**MUMBAI RAINFALL FORECASTING USING TIMESERIES**

**Data Source**

Prediction of Worldwide Energy Resources (POWER) project provides meteorological data from NASA research for support of renewable energy, building energy efficiency and agricultural support. NASA has an Earth Science research program with satellite systems providing research data that are important to study of climate and climate process. These data contain long-term climate averaged estimates and surface solar energy fluxes.

**Data Collection**

Data is collected from [Power Data Access Viewer](https://power.larc.nasa.gov/data-access-viewer/)  
The POWER Meteorological data is prediction or observation given by NASA's GMAO MERRA-2 assimilation model.  
The data collected is monthly frequency data for a particular latitude and longitude in Mumbai for the period 2000 – 2020. The data consists of the following variables:  
• Specific Humidity  
• Relative Humidity:  
• Temperature  
• Precipitation (The data consist of precipitation as monthly sum of rainfall)

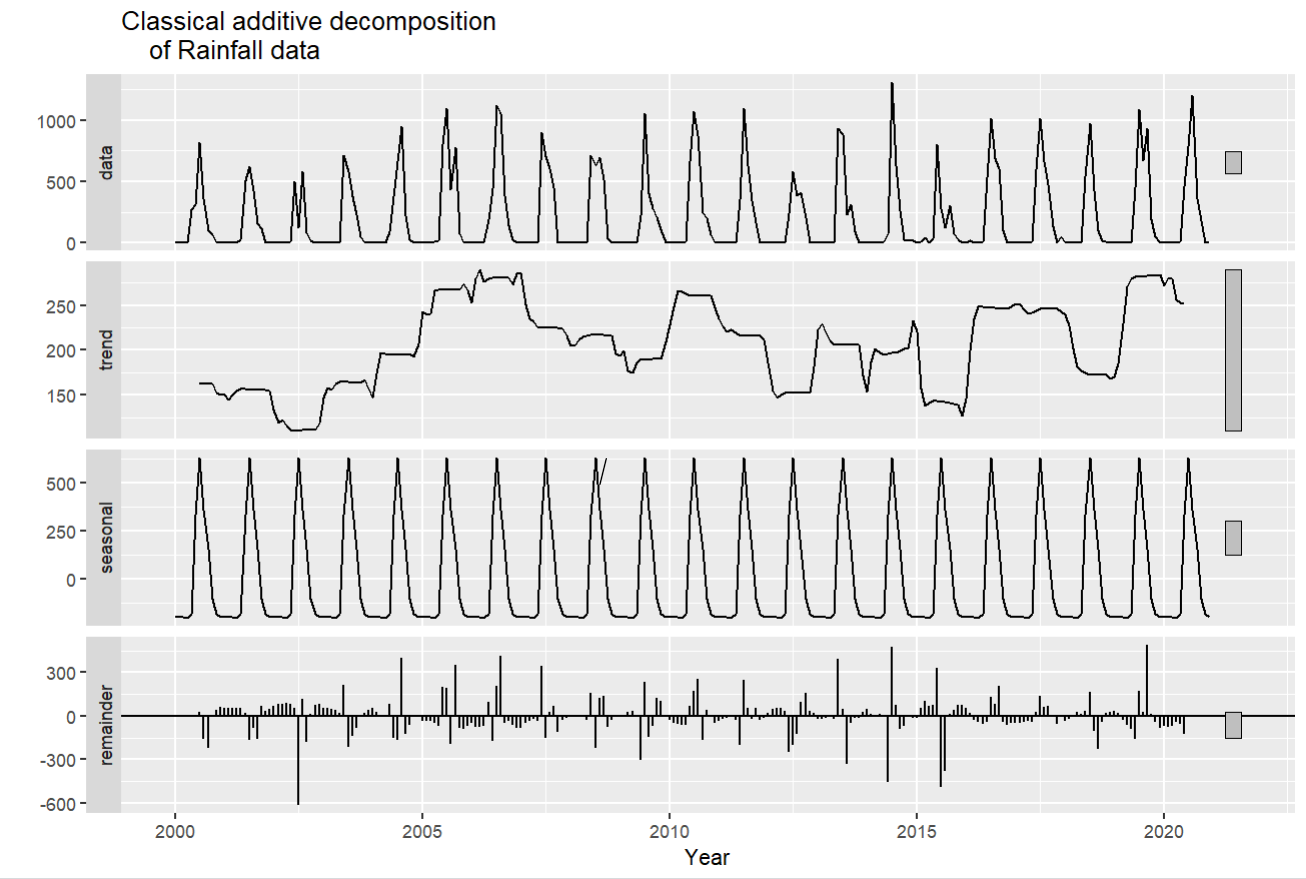
* Plot the time series data from the year 2000 to 2020

