral density provides insights into the cyclical nature of wind power generation data from 2017 to 2020. As shown in Figure 5, the most prominent spectral peak occurs at a frequency of approximately $\frac{0.9629630}{52}$, indicating a cycle duration of roughly 54 weeks. This observation aligns well with the seasonal patterns in our data, which exhibit the highest power generation during the winter months and the lowest in the summer. The seasonal variation captured by the spectral analysis further supports the idea that the wind power generation cycle is closely tied to the annual seasonal cycle.

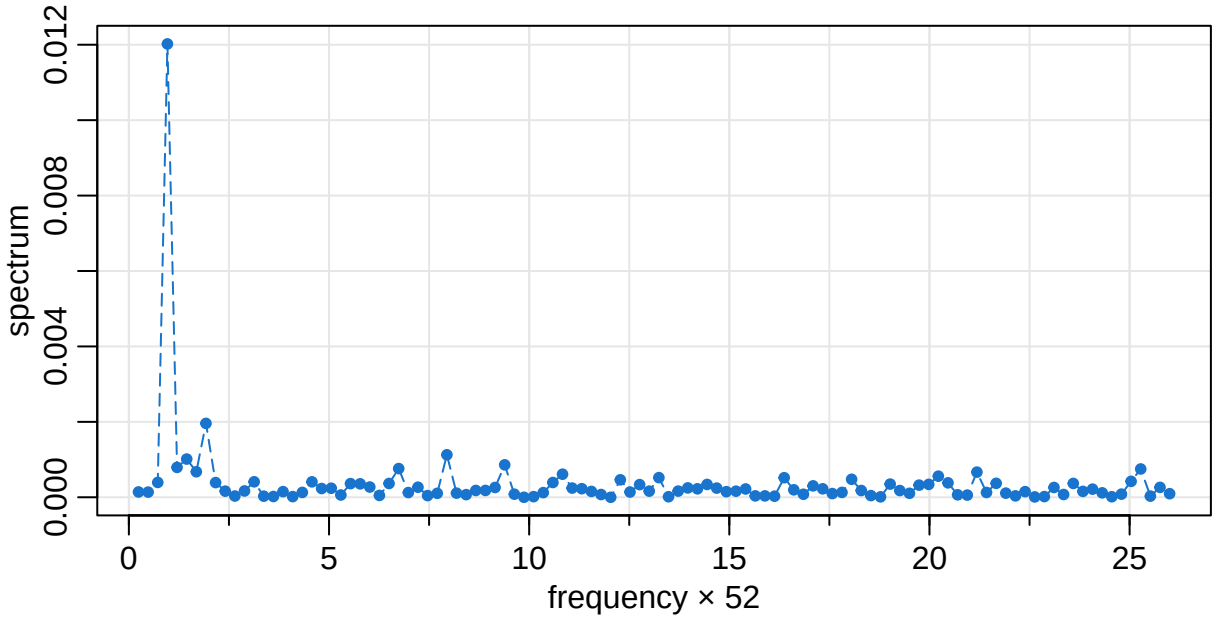

Figure 5: Spectral Density of Wind Power Generated from 2017-2021

We now want to perform a diagnostic analysis of the models by reviewing their summaries. This will allow us to assess key statistical measures, such as the significance of the model parameters, the goodness-of-fit metrics (e.g., AIC, BIC), and the residuals. By examining these factors, we can identify potential issues like overfitting, underfitting, or the presence of autocorrelation in the residuals. The model summary will provide valuable insights into the reliability and performance of each model, helping us determine which one is best suited for accurate forecasting.

```
## <><><><><><><><><><><><><><><><><><>
##
## Coefficients:
##     Estimate     SE  t.value p.value
## ma1  -0.7122 0.0458 -15.5562       0
##
## sigma^2 estimated as 0.01500357 on 207 degrees of freedom
##
## AIC = -1.338957  AICc = -1.338864  BIC = -1.306866
##
```

