Forecasting Cocoa Futures Prices with Time Series Models

Kevin Li, Harrison Park, Justin Wu

April 3, 2025

1 Introduction

Global cocoa markets are subject to frequent fluctuations in price, driven by a multitude of factors such as weather variability, macroeconomic conditions, and evolving supply and demand dynamics. Because cocoa serves as a key ingredient in the chocolate and confectionery industries—sectors that generate billions of dollars in annual revenue—understanding and forecasting cocoa price movements carries considerable real-world significance. Accurate forecasts help stakeholders across the cocoa supply chain make more informed decisions about production, inventory management, and risk mitigation. Moreover, the ability to incorporate environmental data (e.g., temperature and precipitation) into price forecasting models has become increasingly important in light of climate change, which can alter cocoa yields and disrupt established market patterns.

Motivated by these issues, this report presents a comprehensive study on forecasting cocoa futures prices using both classical time series techniques and modern machine learning approaches. To deepen our analysis, we additionally incorporate climate variables and macroeconomic indicators—such as the USD/Ghana exchange rate and consumer price indices (CPI)—as potential predictors. By capturing the relationships between price, climate, and economic factors, we aim to achieve more robust forecasts than could be attained by relying on price history alone.

Our objectives in this work include:

- **Data Collection and Preparation:** Gather relevant monthly cocoa futures prices, temperature, precipitation, and macroeconomic variables. Preprocess these data by cleaning missing observations and aligning dates to ensure consistency across all time series.
- **Exploratory Data Analysis (EDA):** Identify trends, seasonality, and correlations among the variables. This step provides foundational insights that guide model selection and parameter tuning.
- **Model Development and Comparison of Forecasting Results:** Implement a variety of forecasting methods—including SARIMA/SARIMAX, GARCH-type models, and recurrent neural networks (RNNs)—to capture both the linear and nonlinear structure in the data. Compare the forecasts yielded by these approaches using testing metrics such as RMSE and MAPE.
- **Determination of Best Model and Discussion of Next Steps:** Conclude which model is best by comparing the values of testing metrics, and outline future steps for improving the final model (e.g., by integrating additional commodity prices or regional climate data).

Throughout this analysis, we encounter a number of key challenges. First, cocoa price data can exhibit non-stationary behavior and high volatility, requiring careful application of differencing and volatility modeling techniques (e.g., GARCH). Second, climate data for certain regions include missing observations; deciding on an appropriate imputation method is not trivial, as naive approaches can distort the underlying signals. Third, concept drift—where the data-generating process changes over time—can limit the long-term accuracy of models trained on historical data. Finally, while advanced methods such as neural networks might potentially yield superior predictive performance, they can be computationally expensive and less interpretable than classical statistical models.

The remainder of this report proceeds as follows. In Section 2, we review key literature on various kinds of time series models and machine learning models, and contextualize the relevance of climate and macroeconomic variables for cocoa price forecasting. In Section 3, we introduce our overall methodology, including data preprocessing techniques and model formulations. Section 4 provides detailed EDA results, highlighting core trends and stationarity checks. In Section 5, we present forecasting outcomes from each model and offer a comparative analysis of their performance. Finally, in Section 6, we discuss the main findings, reflect on limitations, and propose directions for future research.

2 Literature Review

2.1 Atejombi and Olaleye: Efficacy of ARIMA model on cacao production.

A study conducted by Atejombi and Olaleye [1] on cacao production in Nigeria from 1984-2014 modeled the time series as an ARIMA(1,1,1) model on the log-transform of the cacao production series. The estimated parameters were found at 1% significance level. The findings indicate that cacao production in Nigeria is forecasted to steadily decline going into 2025. Although our study focuses on cocoa commodity prices, we expect the time series equivalent of production levels to be correlated with commodity price. The efficacy of ARIMA models in modelling cacao production levels motivates us to investigate the ARIMA model in our own analysis.

