**Abstract**

This report investigates the solar energy consumption trends from 2001 to 2024, utilizing a diverse set of statistical models to analyze and forecast energy utilization. By employing linear regression, exponential smoothing, and ARIMA models, the analysis elucidates underlying patterns, seasonal variations, and predicts future consumption trends. Each method's efficacy in handling the complexity of energy data, marked by periodicity and trend components, is critically assessed. The findings highlight the strengths and limitations of each methodological approach, offering insights into the most accurate predictive models for solar energy consumption. The report concludes with a recommendation on the optimal modeling approach, advocating for a mixed-model strategy to enhance forecast accuracy and reliability.

**Introduction**

With global emphasis on sustainable energy sources, understanding and predicting solar energy consumption has become crucial for planning and strategic decision-making. This report presents a comprehensive analysis of solar energy consumption data from the U.S. Energy Information Administration, spanning over two decades. It aims to identify significant consumption trends, evaluate the impact of seasonal variations, and develop reliable forecasts using various statistical techniques. The methods employed include linear regression to ascertain long-term trends, exponential smoothing for short-term forecasts, and ARIMA models to address both non-stationarity and seasonality in the data. The overarching goal is to determine which method or combination thereof provides the most robust predictions for future solar energy demands.

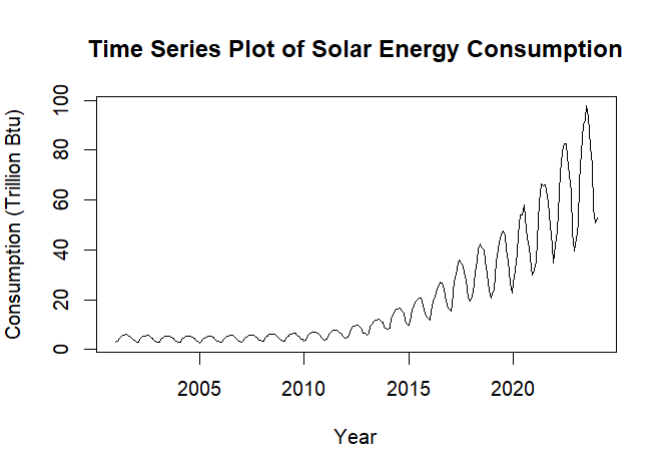
**Part 1 Linear Regression and Exponential Smoothing**

**Data collection Reliability**

The data used in this study was published by U.S Energy Information Administration

**Classification and tabulation of data**

The data here consists of monthly records of solar energy consumption from 2001 onwards, expressed in trillion British thermal units (Btu). For effective analysis, classification and tabulation were employed. The data was classified into years and months, allowing for an examination of trends, seasonal variations, and cyclic behavior over time. Tabulation involved arranging this classified data into structured columns for 'Month' and 'Solar Energy Consumption', facilitating easier visualization and statistical analysis. This structured format aids in identifying monthly and yearly consumption patterns and supports the calculation of growth rates, averages, and forecasting.



The Time Series Plot of Solar Energy Consumption shows the monthly consumption of solar energy measured in trillion British thermal units (Btu) from 2001 onwards. Here's the detailed interpretation of this plot:

Trends:

Overall Increase: There's a clear upward trend in solar energy consumption over the period shown. This suggests increasing utilization of solar energy resources over time, likely due to advancements in technology, decrease in costs, and increased environmental awareness.

Volatility: The plot also exhibits periods of volatility, where consumption spikes and then falls. These could correlate with seasonal variations, changes in economic conditions, or policy changes affecting energy usage.

Seasonality:

While this specific plot does not explicitly show seasonality (as it aggregates all months within each year), the sharp peaks and troughs might hint at seasonal influences. Typically, solar energy consumption would be higher in sunnier months due to more daylight hours available for generating solar power.