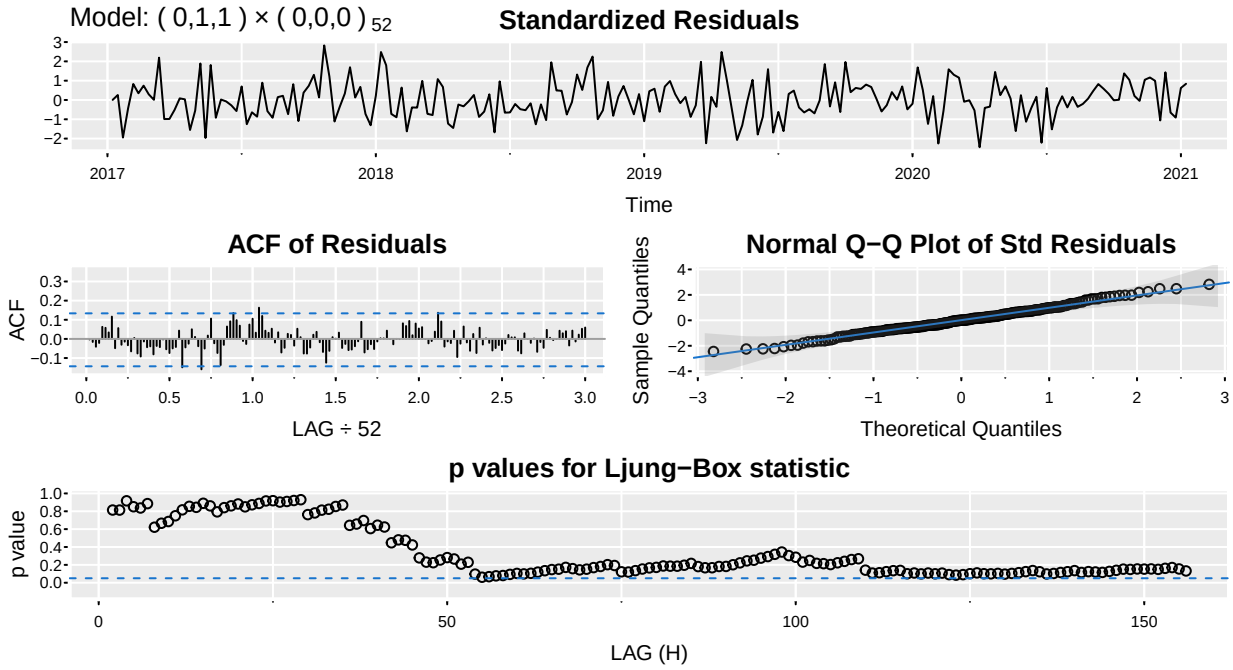

Figure 6: Summary of ARIMA$(0, 1, 1)$ Model

The Model 1 we examine is the ARIMA(0,1,1). According to Figure 6, the MA variable is highly significant, with a p-value of 0, which is less than 0.05. From the diagnostic plots, the standardized residuals appear to be white noise, as all lags have p-values greater than 0.05 in the Ljung-Box statistic. Additionally, the residuals appear to follow a normal distribution, as none fall outside the theoretical quantile range. However, the original time series data exhibits a clear seasonal pattern each year, suggesting that a seasonal term might be worth considering in our time series model. Moreover, it would also be valuable to explore whether external regressors could significantly enhance the model.

For the Model 2, we use the same base ARIMA(0,1,1) structure but include the variables windspeed_100m and dewpoint_2m. As shown in Figure 7, the AIC is significantly lower than Model 1, indicating that the inclusion of these external regressors improves the model's forecasting capability. Additionally, all variables are highly significant. The diagnostic plots confirm that the residuals are white noise, with p-values well above 0.05, and they appear normally distributed. Therefore, the model incorporating external variables outperforms the first model.

```
## <><><><><><><><><><><><><><><><><><>
##
## Coefficients:
##       Estimate     SE  t.value p.value
## ma1    -0.9043 0.0325 -27.8547  0.0000
## xreg1   0.0982 0.0038  25.7140  0.0000
## xreg2  -0.0008 0.0003  -2.6072  0.0098
```

```
##
## sigma^2 estimated as 0.00309082 on 205 degrees of freedom
##
## AIC = -2.894794  AICc = -2.894229  BIC = -2.830611
##
```

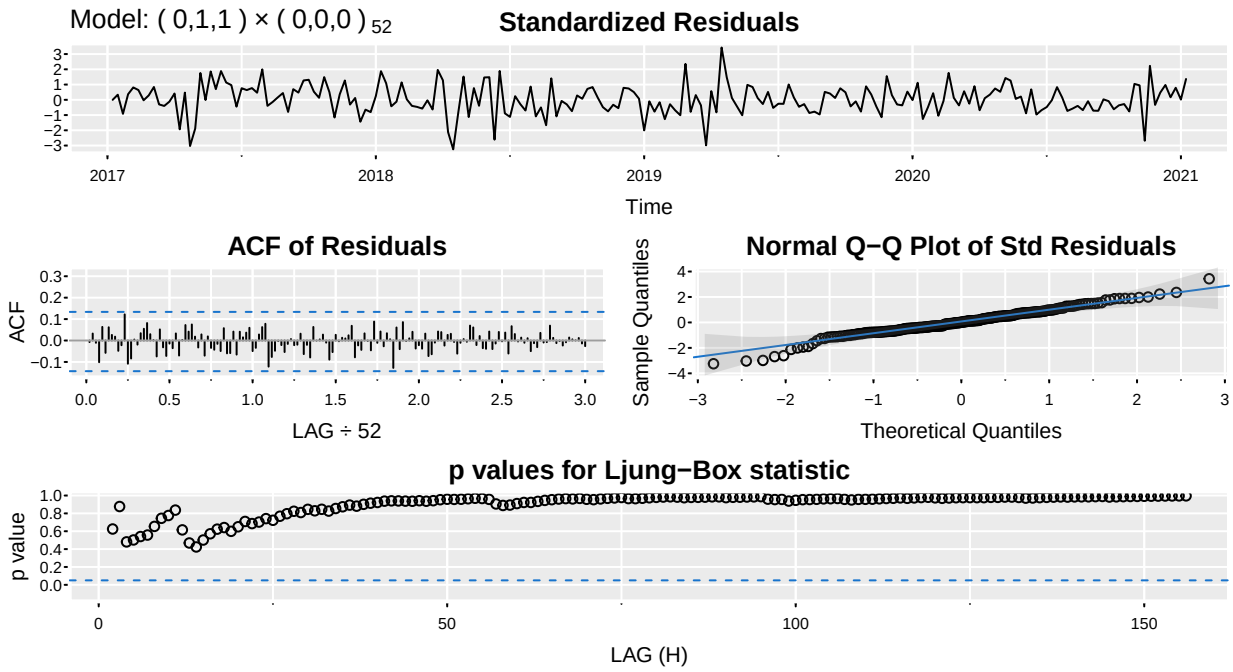

Figure 7: Summary of ARIMA$(0, 1, 1)$ with External Regressors

However, we know that there's a seasonal pattern in the original time series as the high power generated in the winter and the lowest in the summer. Therefore, we will consider incorporate seasonal parameter without any regression terms.

```
## <><><><><><><><><><><><><><><>
##
## Coefficients:
##       Estimate     SE  t.value p.value
## ma1    -0.9645 0.0423 -22.7747  0.0000
## sma1   -0.9952 0.5157  -1.9297  0.0555
##
## sigma^2 estimated as 0.0112419 on 154 degrees of freedom
##
## AIC = -1.136412  AICc = -1.135909  BIC = -1.077761
##
```

