STA457 Group Project

Meer Sisodia, Natalia Lisa George, Helen Zhao, Vraj Shah

Introduction

Cocoa is a key ingredient for the global food industry and plays a critical role for many companies and their food production. Cocoa prices are influenced by a variety of factors such as climatic conditions, market demand and exchange rates. This report aims to explore the prediction of cocoa prices through statistical models by identifying key predicters of price. The motivation for this study stems from the increasing volatility of cocoa prices observed in the last year which are driven by climate change, unpredictable weather patterns, diseases that affect the crop and shifts in global demand. Cocoa prices have skyrocketed to nearly $10,000 per metric ton in 2024. Around 80% of the world's supply comes from Western Africa, particularly Ghana. Hence, understanding the factors that affect the price of cocoa is essential for various stakeholders such as farmers, manufacturers, consumers and governments. The objective of this analysis is to build a predictive model of cocoa prices using historical data on variables such as average temperature and rainfall. The key challenges of this analysis have been the non-stationary prices of cocoa and several economic trends that can affect the price. Nevertheless, this report aims to provide a valuable prediction for future cocoa prices considering the recent volatility of the commodity. These predictions will help stakeholders of the industry to make more informed decisions and help policy makers concerned with the economic well being of farmers in cocoa growing regions.

Literature Review

Price volatility in agricultural goods can be due to several conditions. Cocoa plants require moderate rainfall and temperatures between 65-90 F. The prediction of commodity prices has been a topic of interest in economic research. Commonly used models are Autoregressive Integrated Moving Average (ARIMA), Generalized Autoregressive Conditional Heteroskedasticity (GARCH), and Vector Autoregressive (VAR) models. As shown by Ketut Sukiyono (2018) the ARIMA method was most suitable for predicting both domestic and foreign prices of cocoa. ARIMA assumes a linear process and can be effective when the time series is stationary. However, ARIMA models are not able to account for exogenous variables such as weather conditions or demand shocks. Engle (2012) showed that GARCH models can be used to model inflation and commodity price. These models are useful to study volatility and variability in prices. Kamu et al. (2010) forecast the prices of cocoa and conclude that GARCH models are the best for predictions. There have also been studies that use VAR models to see the dynamic interactions between multiple time series variables. Kutu (2019) highlights that exchange rate volatility has an impact on the cocoa prices in Nigeria.

This study builds on existing research by combining a combination of weather factors such as rainfall, temperature and macroeconomic variables such as exchange rate to develop a comprehensive model to predict the price of cocoa. By using Impulse Response Function this study also explores how shocks to temperature,

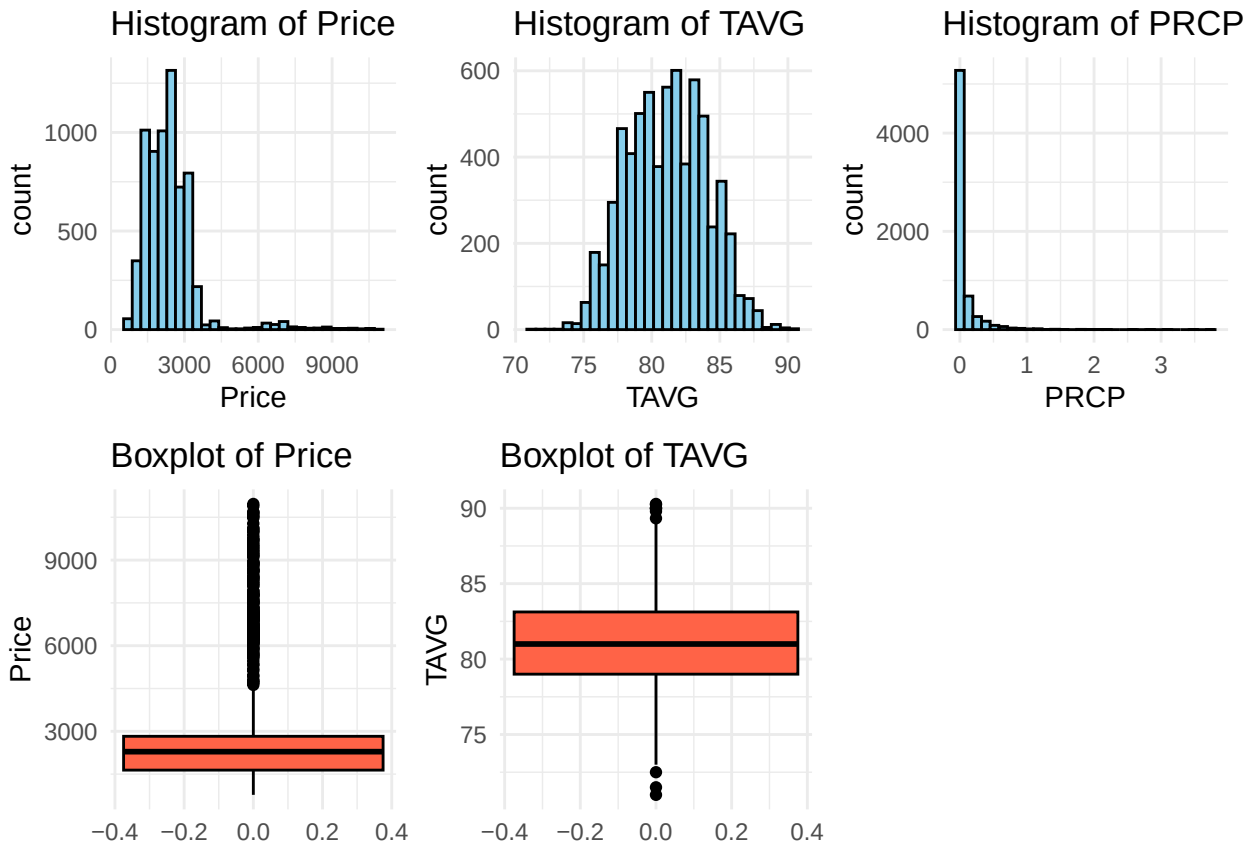
rainfall propagate through the system and affect cocoa price over time which has not been explicitly studied previously.

Data

In the following analyses, we will be analyzing two datasets, the Cocoa Future Prices data and Ghana's climate data. Sourced from the International Cocoa Organization (ICCO), the Cocoa Future Prices dataset consists of time series data regarding the daily trade day closing prices for cocoa futures contracts traded on major commodity exchanges, measured in US dollars, with the timeframe spanning from March 3rd, 1994, to February 2nd, 2025. The Ghana climate dataset contains 53,232 entries of climate data from Ghana, which is the top cocoa-producing country in the world. Sourced from the National Centers for Environmental Information (NCEI), the data includes various parameters recorded by observation stations. Each record includes information consisting of the station ID that the climate was recorded at, the name of the observation station, and the date of the observation. Further, the dataset features daily climatic measurements of each observation station including the precipitation levels, measured in millimeters, and the daily average, maximum, and minimum temperatures, measured at 2 meters above ground level in degrees Fahrenheit. The dataset offers valuable insights into Ghana's daily weather conditions and can be crucial for agricultural planning and for cocoa cultivation, which could potentially be significant for predicting and unveiling the reasons behind cocoa price fluctuations. With the data, we will focus on investigating two main variables of interest: the independent variable of TAVG and PRCP, the daily average temperatures and precipitation levels of each of Ghana's observation stations, and the main outcome variable of interest of Cocoa Future Prices.