Growth Patterns:

The data starts relatively flat with minor fluctuations but starts to exhibit more pronounced growth from around 2010. This phase of accelerated growth could be due to increased investment and interest in renewable energy sources.From around 2015, the growth seems even more rapid, suggesting possible significant policy implementations or technological breakthroughs that made solar energy more accessible and efficient.

**CREATION OF TIME PLOTS AND INTERPRETATION (Seasonal, Subseries, Lag and ACF Plots)**

Seasonal Plot

Observations: Each line represents a year from 2010 to 2024, showing solar energy consumption by month. Consumption peaks in the summer months (May to August) every year, indicating a strong seasonal pattern where more solar energy is utilized during these months, likely due to longer daylight hours and more intense sunlight.

Insights: The successive increase in peaks over the years suggests a growing trend in solar energy usage during these months. The data for 2024 shows a significant leap, suggesting anticipated or projected increases in solar energy consumption.

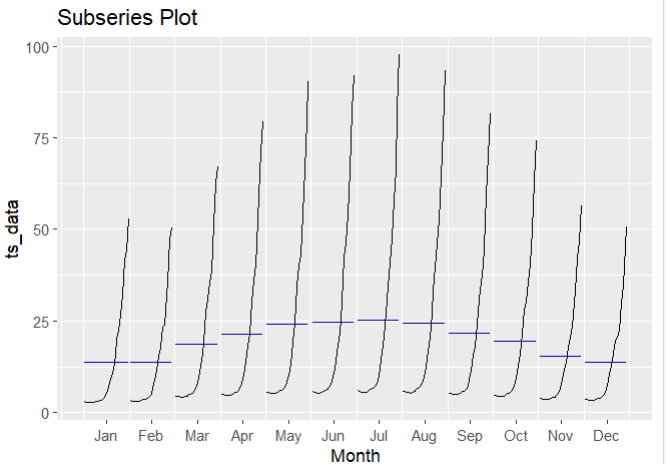
A graph of different colored lines

Description automatically generated

Subseries Plot

Observations: This plot breaks down the monthly data across all years, illustrating the cyclical nature of solar energy consumption within each year.

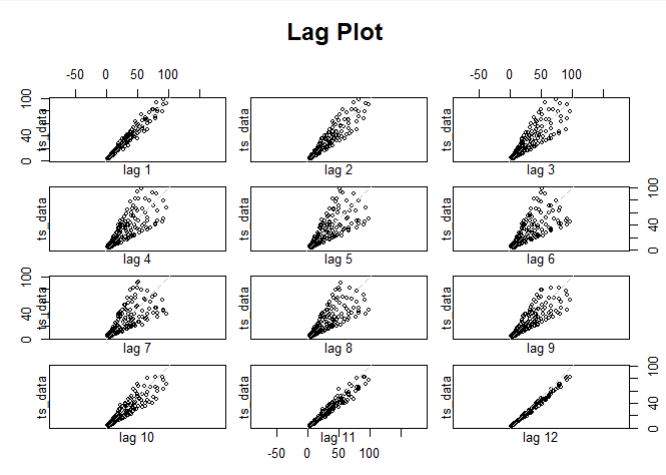
Insights: Consistent peaks during mid-year and troughs during the start and end of each year reinforce the seasonal influence observed in the Seasonal Plot. This visualization is particularly useful for confirming consistency in seasonal patterns across multiple years.



Lag Plot

Observations: Each subplot represents a different lag (from 1 to 12 months). The formation of a linear pattern, especially in the initial lags, suggests a strong correlation between observations and their immediate predecessors.