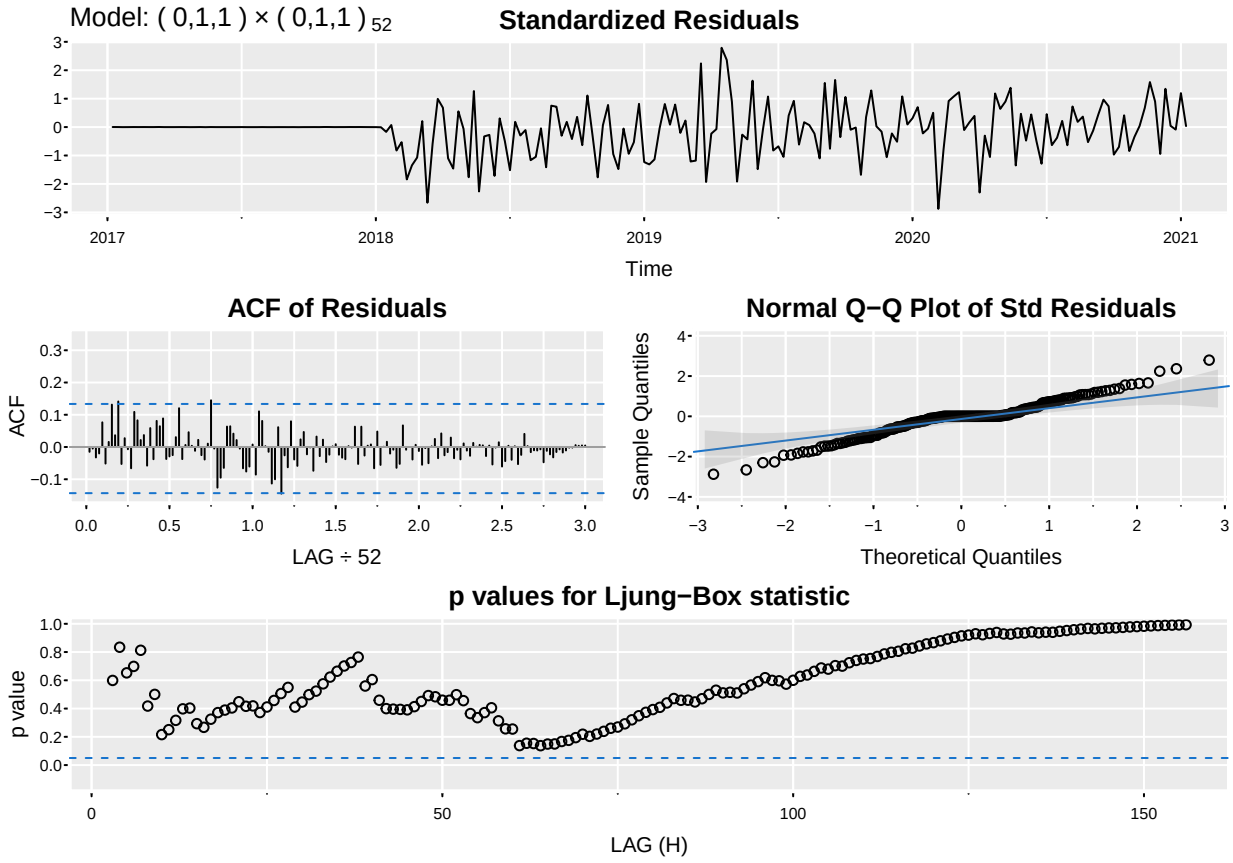

Figure 8: Summary of SARIMA$(0,1,1) \times (0,1,1)_{52}$

According to Figure 8, Model 3 with the seasonal parameter appears to have limited significance, as indicated by a p-value of 0.0555, which slightly exceeds the conventional threshold of 0.05. Additionally, incorporating the seasonal parameter affects the model's residuals, as the standardized residuals deviate from normality, with some observations falling outside the theoretical quantile range. Moreover, the residuals appears less white noise compared to Model 2, suggesting that the seasonal parameter may not improve the model's overall performance substantially. Now, we want to check if the seasonal parameter is significant with external regressers.

```
## <><><><><><><><><><><><><><><>
##
## Coefficients:
##       Estimate     SE  t.value p.value
## ma1    -0.8929 0.0381 -23.4370  0.0000
## sma1   -0.9699 2.1457  -0.4520  0.6519
## xreg1   0.0991 0.0051  19.5985  0.0000
## xreg2  -0.0015 0.0008  -1.9861  0.0488
##
## sigma^2 estimated as 0.003390578 on 152 degrees of freedom
```

```
##
## AIC = -2.342526  AICc = -2.340828  BIC = -2.244774
##
```