2.2 Aklimawati and Wahyudi: Estimating the Volatility of Cocoa Price Return with ARCH and GARCH Models

A study by Aklimawati and Wahyudi [2] examines the use of ARCH/GARCH family models to forecast volatility in cocoa prices. Focusing on daily cocoa futures prices from the ICE U.S. exchange, the authors compare various GARCH-type models in capturing price fluctuations under market liberalization. The study finds that a simple GARCH(1,1) model provides the best fit for cocoa price return volatility, satisfying diagnostic checks (no remaining ARCH effect, uncorrelated and normally distributed residuals). This indicates that cocoa price volatility can be effectively predicted by GARCH models, which capture clustering in variance. For our own analysis, we aim to ascertain whether heteroskedacity in our data can be appropriately handled through the implementation of GARCH models such as those used in this paper.

2.3 Tabe-Ojong, Guedegbe, and Glauber: Soaring Cocoa Prices – Diverse Impacts and Implications for Key West African Producers

A study conducted by Tabe-Ojong, Guedegbe, and Glauber [3] analyzes the drivers behind the recent surge in cocoa prices and its effects on West African producers. Their research points to a significant drop in cocoa output in late 2023, attributed to adverse climate events. The study finds that the price spike was triggered by the dual effects of climate change and weather disasters, which brought erratic rainfall and higher temperatures, fueling outbreaks of pests and diseases and substantially reducing cocoa yields. The relevance of this study to our research lies in its emphasis on how exogenous factors—especially those relating to *climate variability*—directly influence cocoa supply and thus prices. The study thus informs our time series forecasting approach by reminding us that we will need to account for climate behaviour when predicting cocoa price movements.

2.4 Anderson and Gough: Accounting for missing data in monthly temperature series

A study by Anderson and Gough [4] tested how missing daily values affect monthly temperature averages (climate normals). They found that allowing a few days to be missing (and effectively treating them as average) introduces very small error. Under the "3/5 rule" (no more than 3 consecutive or 5 total days missing), the bias in the monthly mean was only about 0.06–0.07 standard deviations. In practical terms, filling up to a handful of missing daily temperatures with climatological means has a negligible impact on the monthly average. The authors conclude that such small gaps can be safely accommodated with minimal error. This study provides precedent for applying the imputation technique of using monthly averages to infill missing temperature data, since it found that the long-term mean for that month is a reasonable surrogate for a few missing days. Since there is missing temperature data in the datasets we are considering for our analysis, we will strongly consider using this technique to fill in missing values.

2.5 Widmer and Kubat: Learning in the Presence of Concept Drift and Hidden Contexts

Widmer and Kubat [5] proposed the idea that target concepts depend on hidden contexts, this can be seen in how weather predictions can vary based on the season. Concept drift defines a change in this hidden context; intuitively, this means that patterns in a variable of interest can change over time, which is especially prevalent in time series application where the historical context may not remain relevant when unforeseen events alter the hidden context. Applying this idea to cocoa futures prices, we can imagine that factors driving these prices represent the hidden context. Over time, it is very likely that consumer habits can change, or the market becomes more speculative, leading to a concept drift. To address this challenge, we can consider training our models on a more recent subset of the data to ensure that the hidden context remains relevant for prediction tasks and forecasting future prices.

2.6 Hewamalage, Bergmeir, and Bandara: Recurrent Neural Networks for Time Series Forecasting: Current Status and Future Directions

A study by Hewamalage, Bergmeir, and Bandara [6] examined the use of recurrent neural networks (RNN) for time series forecasting. The paper compared RNNs against the exponential smoothing (ETS) and autoregressive integrated moving average (ARIMA) models. Furthermore, the authors developed a set of guidelines and best practices for applying RNNs for time series forecasting, including data preprocessing and RNN architecture. In particular, the authors noted that RNNs are more difficult to fit compared to ETS and ARIMA, with higher computation costs and challenges in capturing seasonality in the data. However, RNNs can effectively capture information across multiple series, resulting in competitive forecasting accuracy compared to the other models. Building upon the valuable guidelines provided in the paper, we aim to leverage RNNs as a key benchmark for our research; contrasting their performance with established statistical forecasting methods. Additionally, since RNNs are universal approximators, they offer the flexibility to model complex nonlinear relationships without strong assumptions about the underlying data-generating process.