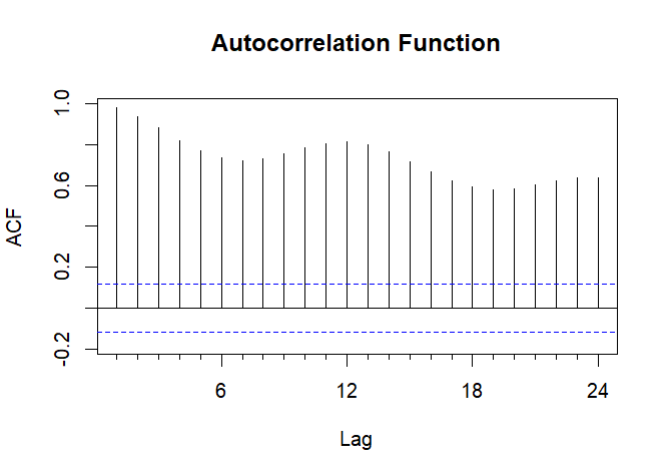
Insights: The clarity of the linear patterns in early lags indicates that past values are good predictors of future values, which is typical for time series data with trends and seasonality. The pattern diminishes as the lag increases, showing less correlation as the gap widens.



Autocorrelation Function (ACF)

Observations: The ACF plot shows significant positive spikes at seasonal intervals (at lags of 12, 24, etc.), which are well above the significance line.

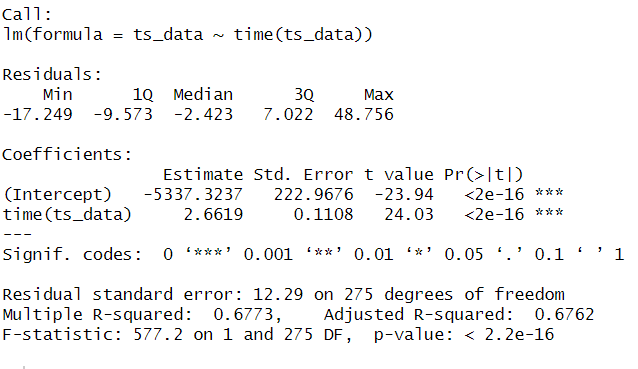
Insights: This confirms the strong seasonal component of the data, as the autocorrelation at yearly intervals remains high. This plot is essential for identifying the need for seasonal differencing when modeling the data.



Regression Model Output

Observations: The linear regression model has a significant positive slope and a high R-squared value (0.6762), indicating a good fit.

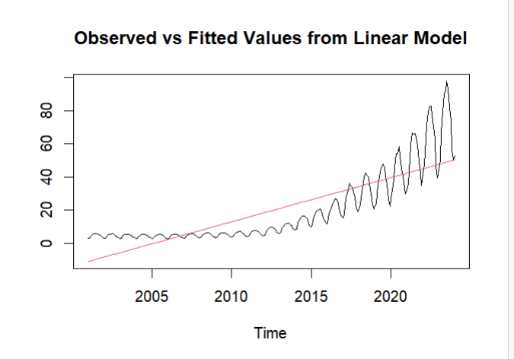
Insights: The model effectively captures the upward trend in solar energy consumption. The high coefficient for time suggests that as time progresses, the consumption significantly increases, which aligns with the trends observed in the plots.



Observed vs Fitted Values from Linear Model

Observations: The fitted line shows the general upward trend but does not capture the seasonal fluctuations.

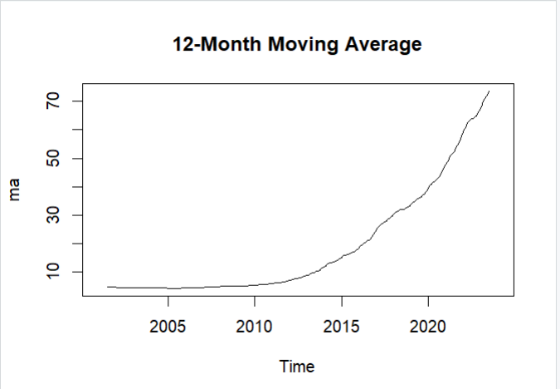
Insights: While the linear model is good for understanding the overall direction and pace of change in solar energy consumption, it does not address the cyclical nature within each year.