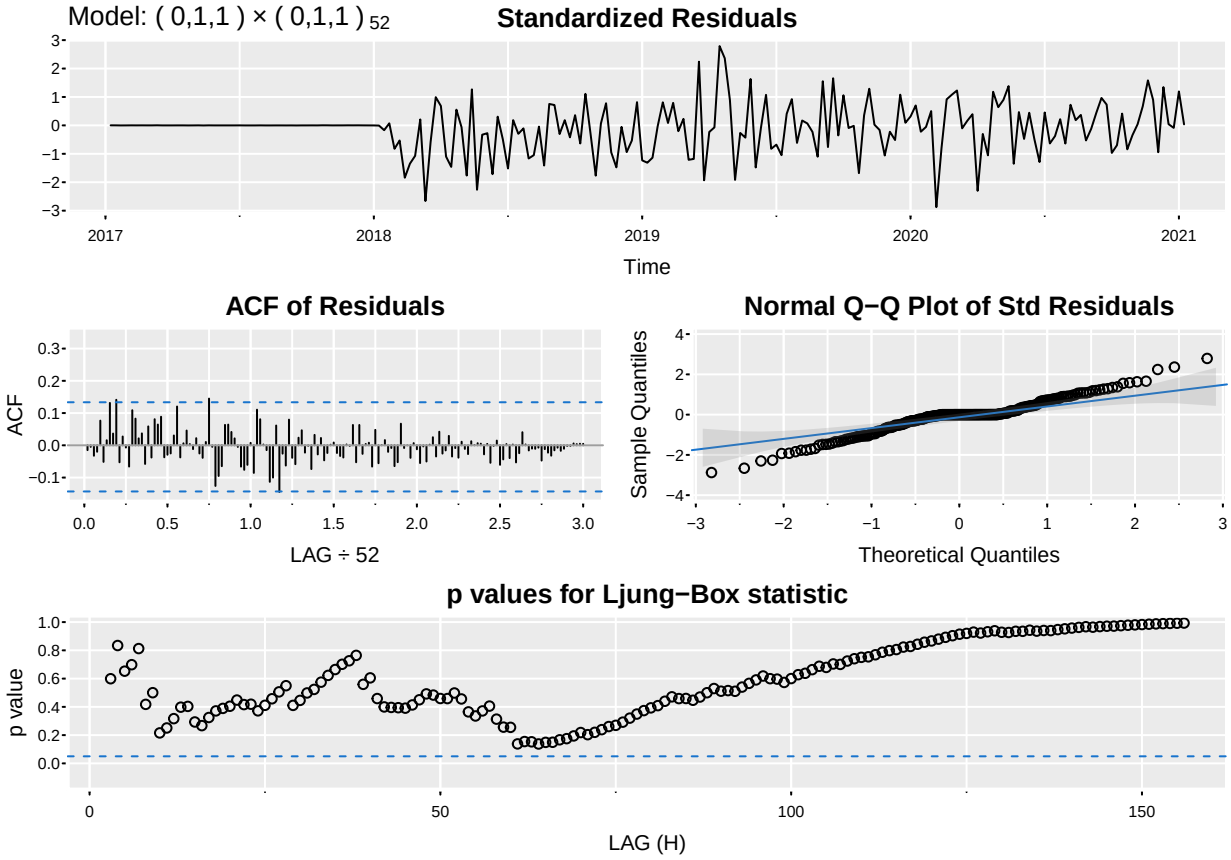

Figure 9: Summary of SARIMA$(0,1,1) \times (0,1,1)_{52}$ with External Regressors

Based on Figure 9, the seasonal parameter remains insignificant, with a p-value of 0.6519, which is much higher than the threshold of 0.05. This strongly suggests that the seasonal parameter should not be included in the model. However, both external regressors, windspeed_100m and dewpoint_2m, are significant contributors. Despite these adjustments, the residuals still deviate from normality, indicating that the model should include only one MA parameter alongside the external regressors.

Overall, after diagnosing four models, Model 2 demonstrates the most promise for predictive analytics of future power generation, which has the lower AIC. Next, we will evaluate the predictive capabilities of these models to determine which one is best suited for forecasting.

Results

First, we will look at the models' plot on predicted and observed power. Then, we will look at the forecasting and MSE (Mean Squared Errors). From these results, we will get to conclude on which model is best suited for forecasting.

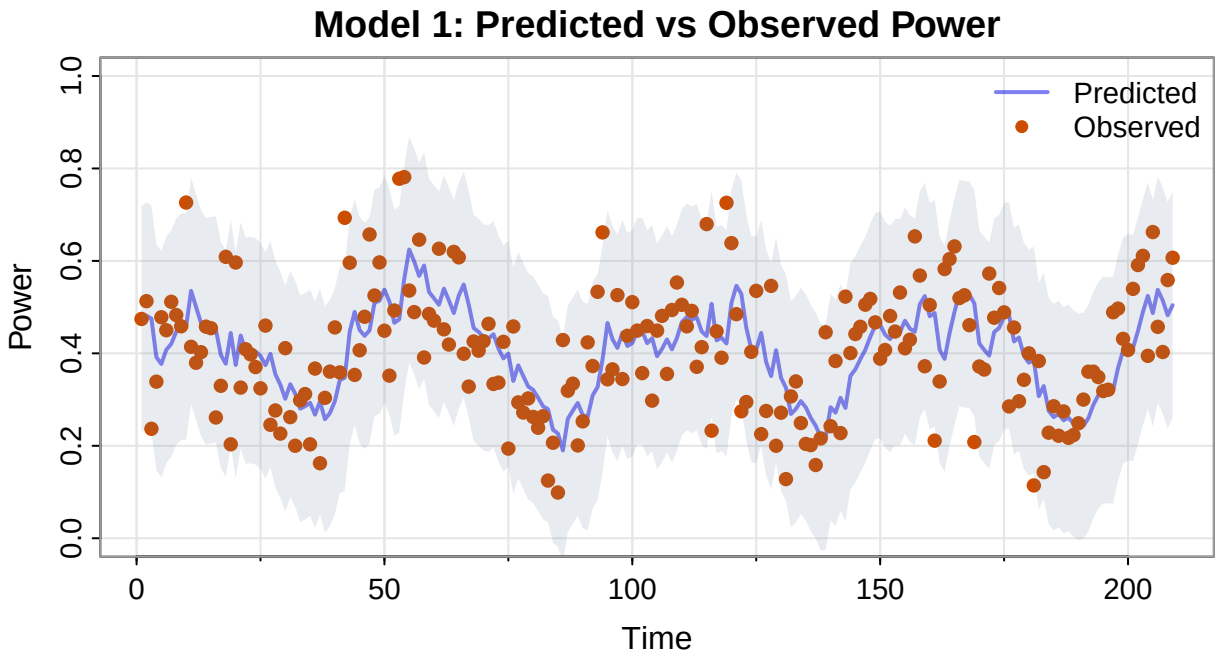

Figure 10: Model 1 Predicted and Observed Power Plot

According to Figure 10, the model effectively captures the overall trend. However, Model 1 shows signs of underfitting, as it struggles to accurately represent data points that deviate significantly from the mean Power. This limitation may affect its ability to forecast power values outside the range of the mean. Additionally, the model exhibits a large prediction error. Next, we will evaluate whether Model 2, incorporating external regressors, provides a better fit to the observed power by capturing deviations from the mean and reducing the prediction error.

Based on Figure 11, Model 2 shows a stronger ability to capture data that deviates from the mean Power. Although there is some concern about potential overfitting, the model appears more promising for forecasting overall. Additionally, the prediction error is significantly smaller compared to Model 1, further supporting its suitability.