Methodology

To obtain meaningful results, we first clean the data. We then check for any NA results in the TAVG variable, PRCP variable and the outcome variable of interest, Cocoa Future Prices, as those invalid responses may distort analyses of data findings. Missing values of PRCP variable are taken as zero as no rainfall occurred on that day. There were no other missing values. We further transform the data by averaging across all observation stations' daily TAVGs and PRCP to obtain a final average daily temperature and precipitation record for each day. Finally, we omit all temperatures and climatic measurements collected during weekends in the Ghana climate dataset; this exclusion ensures that the climate dataset aligns with the trading days when Cocoa Future Prices are recorded, avoiding any discrepancies in data timelines.



The distributions of the independent variable TAVG and PRC (see Figure 1.) and the outcome variable of interest Cocoa Future Prices are displayed above. TAVG's histogram does not display obvious skews, indicating that the dataset mostly resembles a normal distribution. The mean of the variable is 81.31071, with the median being 81, revealing that most of the sampled Ghana observation stations' temperatures tend around 81 degrees.The distribution of precipitation has a heavy right skew.

On the other hand, Cocoa Future Prices' histogram showcases an obvious positive skew, demonstrating that the majority of the Cocoa Future Prices data is centered in the range between 0 to 4000 dollars, with only 339 observations with values ranging from 4000 up to 12000. The discrepancies between the data's mean and median, with the mean being 2374.009 and the median being 2223.98 where the mean is significantly above the median, also indicate the rightward skew of the dataset.This skewness of the price can also be confirmed from the box plot as there are an outlier that can be seen from the graph.

The boxplots of the two datasets further showcase a clear comparison between the two dataset's distributions. As displayed in Figure 1, the boxplot of TAVG does not display obvious outliers on either end, matching with the previous analysis of the histogram, resembling the normal distribution. As displayed in Figure 4, the boxplot of Cocoa Future Prices displays a rather large proportion of anomality, with 288 outliers on the positive end. The positive skew of the distribution indicates that there may have been a potential factor of market shock that resulted in a sudden large increase in Cocoa Future Prices, leading to its deviation from its center.