12-Month Moving Average

Observations: Shows a smoothed curve that rises more steeply in later years.

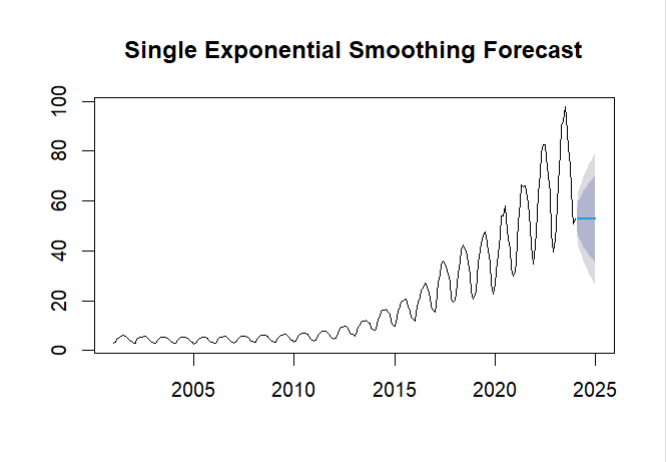
Insights: This plot emphasizes the increasing trend by smoothing out seasonal variations and random fluctuations, making the long-term trend more apparent.



Single Exponential Smoothing Forecast

Observations: This forecast model adapts more responsively to recent changes in the data, shown by the fit and projection.

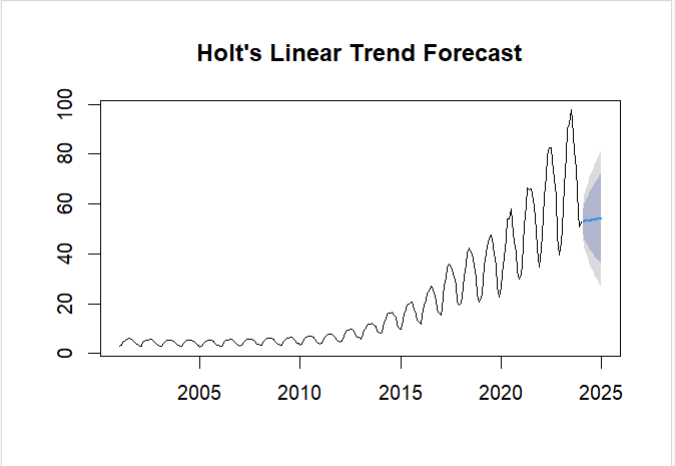
Insights: It provides a smoothed forecast that highlights short-term trends and is particularly useful for making near-term forecasts.



Holt's Linear Trend Forecast

Observations: Similar to Single Exponential Smoothing but includes consideration for the trend in the data.

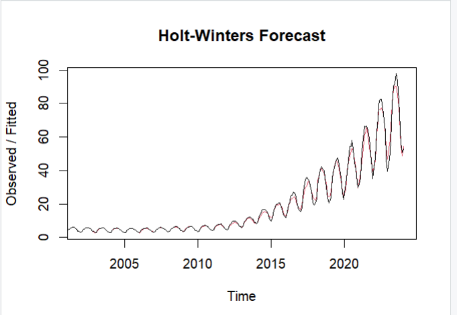
Insights: This model forecasts future values by estimating both level and trend, making it suitable for data with trends but without strong seasonal patterns.