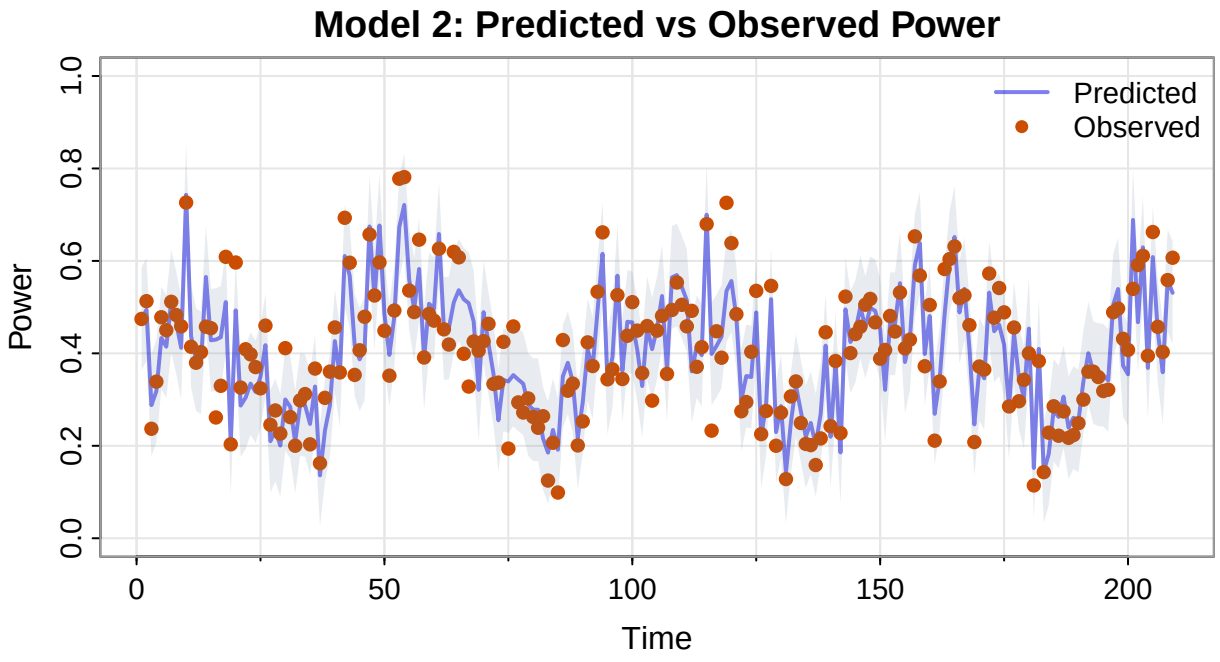

Figure 11: Model 2 Predicted and Observed Power Plot

Now, we want to explore Model 3 with seasonal component. According to Figure 12, Model 3 outperforms Model 1, but its effectiveness diminishes over time as it struggles to capture observations that deviate from the mean Power. This suggests that the model is more suited for fitting recent trends, which may result in weaker forecasts. Additionally, the prediction error is higher than that of Model 2, and the area of prediction error increases as time progresses. This indicates that Model 2 provides a better overall fit than Model 3, which means that the seasonal parameter should not be incorporate into the model.

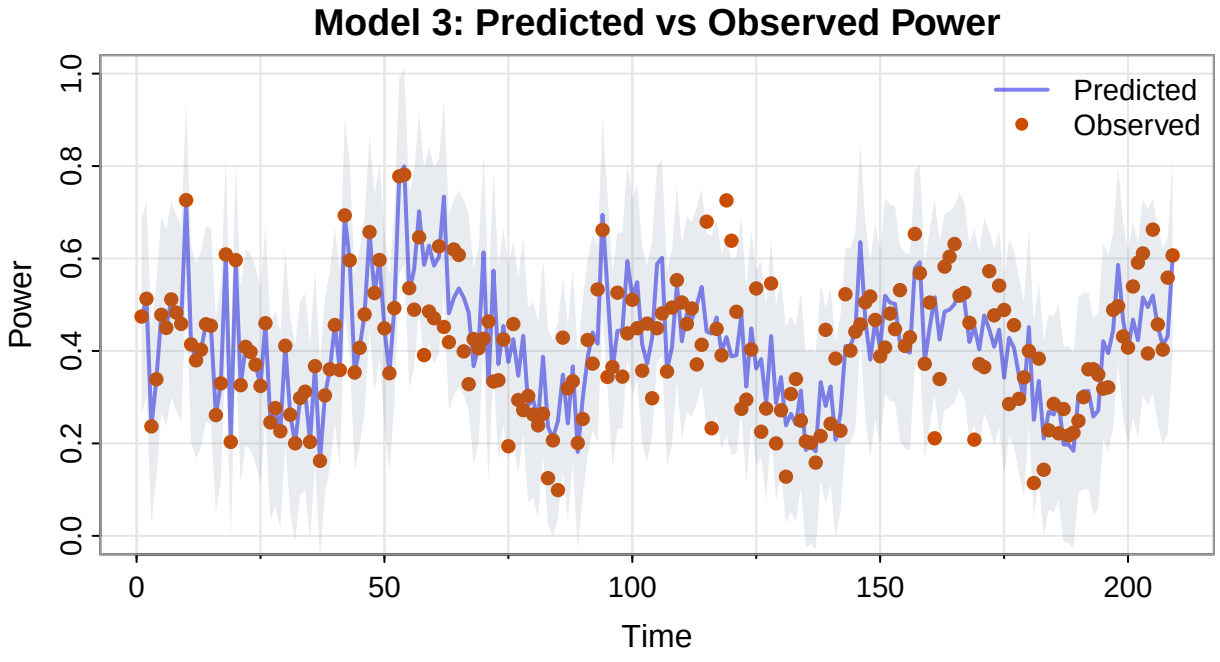

Figure 12: Model 3 Predicted and Observed Power Plot

Now, we want to know whether the external regressors can contribute to the model with seasonal component. Based on Figure 13, the model shows improvement over Model 3, with a lower area of prediction error and consistent performance in fitting the observations over time. However, the model appears to capture more information, including some outlier observations, raising concerns about potential overfitting, and the area of prediction error is large than Model 2.