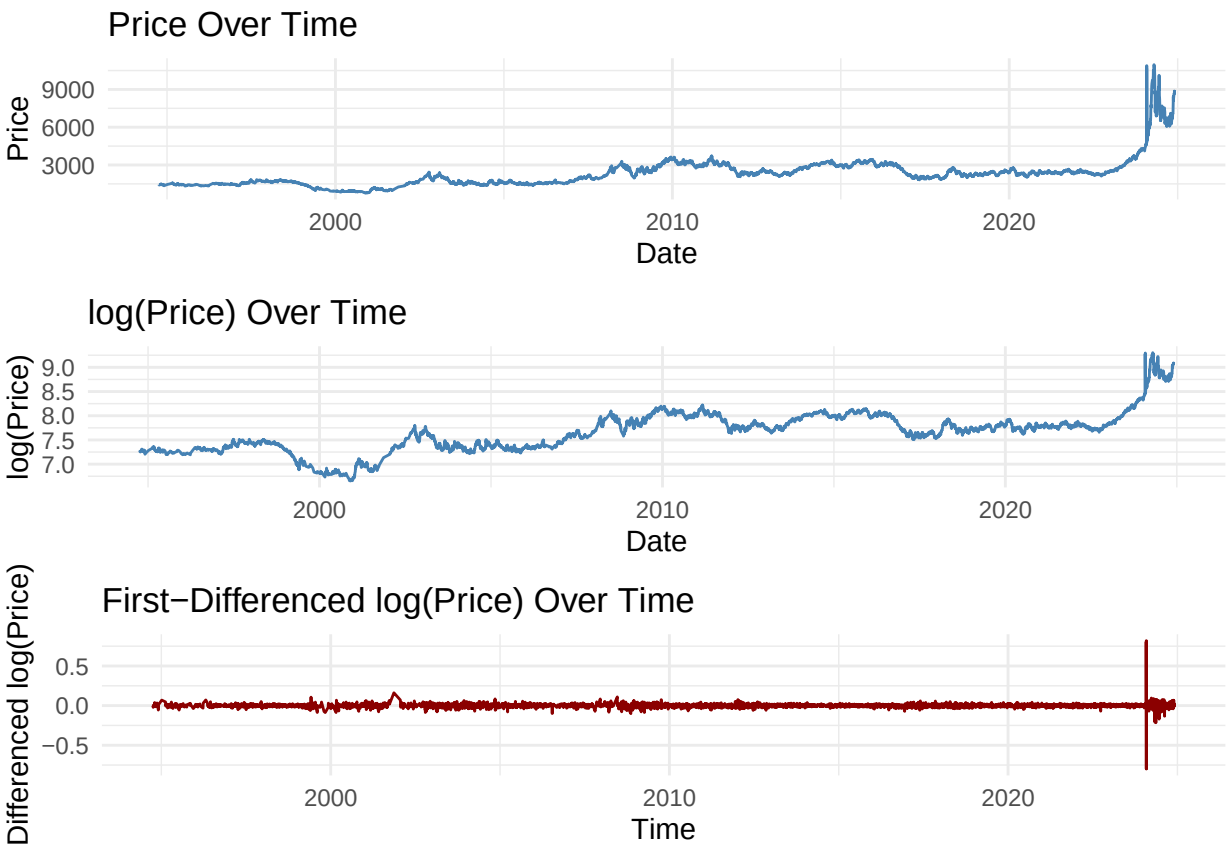
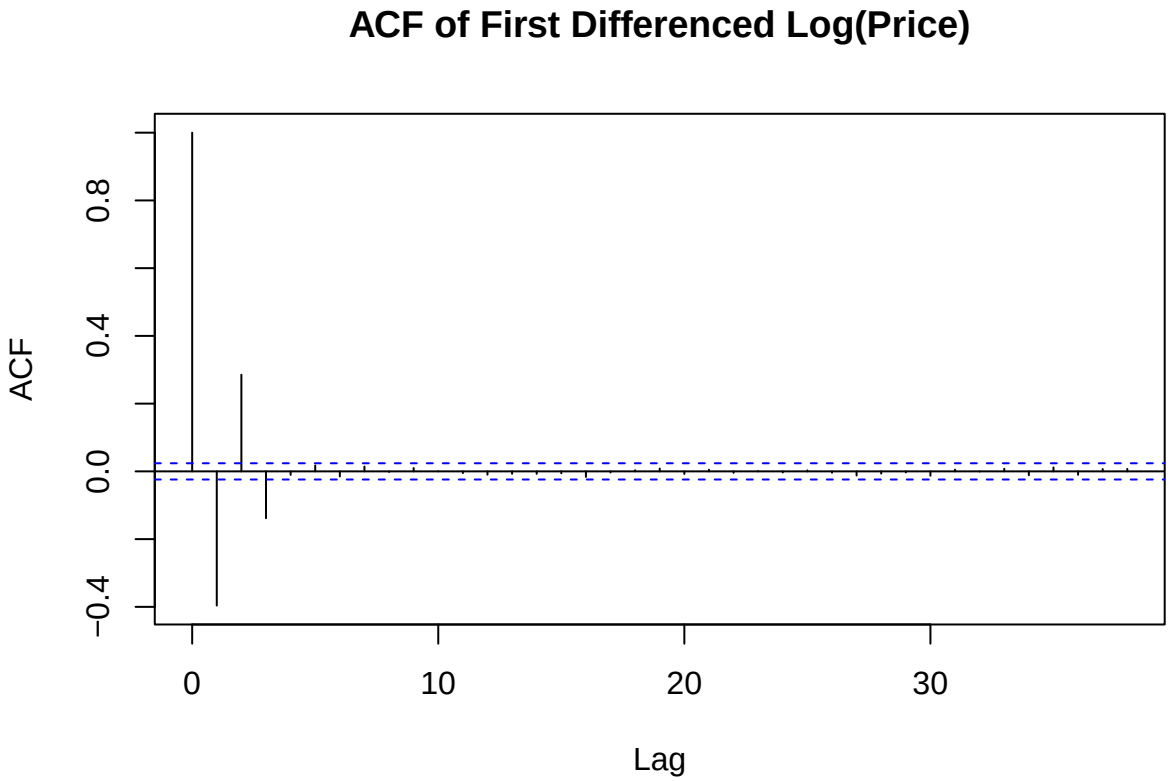
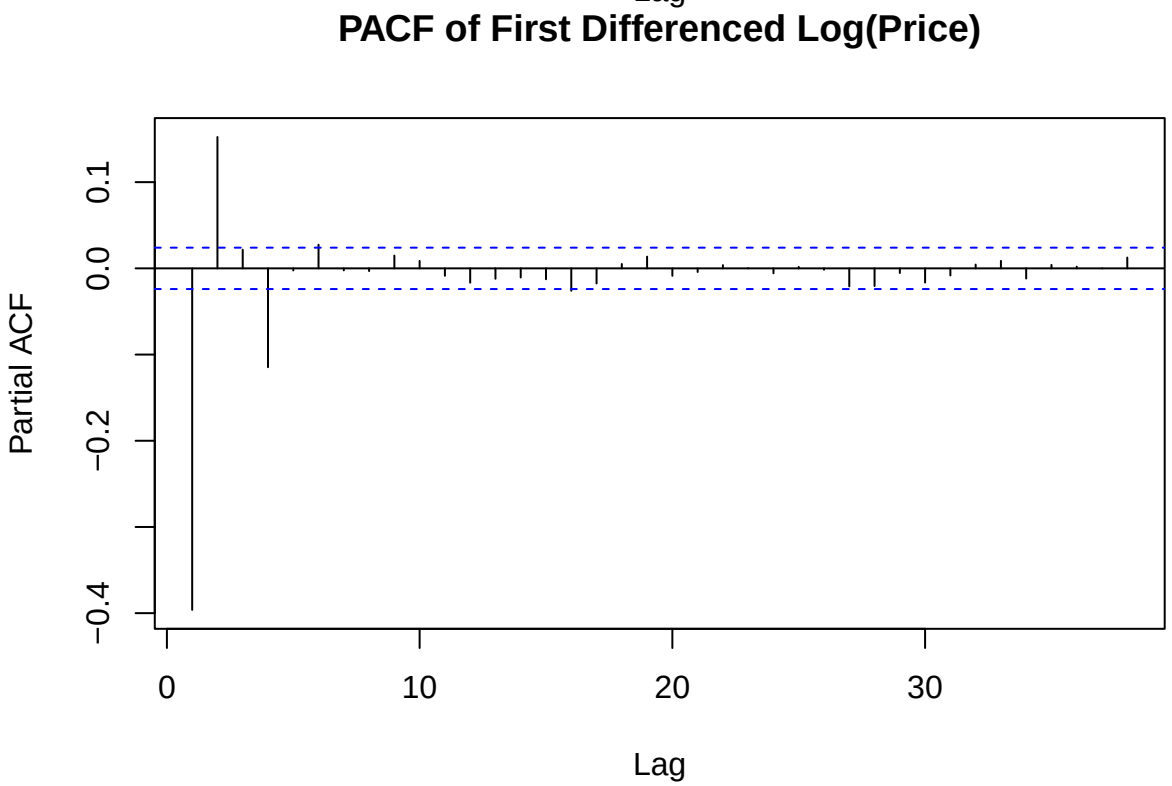


Figure 1: Figure 2





We now take a deepr look at the price variable (See Figures 2 and 3). The original Price Over Time plot (first graph) shows substantial volatility, with prices fluctuating dramatically between 2000 and 2020, ranging from around 2,000 to over 9,000. The erratic peaks and troughs suggest significant instability, making it difficult to discern a clear trend. To better analyze the data, we first applied a logarithmic transformation,

resulting in the log(Price) Over Time plot (second graph). The log-transformed version revealed a more stable pattern, smoothing extreme fluctuations. However, to fully address the remaining non-stationarity, we computed first differences of the log prices, yielding the differenced log(Price) series. This differenced plot now exhibits stationary behavior, with fluctuations centered around zero for reliable time series modeling. Further diagnostics via the ACF plot indicate the differenced series tails off gradually. Meanwhile, the PACF plot displays an erratic spike at lag 4, hinting at a possible ARIMA. Together, these findings justify the use of an ARIMA framework to model the data, so we have fitted the first model as ARIMA (4,1,0).

Next, to incorporate the effects of temperature and precipitation we employ a linear regression model to try to explain variation in prices. Recognizing the problem of correlated errors we incorporate a GARCH(1,1) model for the error term along with the basic regression model. This helps in capturing time varying volatility, allowing for more accurate price prediction. Also, accounting for the ACF and PACF of the error terms, we then employ a AR(4) + GARCH(1, 1) on the error terms in price. The allows us to account for lagged effects where past price fluctuations influence current changes. We also use a machine learning model- XGBoost, to see if other statistical models provide better predictions. Vector Autoregressive Model is used to see the dynamic relationship between the variables and to see the long-term trends. To fit all these models, the data has been split into a training and testing datasets. We use RMSE and MSE values calculated with testing data to evaluate the models. The testing dataset consists of the last 30 values of the dataset.