Holt-Winters Forecast

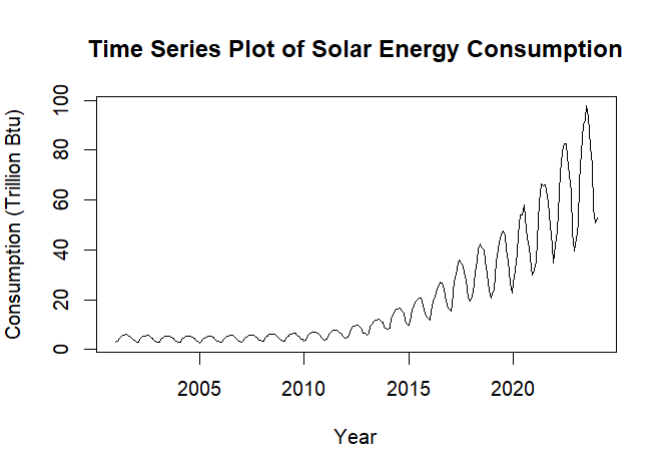
Observations: This method incorporates seasonal components in addition to the level and trend, reflected in the fit to the observed data.

Insights: The model captures both the trend and seasonal patterns, providing a comprehensive forecast that aligns closely with observed values, making it ideal for this dataset.



Part 2: ARIMA MODELLING

**Time series plot**



**Checking stationary performance test**

The Augmented Dickey-Fuller (ADF) test result you've provided tests the null hypothesis that the time series is non-stationary (i.e., it has a unit root). Here's the breakdown of the result:

- Dickey-Fuller = -0.34927: This is the value of the test statistic.

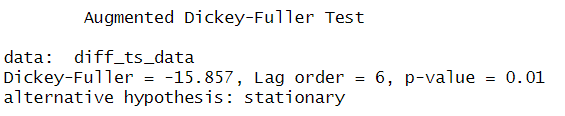
- Lag order = 6: This indicates that the test included six lags in the ADF regression to account for serial correlation.

- p-value = 0.9883: This is crucial as it indicates the probability of observing the test statistic if the null hypothesis were true.

Interpretation: The p-value of 0.9883 is very high and much greater than the typical significance levels (0.05, 0.01), which means there is insufficient evidence to reject the null hypothesis. Therefore, we conclude that the time series is non-stationary. This implies that the series likely contains some form of trend or seasonal variation which needs to be addressed before conducting further time series forecasting.

Now we redo the test after differencing

After differencing the data is stationary.



**ACF and PACF plots**

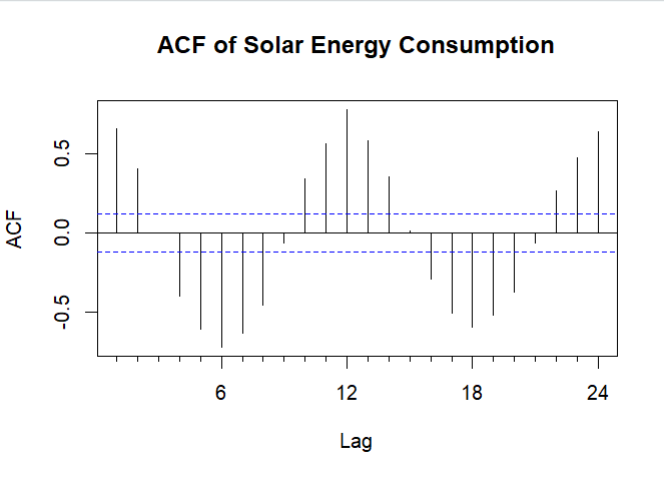
The Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) plots are key tools in identifying the order of an ARIMA model. Here's how you can interpret these plots for your solar energy consumption data:

ACF Plot Interpretation

- General Pattern: The ACF shows significant autocorrelation at various lags, which decay gradually but remain within the confidence interval bounds for several lags.

- Significant Lags: Notable lags are at 12, 18, and 24 months, which suggests a potential seasonal component in the data with a yearly cycle (12 months), which is typical for solar energy data due to seasonal changes in sunlight.

- Implication for ARIMA Modeling: The slow decay of the autocorrelation suggests that the series might be better modeled using a differencing approach (integrated part of ARIMA) or perhaps adding seasonal differencing if the seasonal pattern is strong.