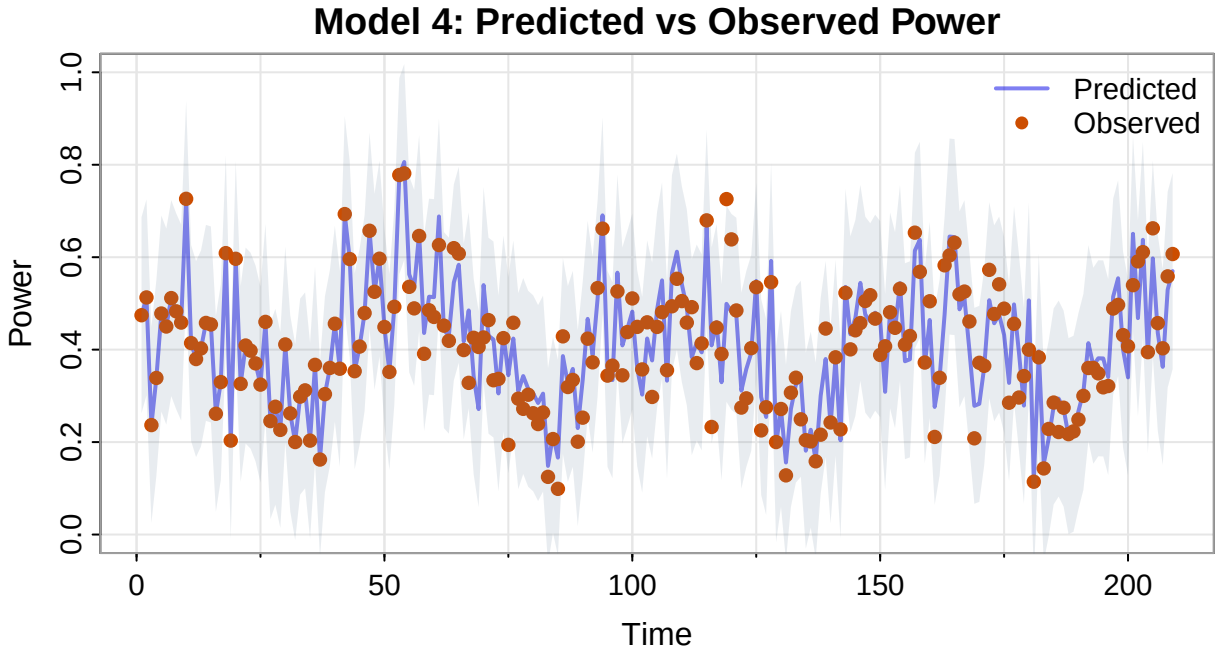

Figure 13: Model 4 Predicted and Observed Powers Plot

In summary, Model 2 performs the best at fitting the observations, with a low prediction error and a balanced fit that avoids both overfitting and underfitting. In contrast, Model 1 shows underfitting, Model 3 shows inconsistency in fitting over time, and Model 4 displays high variance in its fit.

We will now assess the forecasting performance of these models by examining their forecast plots with a 52-step ahead forecast. This analysis will use the 2021-2022 test set to evaluate how well each model predicts future power generation and how accurately they capture the trends over time. By comparing the forecasted values with the observed data, we can identify any discrepancies, patterns, or trends that highlight the strengths or weaknesses of each model. This evaluation will provide valuable insights into which model is most effective at forecasting future power generation.

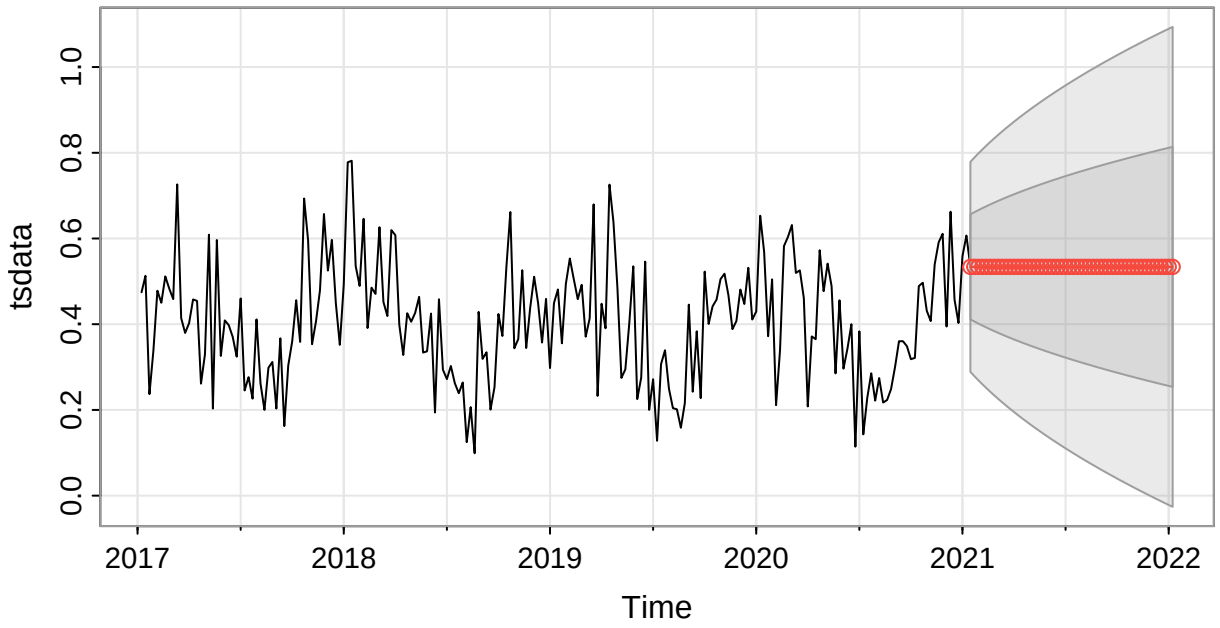

Figure 14: Forecasting Results for Model 1

According to Figure 14, the forecasting performance of Model 1 is poor, with the forecasted values forming a flat line around the mean. Additionally, the prediction error is quite large. This is consistent with the findings in Figure 10, where the model's predictions fluctuate slightly around the mean, causing it to forecast consistently at the mean level. Despite Model 1's summary indicating a significant MA parameter and a low AIC, its forecasting ability is insufficient. Therefore, Model 1 should not be relied upon for accurate predictions of future power generation.