Results

The ARIMA(4,1,0) model fits the first-differenced log prices using the following AR(4) process:

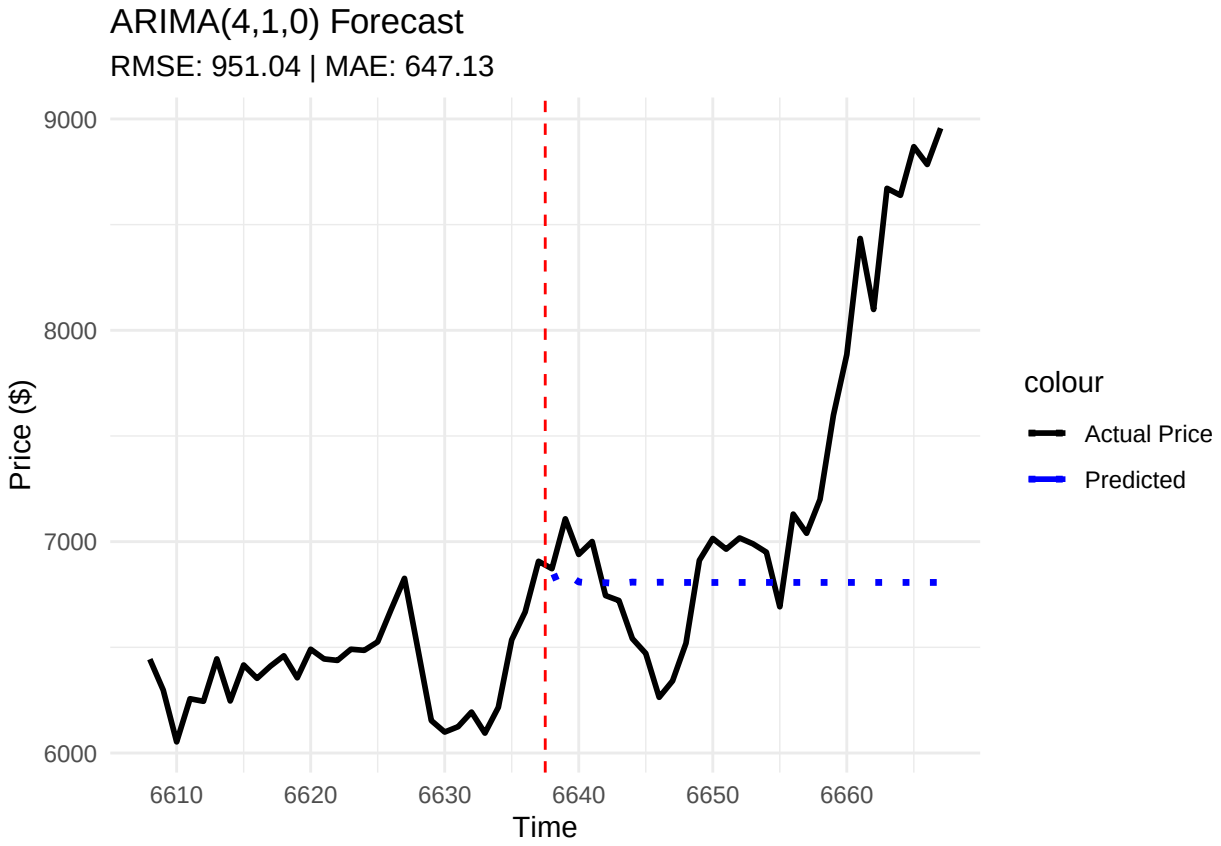
$$\Delta y_t = \phi_1 \Delta y_{t-1} + \phi_2 \Delta y_{t-2} + \phi_3 \Delta y_{t-3} + \phi_4 \Delta y_{t-4} + \varepsilon_t$$

where
- $\Delta y_t = y_t - y_{t-1}$ is the first difference of the log price,
- $\phi_1, \phi_2, \phi_3, \phi_4$ are the autoregressive coefficients,
- ε_t is white noise.

The figure below shows us the values of the RMSE and MAE for this model. We can also see the forecast that this model gives for the test data.