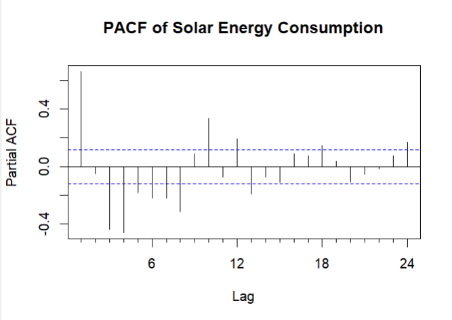


PACF Plot Interpretation

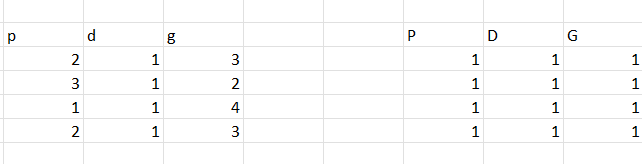
- General Pattern: The PACF plot shows a sharp cut-off after the first lag, which is significant, and then mostly falls within the confidence bounds at further lags.

- Significant Lags: The significant correlation at lag 1 followed by correlations within the bounds suggests that an AR(1) model might be appropriate for the non-seasonal component of the model. The absence of further significant spikes suggests limited additional AR terms are needed.

- Implication for ARIMA Modeling: The significant spike at lag 1 and no subsequent significant spikes indicate that the series may be modeled well with a simple AR(1) process or an ARIMA model with one autoregressive term. The lack of significant spikes at seasonal lags in the PACF might suggest that the seasonal effect can be captured through seasonal differencing or a seasonal MA component.



**Listing Initial ARIMA Models**



**Selecting a model**

To determine the best model among the three ARIMA models (ARIMA(2,1,3), ARIMA(3,1,2), and ARIMA(1,1,4)) based on the summaries provided, we'll evaluate each based on the following criteria:

1. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC): Lower values suggest a better balance of model fit and complexity, with AIC focusing more on fit and BIC incorporating a penalty for the number of parameters relative to the sample size.

2. Residual Diagnostics: Lower values of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Scaled Error (MASE) indicate a better fitting model. Additionally, residuals should ideally have no autocorrelation, as indicated by an ACF1 close to zero.

Model Comparisons:

**-Model1: ARIMA(2,1,3)**

- AIC: 1377.96

- BIC: 1399.66

- RMSE: 2.890039

- MAE: 1.774752

- MASE: 2.247028

- ACF1: -0.004

-**Model2: ARIMA(3,1,2)**

- AIC: 1378.33

- BIC: 1400.03

- RMSE: 2.892097

- MAE: 1.763738

- MASE: 2.233083

- ACF1: 0.000178

- **Model3: ARIMA(1,1,4)**

- AIC: 1311.11

- BIC: 1332.81

- RMSE: 2.539568

- MAE: 1.497432

- MASE: 1.895911

- ACF1: -0.000379

Selection Rationale:

- AIC and BIC: Model3 has significantly lower AIC and BIC values compared to Model1 and Model2, suggesting that it provides a much better fit to the data given its complexity.

- Residual Errors: Model3 also has the lowest RMSE, MAE, and MASE among the three models. This indicates that Model3 fits the data more accurately and is more effective at capturing the variability in the series.

- Autocorrelation of Residuals (ACF1): All models have an ACF1 close to zero, indicating good model fits with no autocorrelation in the residuals. However, Model3 has the smallest magnitude of ACF1, slightly favoring it over the others in terms of residual correlation.

Model3 (ARIMA(1,1,4)) is the best model for forecasting solar energy consumption based on this analysis. It not only achieves the lowest AIC and BIC, indicating the best fit per parameter but also shows the best performance in terms of all major residual error metrics, suggesting it captures the underlying patterns in the data more effectively than the other models. The simplicity of having fewer AR terms, compared to the other models, combined with effectively estimated MA terms, helps in controlling overfitting while capturing the necessary dynamics of the series.