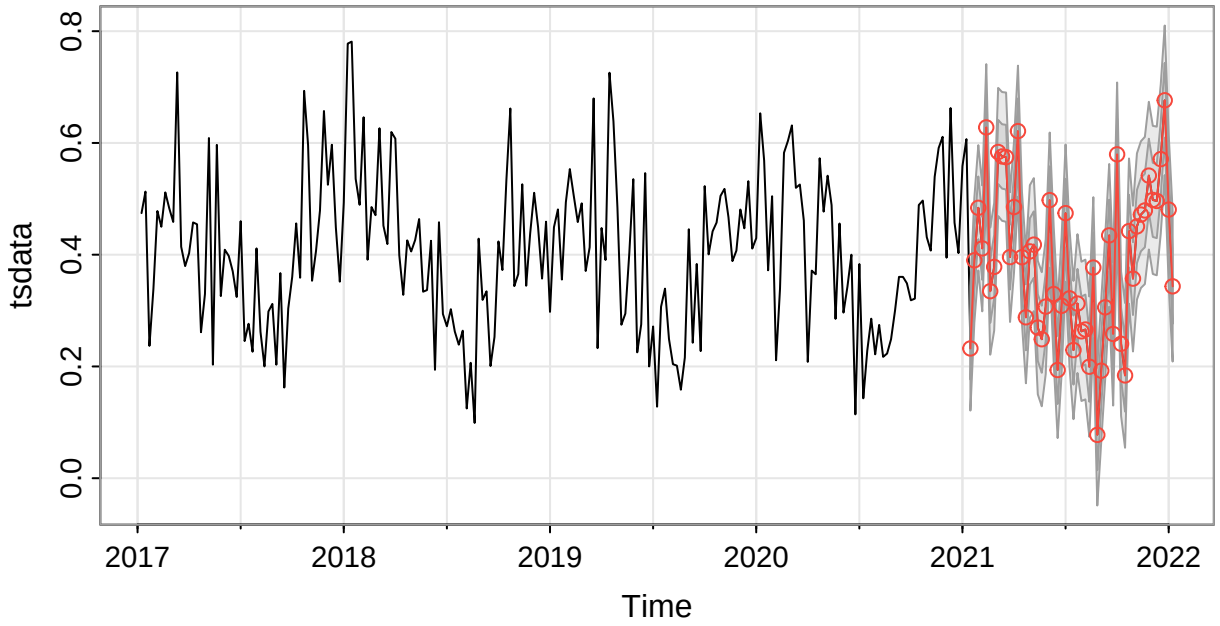

Figure 15: Forecasting Results for Model 2

Now, we will evaluate the forecast performance of Model 2. As shown in Figure 15, the forecast shows a significant improvement over Model 1, with a more realistic forecast and smaller prediction error. Additionally, the forecast closely follows the trends observed in the 2017-2020 period, further suggesting that Model 2 is more effective at capturing the underlying patterns and providing accurate predictions.

Next, we will examine the forecast results of the model with the seasonal parameter. As shown in Figure 16, the forecast appears reasonable overall, but it fails to capture observations that deviate significantly from the mean. This aligns with the findings in Figure 12, where the model's ability to capture information diminishes over time. As a result, the model produces a high prediction error and a weaker forecast.

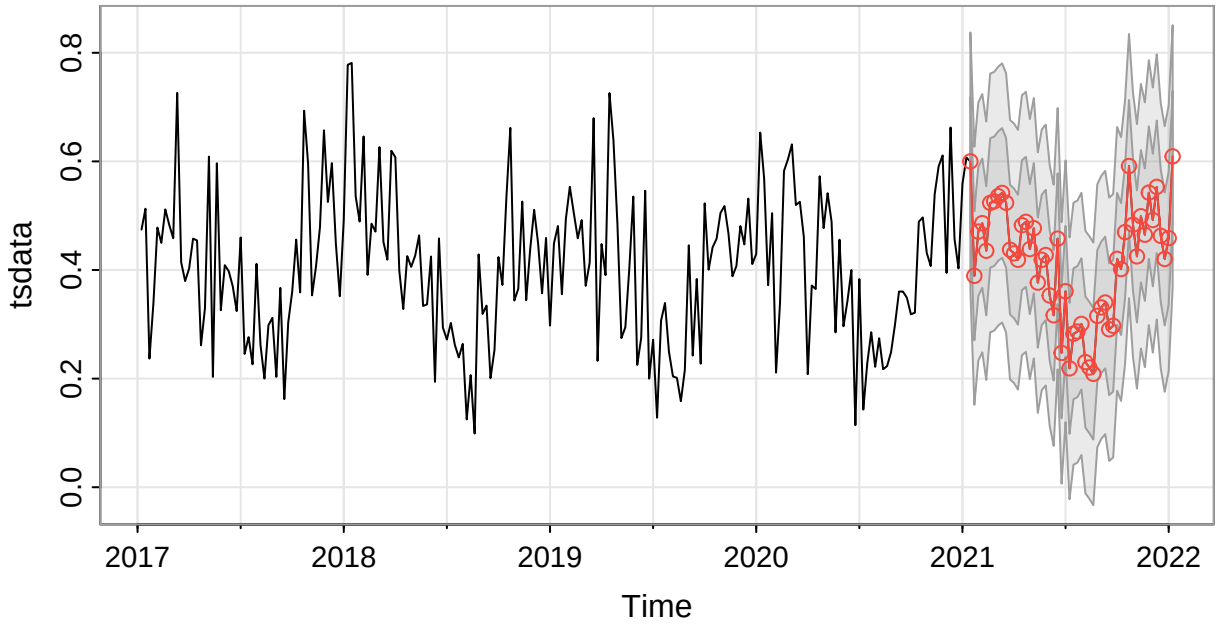

Figure 16: Forecasting Results for Model 3

Thus, we want to assess whether including external regressors with the seasonal parameter leads to a better forecast. As shown in Figure 17, the forecast is better than Model 3 and closely resembling the forecast from Model 2. This suggests that both Model 2 and Model 4 are effective for forecasting, but Model 2 offers a better goodness-of-fit, as indicated by its lower AIC and the significance of the parameters.