Table 1: ARIMA(4,1,0) Coefficients

	x
ar1	-0.3395
ar2	0.1743
ar3	-0.0193
ar4	-0.1164

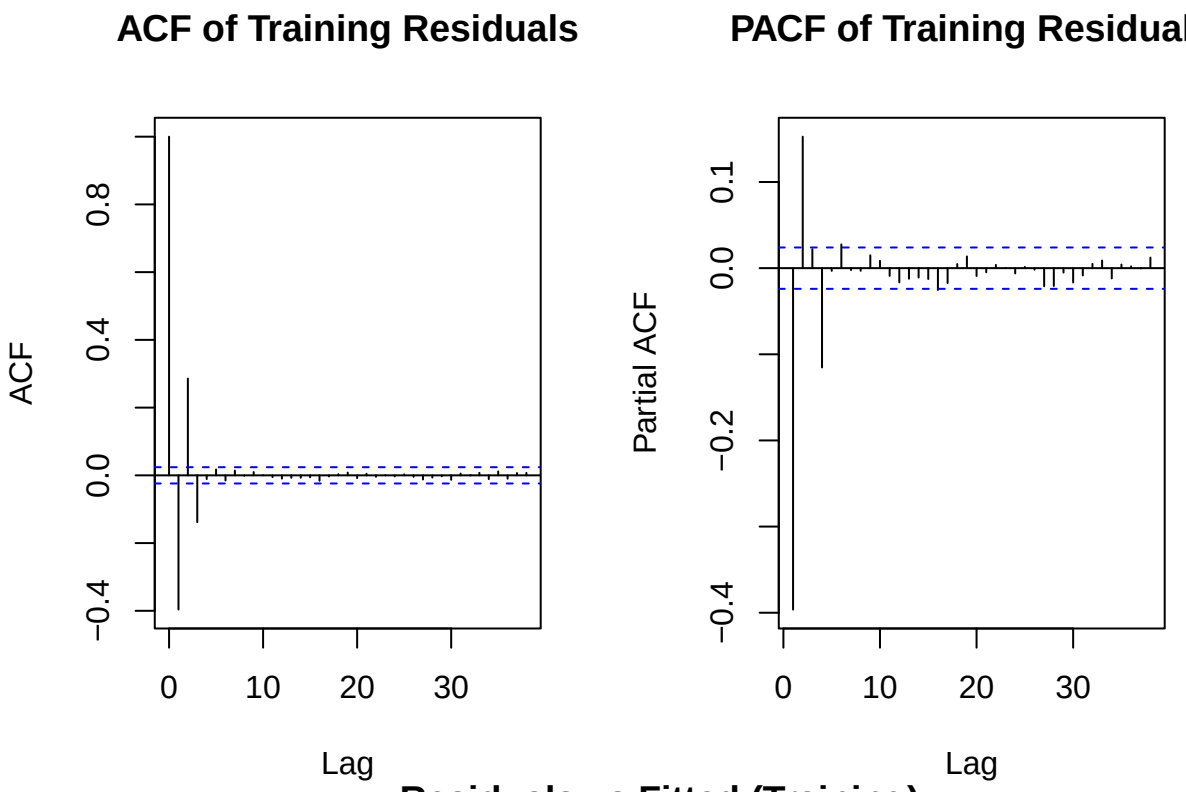


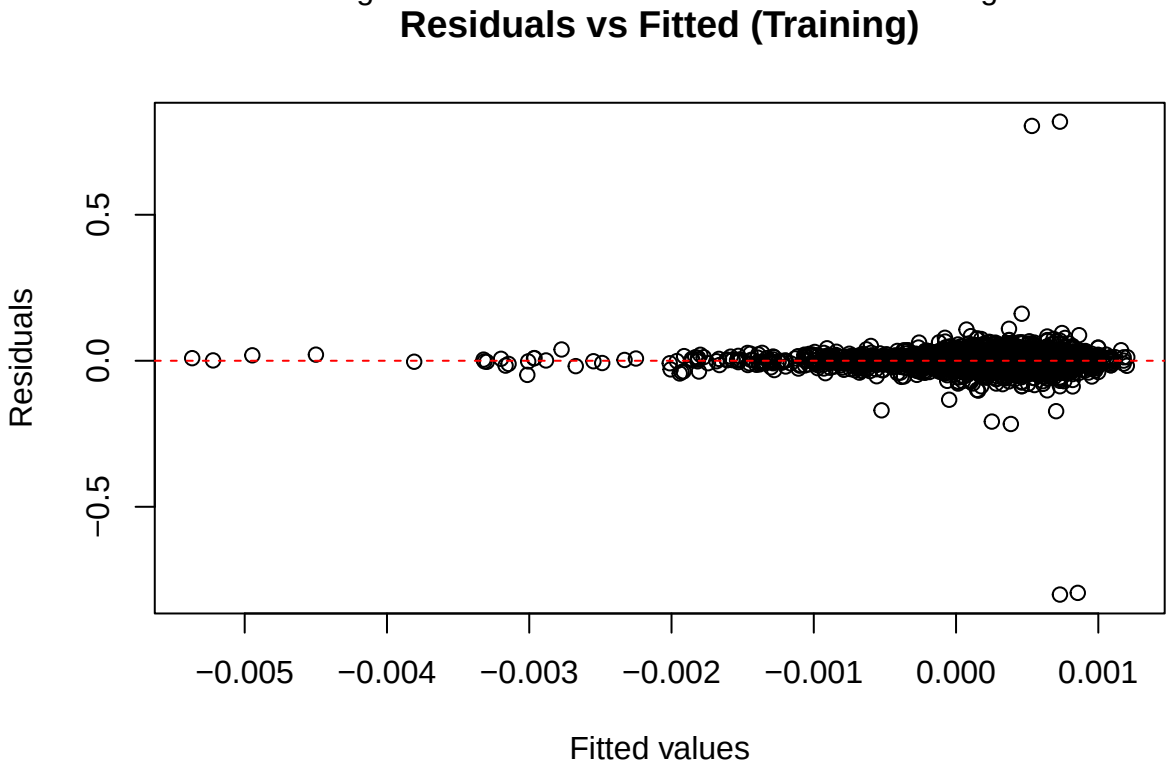
Now we move to fitting models which incorporate a regression component. We start by fitting a simple linear regression model and we analyse the residuals.

$$\Delta \log(\text{Price}_t) = \beta_0 + \beta_1 \cdot \text{PRCP}_t + \beta_2 \cdot \text{TAVG}_t + \varepsilon_t$$

where

- $\Delta \log(\text{Price}_t)$ is the change in log price (i.e., log return),
- PRCP_t is the precipitation at time t,
- TAVG_t is the average temperature at time t,
- $\beta_0, \beta_1, \beta_2$ are regression coefficients,
- ε_t is the error term.





Looking at the ACF and PACF of the residuals, it is quite evident that they are correlated. The residual vs fitted graph shows funneling suggesting heteroskedasticity or that the variances of the errors are not the same. A possible solution for this is to incorporate a GARCH(1, 1) model on the errors. Furthermore, it seems that the ACF is tailing off, however the PACF of the errors seems to abruptly cut off at lag 4. Thus, we may also fit an AR(4) + GARCH(1, 1) model on the errors.

The GARCH(1,1) errors model is given by the equations:

$$\Delta \log(\text{Price}_t) = \beta_0 + \beta_1 \cdot \text{PRCP}_t + \beta_2 \cdot \text{TAVG}_t + y_t$$

and

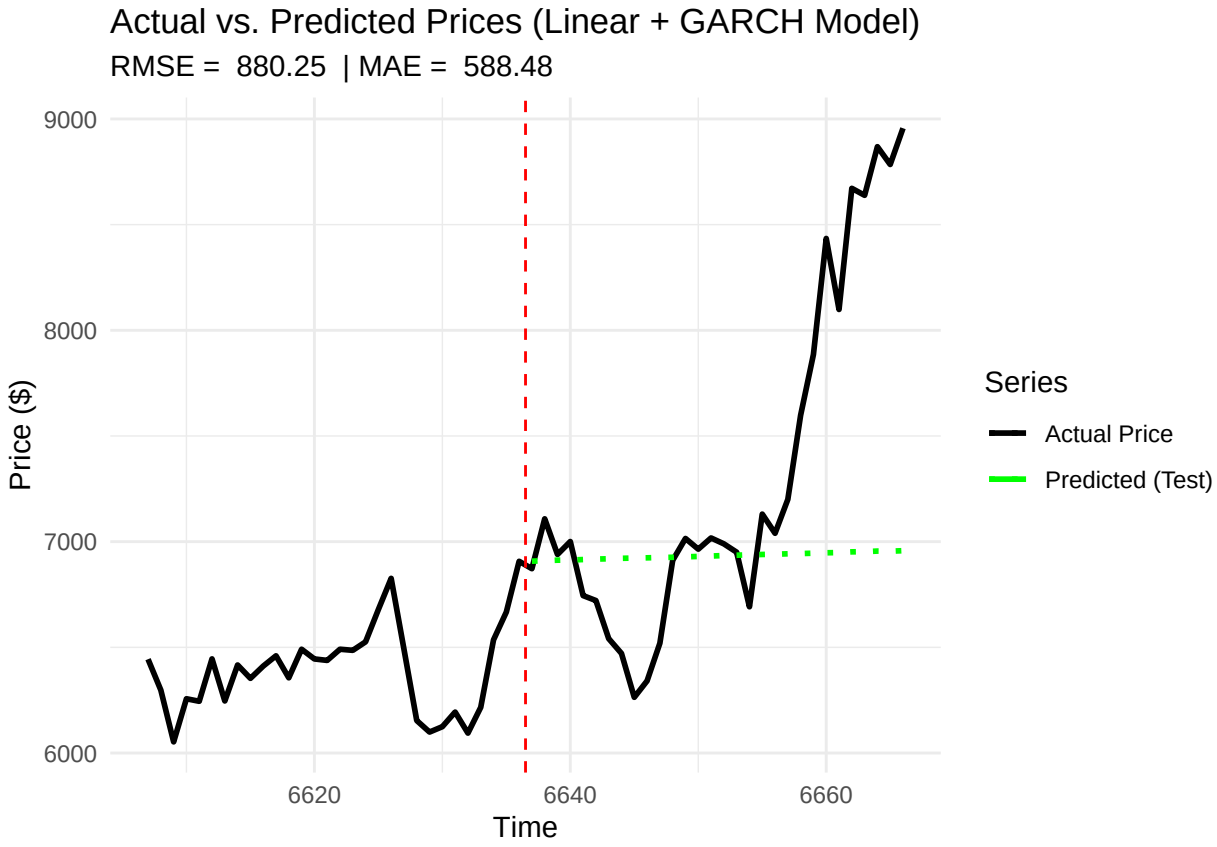
$$y_t = \sigma_t \epsilon_t \quad \sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