A picture containing bar chart

Description automatically generated

* **Classical Decomposition of data**

A time series data is composed of

* + Trend: Trend is the overall direction of the data.
  + Seasonlity: Seasonality is a periodic component which repeats itself within a particulat time period.
  + Residuals: the residual is what’s left over when the trend and seasonality have been removed. Residuals are random fluctuation
* 

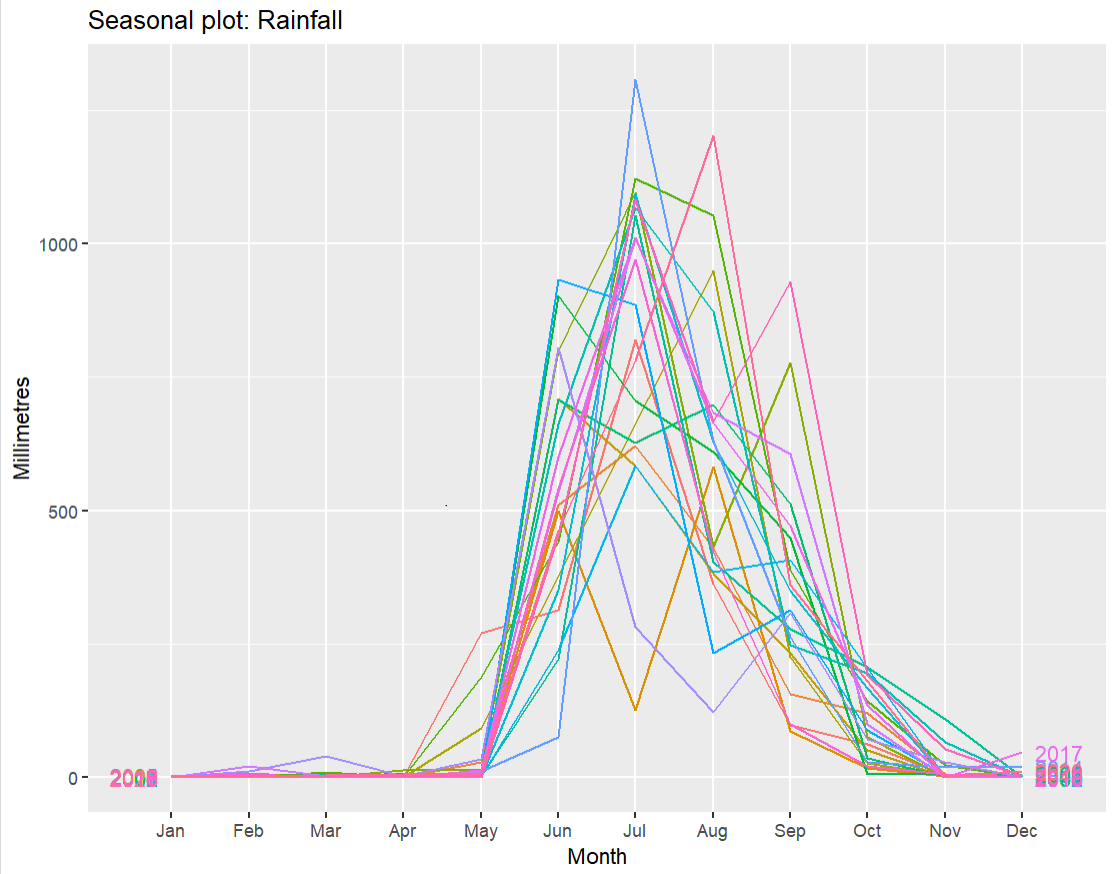
Null Hypothesis: Data is Stationary

Alternative Hypothesis: Data is not stationary

From KPSS test we got p-value = 0.1 which means:

We fail to reject null hypothesis.