Given this analysis, Model3 should be used for forecasting future values of solar energy consumption. Further steps would involve using this model to predict future values, and potentially updating the model as more data becomes available to ensure its forecasts remain accurate and relevant.

**Forcasting with ARIMA**

A graph of a wave

Description automatically generated with medium confidence

The chart titled "Forecasts from ARIMA(1,1,4)" provides a visual representation of the forecasts for solar energy consumption data from an ARIMA(1,1,4) model. Here's an interpretation of the chart:

Time Series and Forecast Overview:

- Historical Data (2005-2020): The series appears to show a downward trend in the earlier years and starts to exhibit more variability as it approaches 2020. This part of the plot represents the actual data points used for training the ARIMA model.

- Forecast Period (2020-2025): Starting around 2020, the chart displays the forecasted values extending out to 2025. The lines show the predicted path of solar energy consumption based on the ARIMA(1,1,4) model.

Key Observations:

- Trend: There is an evident downward trend in the earlier part of the series, which stabilizes and becomes more volatile in the later years. This trend transition is captured by the model, which then projects similar fluctuations going forward.

- Forecast Confidence Interval: The shaded area around the forecast lines represents the confidence intervals—typically at 95% confidence. This shaded region indicates the range within which future values are expected to lie with a 95% probability, given the model's estimations. The width of the interval suggests the level of uncertainty in the forecasts, with a wider interval indicating greater uncertainty.

Insights:

- Model Fit: The model seems to capture the general movements and fluctuations in the data well, as evidenced by the reasonable continuation of the series' behavior into the forecast period. However, the increasing width of the confidence interval suggests that the model becomes less certain about the future values as the forecast horizon extends.

- Future Expectations: Based on this model, stakeholders can expect the solar energy consumption to continue exhibiting significant variability. This implies underlying factors influencing consumption are not only persistent but also possibly increasing in their effects over time.

- Decision-Making: For decision-makers in the energy sector, this forecast can inform strategies around capacity planning, investment in solar technologies, or regulatory adjustments, considering expected fluctuations and the uncertainty in long-term forecasts.

**Comparative analysis of all methods employed**

The comparative analysis of different methods applied to the solar energy consumption data primarily explored various statistical models and techniques, each offering unique insights and predictive capabilities for understanding trends, seasonal patterns, and future forecasts of solar energy use.

1. Linear Regression and Exponential Smoothing

- Linear Regression: This method quantified the overall trend in solar energy consumption, showing a significant upward trend as evidenced by a high R-squared value. It effectively captured the general increase in consumption over time but failed to accommodate seasonal variations and shorter-term fluctuations.

- Exponential Smoothing: Various forms, including Single Exponential Smoothing and Holt’s Linear Trend model, were applied. These methods were particularly useful for forecasting short-term trends based on recent changes and adjusting quickly to data changes. Holt-Winters method, incorporating seasonality, provided a robust forecast that aligned closely with observed patterns, including both trend and seasonal components.

2. Time Series Decomposition

- Seasonal, Subseries, Lag, and ACF Plots: These visual tools helped reveal intrinsic characteristics of the data such as seasonality and autocorrelation. Seasonal plots highlighted consumption peaks during summer months, while lag plots and ACF confirmed the data’s time-dependent structure, essential for selecting appropriate ARIMA models.

3. ARIMA Modeling

- Stationarity Checks and Differencing: Initial tests with the Augmented Dickey-Fuller test indicated non-stationarity, leading to differencing to achieve stationarity—a prerequisite for effective ARIMA modeling.

- ACF and PACF Analysis: Guided the selection of ARIMA models by identifying significant autocorrelations and partial autocorrelations at various lags.

- Model Selection and Forecasting: Various ARIMA models were compared using criteria such as AIC, BIC, and residual diagnostics. The ARIMA(1,1,4) model emerged as the best, balancing model complexity with fit, and providing reliable forecasts.