where
- $\Delta \log(\text{Price}_t)$ is the change in log price (i.e., log return),
- PRCP_t is the precipitation at time t,
- TAVG_t is the average temperature at time t,
- ϵ_t is the error term. - σ_t is the variance of the residuals.

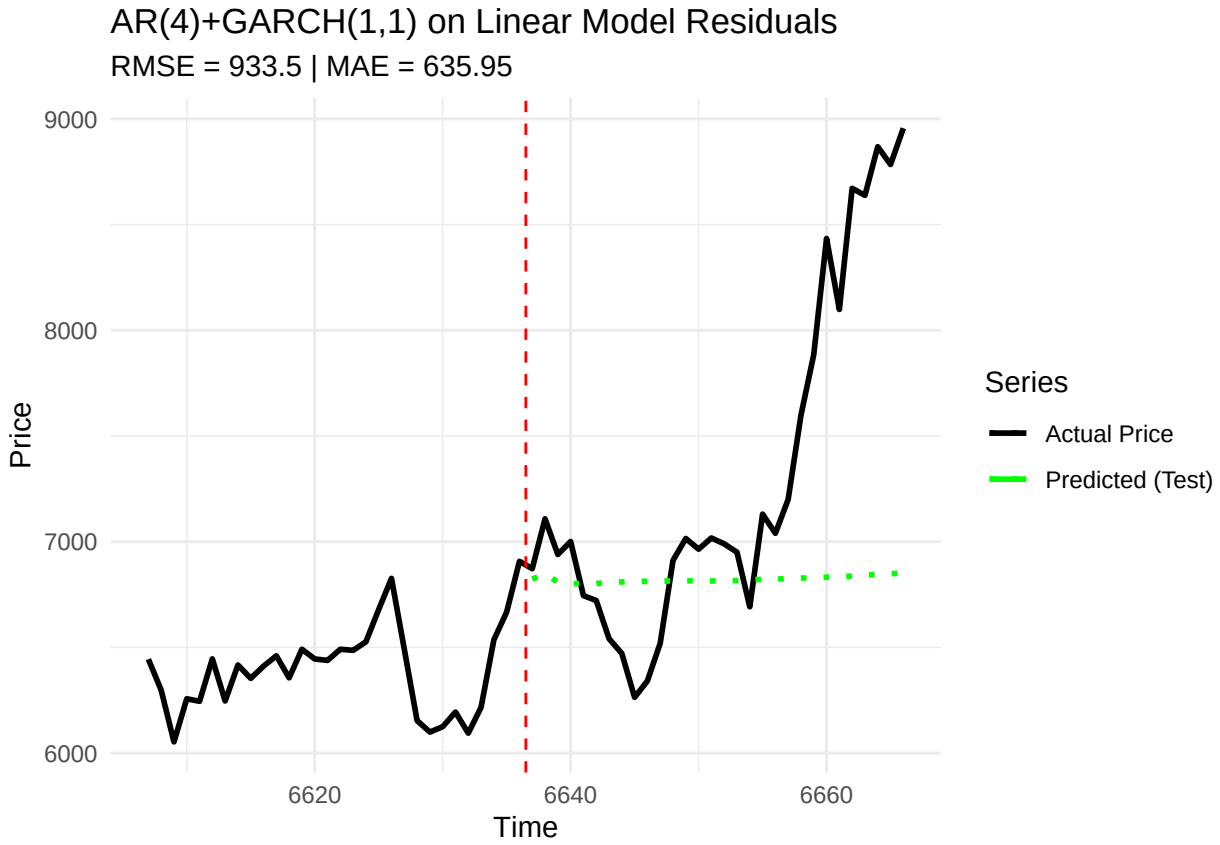
Below we find the results of the fit.