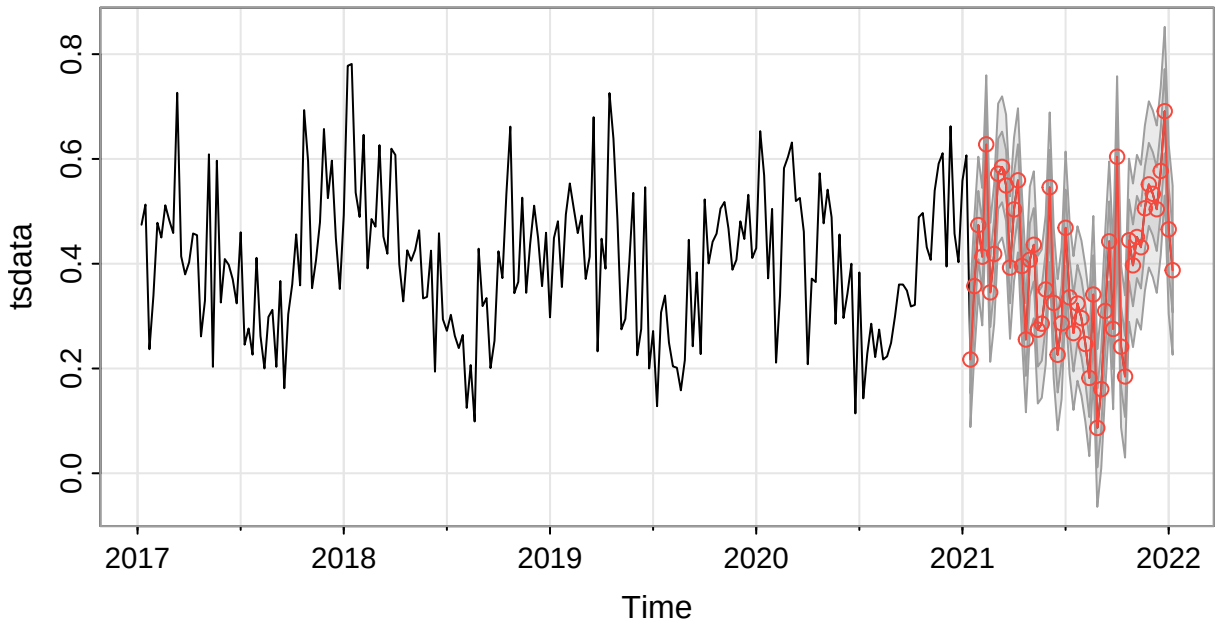

Figure 17: Forecasting Results for Model 4

Lastly, we evaluate the models by comparing their MSE with the test set data from 2021-2022. According to Table 3, Model 2 has the lowest MSE of 0.0034, indicating it performs the best in forecasting. Model 4 follows closely with an MSE of 0.0041, showing similar forecasting accuracy. In contrast, Model 1 has the highest MSE of 0.0435, suggesting it should not be used for forecasting. Model 3 has a lower MSE of 0.0243, which is better than Model 1 but still indicates that it is not ideal for forecasting. Overall, Model 2 stands out as the best model for forecasting, though both Model 2 and Model 4 show promising results.

Model	MSE
ARIMA(0, 1, 1)	0.0435
ARIMA(0, 1, 1) with External Regressors	0.0034
SARIMA(0, 1, 1)x(0, 1, 1)_52	0.0243
SARIMA(0, 1, 1)x(0, 1, 1)_52 with External Regressors	0.0041

Table 3: Mean Squared Error of Models

Discussions

In this project, we identified key external factors influencing wind power generation, such as wind speed and dew point, and analyzed the data patterns. However, the specific location of the dataset remains unknown, which limits our analysis. A potential expansion of this project could involve examining how the location of wind turbines impacts the trends in power generation.

For example, Texas, with the highest number of wind turbines and the greatest installed capacity, presents a different climate compared to Iowa, which ranks second in turbine count and is typically colder. This raises intriguing questions: Will the seasonal patterns observed in Iowa align with those in Texas? How do these differences affect the forecasting ability of the model? Incorporating geographic variability could provide deeper insights of the renewable energy efficiency, offering a more comprehensive understanding of location-based trends in wind power generation.

Conclusions

In conclusion, we found that the Model 2 is the best model at forecasting (ARIMA(0,1,1) with external regressors) with the lowest MSE. Throughout this project, we analyzed the characteristics of the data and we found out that there's a clear pattern every year, where the wind turbine produces the highest power in the winter and the lowest power in the summer. Additionally, the spectral density aligns the data's pattern with the most prominent peak occurs about $\frac{0.9629630}{52}$, which means that it completes it's cycle approximately 54 weeks.

The goal of this project is to determine whether external variables can enhance the forecasting ability of the model. We found that including the external variables, dewpoint_2m and windspeed_100m, significantly improves the model's prediction accuracy and its fit to the observed data. Notably, there is a marked difference between Model 1 (ARIMA(0,1,1) without external regressors) and Model 2. Model 1 struggles to capture observations that deviate from the mean, whereas adding the external regressors in Model 2 strengthens the model's ability to capture such deviations, resulting in a more accurate and robust prediction. We also tested the impact of a seasonal parameter, and found that Model 3 (with the seasonal parameter) performed worse than the model without it. However, Model 4 (with both the seasonal parameter and external regressors) showed significant improvement over Model 3, providing better consistency in fitting the data over time, while the fit of Model 3 diminished. After reviewing the forecast plot and MSE for all four models, we conclude that Model 2 and Model 4 are the best for forecasting. This indicates that including the external regressors, dewpoint_2m and windspeed_100m, is crucial for improving the model's performance.

References

[1] U.S. Department of Energy. (n.d.). Advantages and challenges of wind energy. U.S. Department of Energy. Retrieved December 4, 2024, from https://www.energy.gov/eere/wind/advantages-and-challenges-wind-energy

[2] Rahim, M. (2021). Wind power generation data forecasting. Kaggle. Retrieved December 4, 2024, from https://www.kaggle.com/datasets/mubashirrahim/wind-power-generation-data-forecasting/data