* To check the seasonality we plot:



* Now to see the lag: we can see there is good correlation at lag 12 which is expected also.

Calendar

Description automatically generated

* ACF and PACF plot:

Histogram

Description automatically generated

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are important tools for analyzing time series data. These plots help to identify the presence of autocorrelation and provide insights into the order of the Autoregressive (AR) and Moving Average (MA) components of a time series.

Exponential Smoothing

* Exponential smoothing is generally used to make short-term forecasts
* More recent observations are given larger weights by exponential smoothing
* Exponential Smoothing Methods
  + Simple or Single Exponential Smoothing
  + Double Exponential Smoothing
  + Triple Exponential Smoothing

**Single Exponential Smoothing**

* If the data has no trend and no seasonal pattern, then this method of forecasting the time series is essentially used.
* This method uses weighted moving averages with exponentially decreasing weights.

ME RMSE MAE

0.05 290.49 168.45

A picture containing text, stationary, writing implement, pencil

Description automatically generated

**Double Exponential Smoothing**

* This method is also called as Holt’s trend corrected or second-order exponential smoothing.
* This method is used for forecasting the time series when the data has a linear trend and no seasonal pattern.
* It introduces a new smoothing factor beta that addresses trend in data

Chart, bar chart

Description automatically generated

ME RMSE MAE

-1.246047 290.4758 169.0363

**Triple Exponential Smoothing**

* In this method, exponential smoothing applied three times.
* This method is used for forecasting the time series when the data has both linear trend and seasonal pattern.
* This method is also called **Holt-Winters exponential smoothing.**
* Chart, bar chart

  Description automatically generatedIt introduces a new smoothing factor γ that addresses seasonality in data

ME RMS MAE

9.075332 147.3524 88.9097

Residual Plot for HW method

Graphical user interface, chart

Description automatically generated There are only two spikes outside the upper and lower bound range in ACF plot, that too after a significant lag.

Which mean there is no residual error.