Table 2: Estimated GARCH(1,1) Coefficients

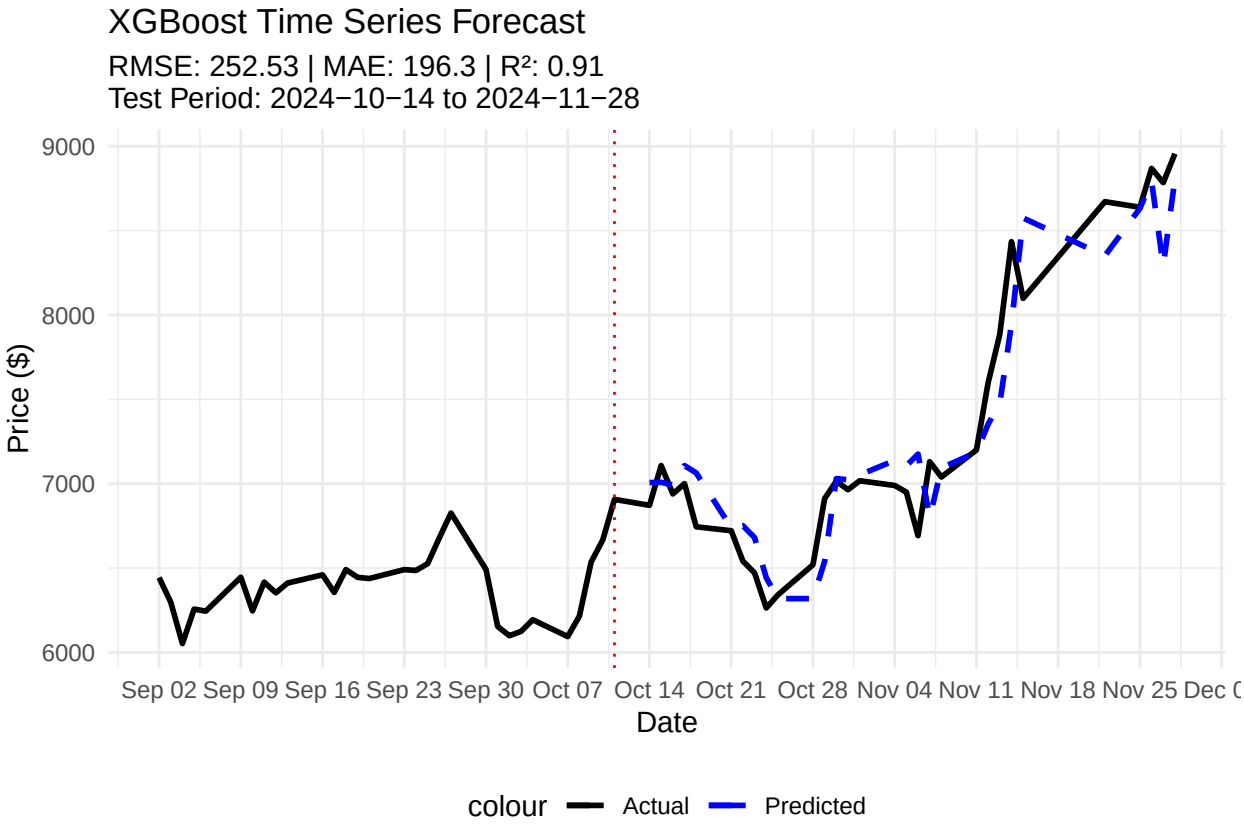
	x
omega	0.0001
alpha1	0.0984
beta1	0.7929



Now we fit the second model with an additional AR(1) component on the errors. The results are displayed below.



The last model we fit is XGBoost.



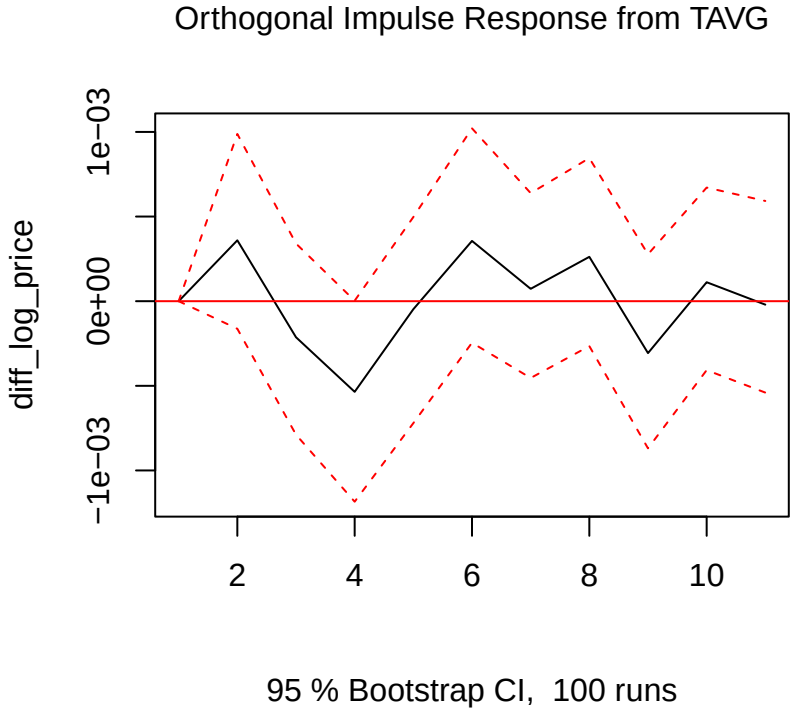
In comparing the results of the models used to predict cocoa prices, The XGBoost model had a significantly better prediction than the time series models.However within the time series models, the Linear Regression model with GARCH(1,1) demonstrated the best performance in capturing the overall trend of the price movements. It incorporates time-varying volatility, provided more accurate predictions during volatile periods. This model's lower RMSE and MAE suggest that it was more adept at modeling the conditional variance in the price data.The other models have done done moderately well but their RMSE values show room for improvement.The other models have done done moderately well but their RMSE values show room for improvement

Analyzing the trends of the forecasted series from the models, we expect an initial rise in the immediate future followed by a sharp decline in the growth rate of Cocoa Future Prices due to its volatile nature. From the estimates of the model, we conclude that the main influencers of the growth rate of Cocoa Future Prices include past growth rates, past volatility of growth rates, and temperature fluctuations in Ghana. The high estimated beta of GARCH(1,1) of a value close to 1 also indicates that the volatility of the financial series is rather persistent, where there is a high potential for sustained periods of fluctuations and risk for Cocoa Future Prices and that shocks in the financial market will have dominating and significant impacts on the series.