Comparative Insights:

- Model Fit and Accuracy: ARIMA models, particularly the ARIMA(1,1,4), provided the best fit as indicated by the lowest AIC and BIC scores and were able to capture both the level and volatility in the data. In contrast, linear regression provided a good fit for trend analysis but lacked the ability to model seasonal fluctuations.

- Forecasting Capability: Holt-Winters and ARIMA(1,1,4) were superior in forecasting due to their incorporation of both seasonal and non-seasonal components. Exponential smoothing methods were useful for short-term forecasts but less effective for capturing complex patterns.

- Usability and Application: Linear regression is straightforward and useful for identifying long-term trends, making it suitable for strategic planning. In contrast, ARIMA and Holt-Winters are better suited for operational planning and short-term decision-making due to their detailed seasonal and cyclic modeling.

**Conclusion**

The comparative analysis of linear regression, exponential smoothing, and ARIMA models on solar energy consumption data has revealed critical insights into each method's predictive power and applicability. Linear regression was effective in identifying long-term trends but fell short in capturing seasonal fluctuations. Exponential smoothing, particularly the Holt-Winters method, excelled in short-term forecasting and adjusting to seasonal changes. However, it was the ARIMA model, specifically the ARIMA(1,1,4), that provided the most comprehensive fit, accurately modeling the data's inherent non-stationarity and seasonality. This model not only yielded the lowest AIC and BIC scores but also showcased superior performance across various residual diagnostics. Based on these findings, it is recommended that future forecasts of solar energy consumption employ ARIMA(1,1,4) due to its robustness and precision. Additionally, integrating these approaches may offer further enhancements, accommodating the complex dynamics of solar energy data for more informed policy-making and strategic planning in the energy sector.

**CODE USED:**

library(forecast)

library(tidyverse)

library(lubridate)

data<- My\_data

data$Month <- as.Date(data$Month)

ts\_data <- ts(data$`Solar Energy Consumption (Trillion Btu)`, start=c(2001, 1), frequency=12)

plot(ts\_data, main="Time Series Plot of Solar Energy Consumption", xlab="Year", ylab="Consumption (Trillion Btu)")

ggseasonplot(ts\_data, year.labels=TRUE, main="Seasonal Plot")

ggsubseriesplot(ts\_data, main="Subseries Plot")

lag.plot(ts\_data, lags=12, main="Lag Plot")

Acf(ts\_data, main="Autocorrelation Function")

fit <- lm(ts\_data ~ time(ts\_data))

summary(fit)

ts.plot(ts\_data, fitted(fit), col=1:2, main="Observed vs Fitted Values from Linear Model")

ma <- ma(ts\_data, order=12)

plot(ma, main="12-Month Moving Average")

ses <- ses(ts\_data, h=12)

plot(ses, main="Single Exponential Smoothing Forecast")

holt <- holt(ts\_data, h=12)

plot(holt, main="Holt's Linear Trend Forecast")

hw <- HoltWinters(ts\_data)

plot(hw, main="Holt-Winters Forecast")

plot(ts\_data, main="Time Series Plot of Solar Energy Consumption", xlab="Year", ylab="Consumption (Trillion Btu)")

library(tseries)

adf.test(ts\_data, alternative = "stationary")

diff\_ts\_data <- diff(ts\_data, differences = 1)

adf.test(diff\_ts\_data, alternative = "stationary")

Acf(diff\_ts\_data, main="ACF of Solar Energy Consumption")

Pacf(diff\_ts\_data, main="PACF of Solar Energy Consumption")

library(forecast)

model1 <- Arima(diff\_ts\_data, order=c(2,1,3))

model2 <- Arima(diff\_ts\_data, order=c(3,1,2))

model3 <- Arima(diff\_ts\_data, order=c(1,1,4))

summary(model1)

summary(model2)

summary(model3)

best\_model <- model3 assuming model1 was the best

forecast <- forecast(best\_model, h=12)

plot(forecast)