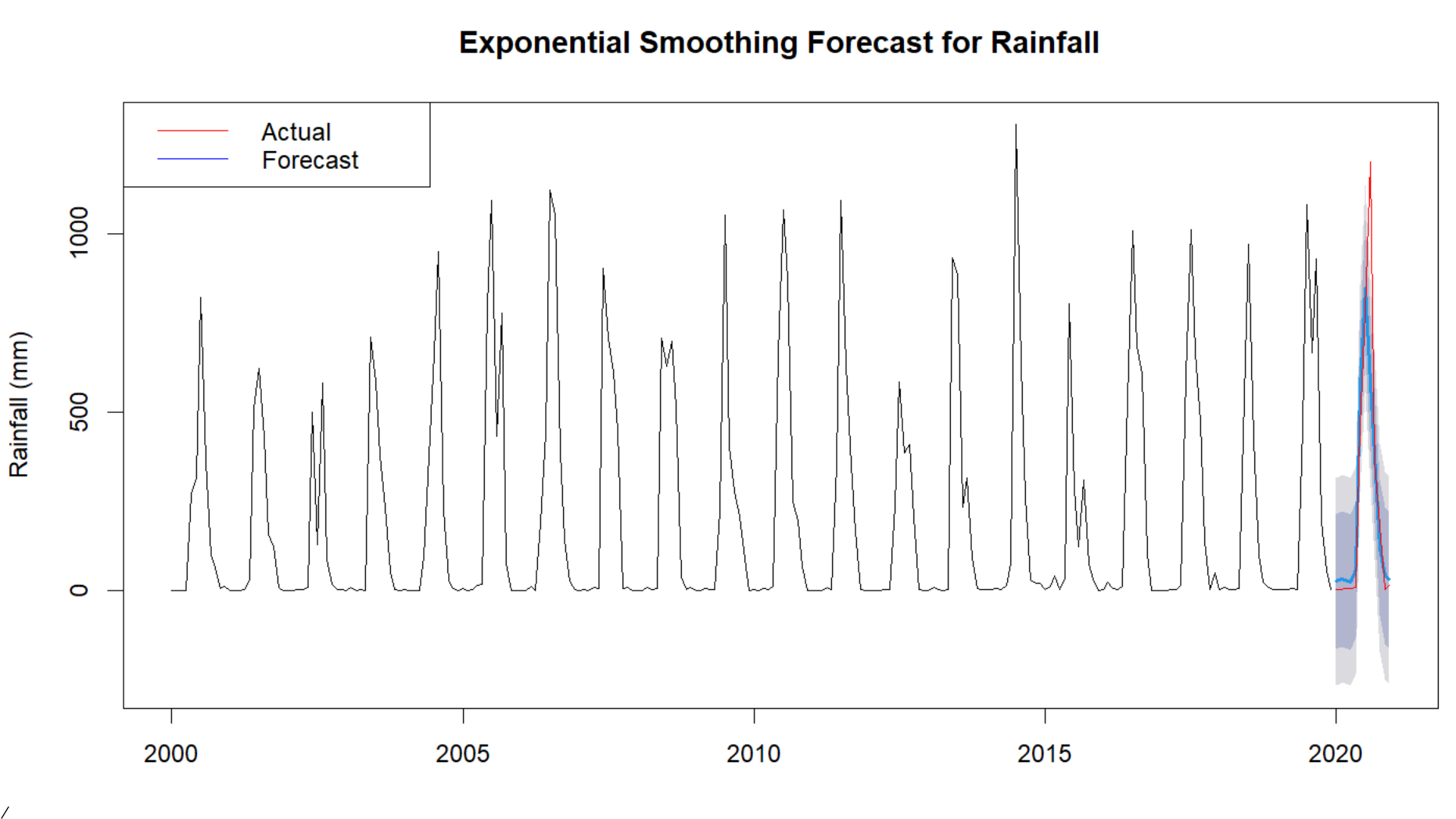
**ETS Model:**

ETS models, or Error-Trend-Seasonality models, are a class of time series models used to forecast future values of a time series based on its historical values. The ETS framework was introduced by Rob Hyndman and Anne Koehler in 2006 as an extension of the popular ARIMA models.

ETS models decompose a time series into three components: error (E), trend (T), and seasonality (S). The error component represents the random fluctuations or noise in the data that cannot be explained by the trend or seasonality. The trend component represents the long-term direction of the data, while the seasonality component captures the periodic fluctuations that occur within a given time frame