Long Term Trends

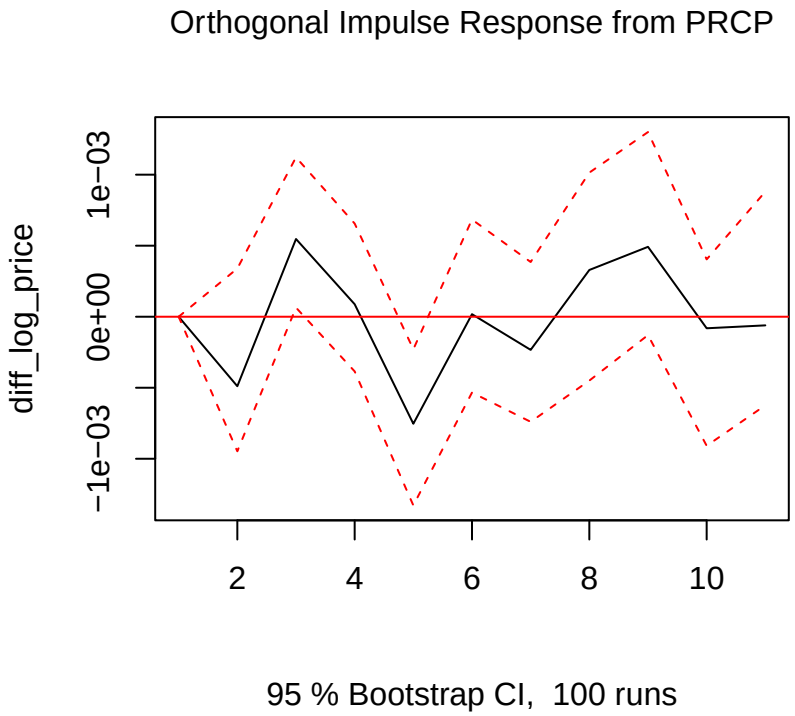
```
##
##  Augmented Dickey-Fuller Test
##
```

```
## data:  data$TAVG
## Dickey-Fuller = -7.1482, Lag order = 18, p-value = 0.01
## alternative hypothesis: stationary

## $selection
## AIC(n)  HQ(n)  SC(n) FPE(n)
##     12      8      8     12
##
## $criteria
##                  1           2           3           4           5
## AIC(n) -5.966416538 -6.08895823 -6.160817242 -6.201877262 -6.210110783
## HQ(n)  -5.964297350 -6.08542625 -6.155872470 -6.195519699 -6.202340428
## SC(n)  -5.960282206 -6.07873434 -6.146503802 -6.183474268 -6.187618235
## FPE(n)  0.002563411  0.00226777  0.002110528  0.002025624  0.002009015
##                  6           7           8           9          10
## AIC(n) -6.221603782 -6.229198713 -6.238645772 -6.237977933 -6.239929128
## HQ(n)  -6.212420636 -6.218602775 -6.226637042 -6.224556411 -6.225094815
## SC(n)  -6.195021680 -6.198527056 -6.203884561 -6.199127168 -6.196988809
## FPE(n)  0.001986057  0.001971031  0.001952498  0.001953802  0.001949994
##                 11          12
## AIC(n) -6.240648051 -6.241113104
## HQ(n)  -6.224400946 -6.223453206
## SC(n)  -6.193618178 -6.189993676
## FPE(n)  0.001948592  0.001947686
```




```
##
```

```
##  Augmented Dickey-Fuller Test
##
## data:  data$PRCP
## Dickey-Fuller = -14.015, Lag order = 18, p-value = 0.01
## alternative hypothesis: stationary

## $selection
## AIC(n)  HQ(n)  SC(n) FPE(n)
##      8      6      4      8
##
## $criteria
##                  1             2             3             4             5
## AIC(n) -1.059343e+01 -1.061964e+01 -1.062251e+01 -1.063773e+01 -1.063995e+01
## HQ(n)  -1.059131e+01 -1.061610e+01 -1.061756e+01 -1.063137e+01 -1.063218e+01
## SC(n)  -1.058730e+01 -1.060941e+01 -1.060820e+01 -1.061933e+01 -1.061746e+01
## FPE(n)  2.508015e-05  2.443151e-05  2.436144e-05  2.399346e-05  2.394017e-05
##                  6             7             8             9            10
## AIC(n) -1.064239e+01 -1.064193e+01 -1.064326e+01 -1.064250e+01 -1.064262e+01
## HQ(n)  -1.063321e+01 -1.063133e+01 -1.063126e+01 -1.062908e+01 -1.062779e+01
## SC(n)  -1.061581e+01 -1.061126e+01 -1.060850e+01 -1.060365e+01 -1.059968e+01
## FPE(n)  2.388187e-05  2.389287e-05  2.386102e-05  2.387932e-05  2.387637e-05
##                 11            12
## AIC(n) -1.064296e+01 -1.064266e+01
## HQ(n)  -1.062671e+01 -1.062500e+01
## SC(n)  -1.059593e+01 -1.059154e+01
## FPE(n)  2.386826e-05  2.387540e-05
```



To clearly see dynamic relationships between cocoa prices and average temperature and rainfall we use VAR analysis. Using the Dickey-Fuller test we ensure all the variables are stationary. We used first differenced log of price to ensure stationarity. The VAR model allows us to see different lags of variables hence capturing feedback effects over time. More importantly we look at the Impulse Response Functions derived from the VAR reveal how shocks to temperature and rainfall influence cocoa price over the next 10 months. We have also used VARselect to identify the optimal lag to use for the impulse functions. As seen in the figures below, the black solid line represents the estimated response of price to a one-time shock in temperature or rainfall. The red lines are the 95% confidence intervals. The red line at zero is the baseline, indicating no effect. If the black line deviates from this line outside the confidence intervals the effect is statistically significant.

The impulse response for precipitation is negative in the first period response, so an decrease in precipitation causes a small decrease in cocoa price. The opposite is shown in temperature with a positive response in the first period, so a increase in temperature causes an increase in price. However, in the long term the response oscillates and does not settle on a single sign, suggesting no clear persistent effect. The confidence intervals are large and so there is no strong predictable long-term impact of temperature on cocoa prices.

Discussion and Conclusion

Our analysis reveals complex interactions between climate variables and cocoa price volatility. The GARCH(1,1) model effectively captured the persistent volatility in cocoa prices. However, when examining the VAR impulse response functions, we found that temperature shocks produce unstable and unpredictable effects over longer time periods.

Our study faces several limitations worth noting. First, our regression framework with time series error components assumes known future values of weather variables, which is unrealistic for practical forecasting scenarios. This constrains our ability to predict prices under changing climate conditions. Second, while our classical time series models provide valuable insights, they may not capture nonlinear relationships as effectively as machine learning approaches like XGBoost, potentially affecting forecast accuracy. Third, our dataset lacks information on cocoa crop diseases, which represent significant supply-side disruptions and likely contribute to the dramatic price spikes observed recently, with cocoa prices reaching nearly $10,000 per metric ton in 2024.

Future research could benefit by employing more sophisticated machine learning techniques to model complex interactions, integrating data on plant diseases affecting cocoa crops, and expanding the model to include exchange rate dynamics. As cocoa is a globally traded commodity, exchange rate fluctuations—particularly between the US dollar and Ghanaian cedi, as highlighted in the literature—could further elucidate price volatility by capturing currency-driven cost variations in international trade.

Despite these constraints, our combined linear-GARCH model demonstrates promising performance for short-term price forecasting, while the VAR analysis clarifies how weather-related shocks propagate through the cocoa market system. Industry stakeholders should remain aware of cocoa's inherent price volatility and develop adaptive strategies to navigate market uncertainties.

References

1. Sukiyono, Ketut & Nabiu, Musriyadi & Sumantri, Bambang & Novanda, Ridha & Arianti, Nyayu & Sriyoto, & Yuliarso, M. & Badrudin, Redy & Romdhon, Muhamad & Mustamam, H.. (2018). Selecting

an Accurate Cacao Price Forecasting Model. Journal of Physics: Conference Series. 1114. 012116. 10.1088/1742-6596/1114/1/012116.

2. Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica, 50(4), 987–1007
3. Lintang, K. L., & Kurniawan, M. L. A. (2023). Vector Autoregressive (VAR) Analysis of Cocoa Export in Indonesia. Journal of Economics Research and Social Sciences, 7(2),192-205. 4.Ketut Sukiyono et al 2018 J. Phys.: Conf. Ser. 1114 012116
4. Alori, Alaba & Kutu, Adebayo. (2019). Export Function of Cocoa Production, Exchange Rate Volatility and Prices in Nigeria. Journal of Economics and Behavioral Studies. 11. 1. 10.22610/jebs.v11i2(J).2813.