ME RMSE MAE

6.572712 144.0193 83.08158



**SARIMA (Seasonal Autoregressive Integrated Moving Average)** :

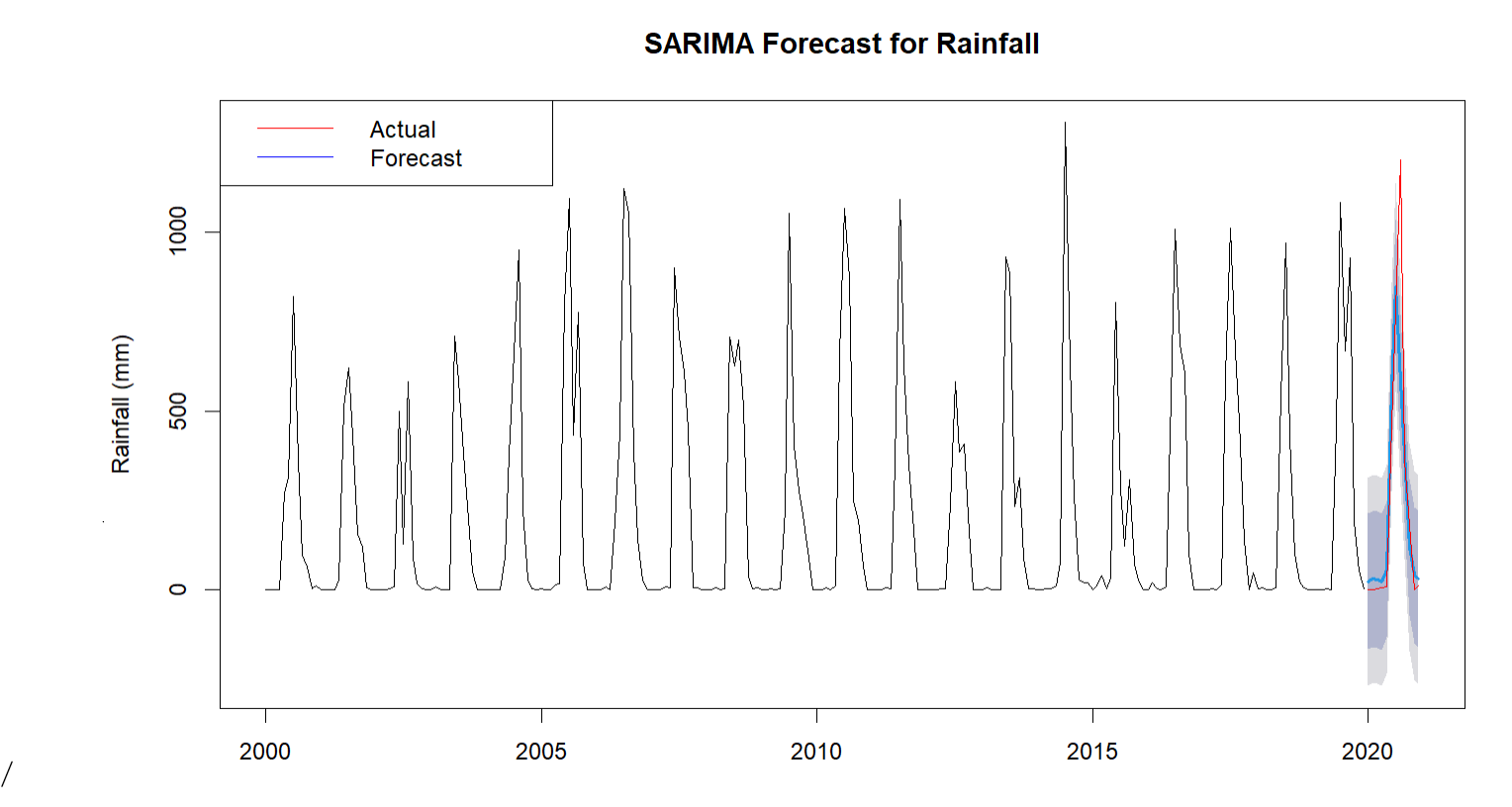
It is a time series forecasting model that extends the ARIMA (Autoregressive Integrated Moving Average) model to incorporate seasonality. SARIMA models are used to analyze and forecast data that exhibits both trends and seasonal patterns.

The SARIMA model is specified by three components:

**The seasonal part** (S): this describes the repeating pattern over the course of a year (or other period of time) and is denoted by the notation (p,d,q)x(P, D, Q)s, where P, D, and Q are the orders of the autoregressive, integrated, and moving average components, respectively, and s is the length of the seasonal cycle (the number of periods in a season).

**The non-seasonal part (ARIMA):** This describes the non-repeating patterns in the data and is denoted by the notation (p, d, q), where p, d, and q are the orders of the autoregressive, integrated, and moving average components, respectively.

**The trend component**: This describes the overall trend in the data and can be incorporated into either the seasonal or non-seasonal part of the model

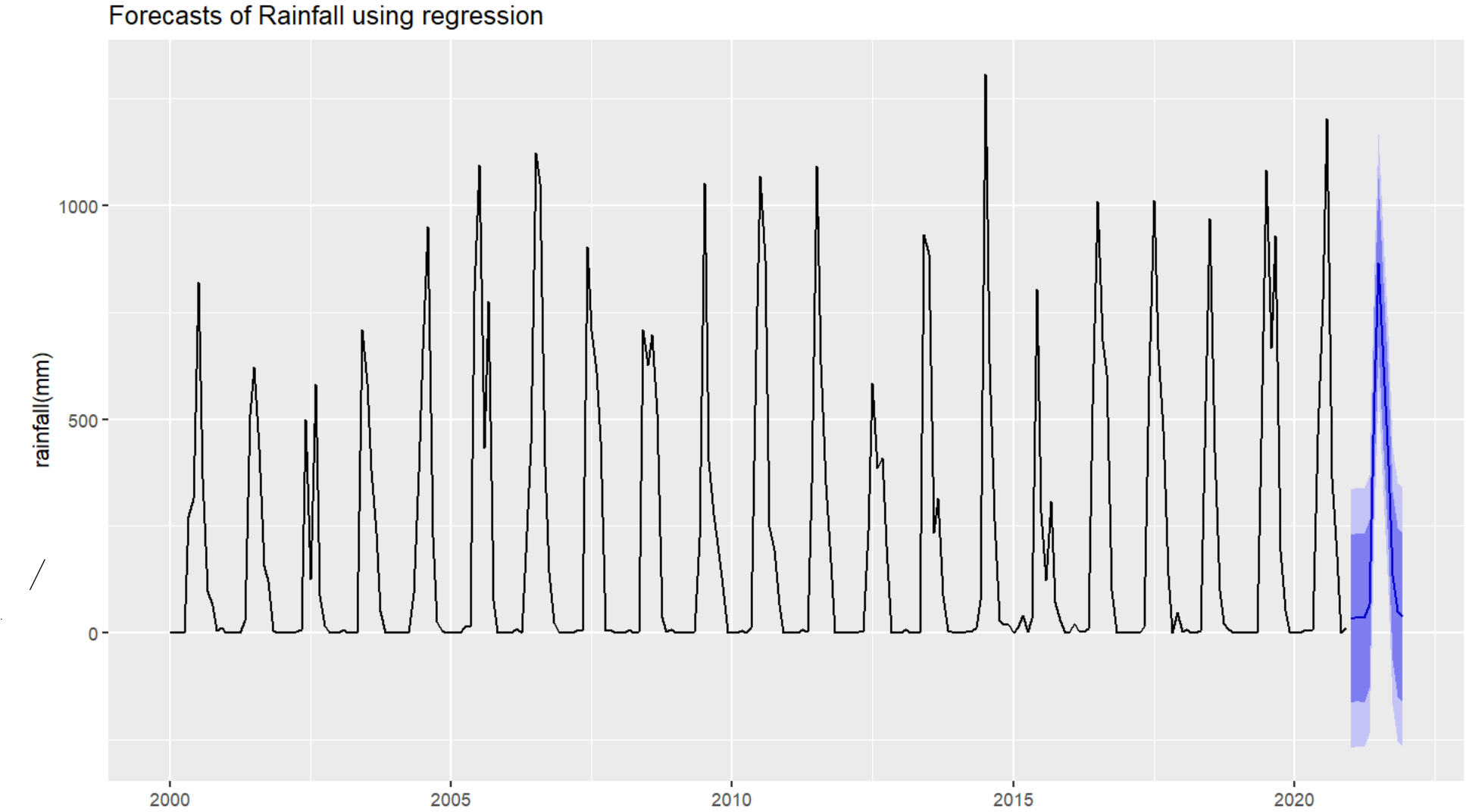


|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Non-seasonality order (p,d,q)** | **Seasonal order (P,D,Q)** | **RMSE** | **AIC** |
| sarima1 | (1,0,1) | (1,1,1) | 149 | 3122.98 |
| sarima2 | (1,0,0) | (1,1,2) | 149.732 | 3125.84 |
| sarima3 | (1,0,1) | (1,1,2) | 148.83 | 3124.83 |
| sarima4 | (2, 0, 0) | (1, 1, 1) | 149.474 | 3124.79 |
| sarima5 | (2, 0, 0) | (1, 1, 2) | 149.194 | 3126.5 |
| sarima6 | (1, 0, 0) | (0, 1, 1) | 149.911 | 3126.51 |
| auto.arima | (1,0,0) | (1,1,1) | 147.05 | 2959.32 |

**Linear regression time series forecasting**:

It is a method of using historical data to make predictions about future values of a time series. It involves fitting a linear regression model to the historical data and then using that model to make predictions about future values.

The basic idea behind linear regression time series forecasting is to use past observations of a variable to predict future values of that variable. This is done by fitting a linear regression model to the historical data, where the independent variable is time and the dependent variable is the variable being forecasted. The resulting model can then be used to predict future values of the dependent variable.



ME RMSE MAE

-1.244816e-15 144.8484 81.35416

|  |  |
| --- | --- |
| Model | RMSE |
| linear regression | 144.8484 |
| Holt | 290.4758 |
| Holt-winter's | 147.352 |
| simple exp smoothing | 290.49 |
| SARIMA | 147.05 |
| ETS | 144.0193 |

On the basis of best RMSE score the fit model for our forecasting is the ETS model.

* Checking the residual plot for our best fitted model
* Graphical user interface, chart

  Description automatically generated
* The plot shows almost no error as the spike is after significant lag of 12
* Forecast for next 12 periods.

Text

Description automatically generated