2.7 Alori and Katu: Export Function of Cocoa Production, Exchange Rate Volatility and Prices in Nigeria

A study by Alori and Katu [7] analyzes the impact of inflation and exchange rate volatility on Nigeria's cocoa export. The variables considered were the value of cocoa exports, exchange rate, domestic cocoa output, real GDP, and the consumer price index. The authors found that none of the variables were stationary and required taking the first difference. The authors specified an export-supply model, where the value of the cocoa export at a given time t is represented by a linear relationship of the other variables; this is estimated using ordinary least squares. The authors also tested a structural vector auto-regressive (SVAR) model in order to examine the effects of shocks in one of the variables on the overall cocoa export value over the following months. The authors found that CPIs (Consumer Price Indices), GDP, cocoa output, exchange rate, and international cocoa prices were all statistically significant in explaining the variation in cocoa export value, with a R^2 value of 0.91. Furthermore, the SVAR analysis showed that a shock in the exchange rate leads to the most volatility in cocoa export earnings over a 12 month window. Based on the study, we will incorporate economic indicators such as the exchange rate and CPI as possible exogenous variables in our methodology.

2.8 Lestari and Dini: Forecasting The Price Of Shallots And Red Chilies Using The ARIMAX Model

Lestari and Dini [8] forecast spice prices using an ARIMAX model which incorporates exogenous variables. They recommend using exogenous variables whose values are capable of being *known in advance*; in particular, they report that a model incorporating *exogenous monthly indicators* successfully captured seasonal price fluctuations tied to the calendar, leading to improved predictive results. This study motivates us to consider using exogenous variables in our own ARIMA/SARIMA model construction process.

3 Methodology

3.1 Data Preprocessing

Before applying any time series models, the data must be cleaned and prepared for analysis. The preprocessing steps include handling missing data, aligning dates, and performing feature engineering. Missing values in the cocoa futures price data will be addressed using linear interpolation or forward/backward filling. For climate data, missing precipitation (PRCP) and missing temperature values (TAVG, TMAX, TMIN) will be averaged based on monthly bins, in alignment with the practice discussed in Section 2.4 of our literature review [4]. To ensure consistency, the cocoa futures price data and climate data will be aligned by date. Additionally, feature engineering will be performed to enhance predictive power. This may include aggregating data into monthly averages, along with transforming variables if it is necessary for the chosen model. External predictors such as macroeconomic indicators (e.g., GDP, inflation), will have lagged versions created to account for delayed effects on cocoa prices. Finally, stationarity checks will be conducted, and transformations such as differencing or logarithmic transformations will be applied if non-stationarity is detected. For forecasting purposes, we take the validation dataset to be the last 4 recorded months across our variables, and take the rest to be our training set.

3.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) will be conducted to understand patterns and relationships in the data. Time series plots of cocoa prices and climate variables will be examined to identify trends, seasonality, and outliers. Seasonal decomposition, including STL decomposition, will be applied to separate the cocoa price series into trend, seasonal, and residual components, aiding in model selection and seasonality identification. The ACF/PACF plots of the ICCO prices and other predictors will be examined to further investigate stationarity. Cross-correlation analysis will be performed to explore the relationship between cocoa prices and climate variables, incorporating appropriate lags to account for delayed effects.

3.3 Model Selection, Description, and Justification

Based on the insights from EDA, various classical time series models, as well as a neural network model, will be considered for forecasting.

3.3.1 SARIMA/SARIMAX Model

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model will be implemented, with parameters (p, d, q) and seasonal parameters (P, D, Q, S). Taking inspiration from Section 2.8 of our literature review [8], we will also attempt to incorporate exogenous predictors and implement a SARIMAX model, since Lestari & Dini achieved superior predictive results with this kind of model in their paper (relative to ordinary SARIMA). In order to weight the most recent trends in the ICCO price dataset effectively, we will also choose to window the training set to observations recorded from 2020 onwards in order to build more accurate forecasts.

To determine a final SARIMA model, we will try to consider an initial set of potential values of p, d, q and P, D, Q, S for the model, and use backwards AIC selection to find the model with the lowest AIC. We will then perform residual diagnostics to ensure that residual assumptions are satisfied for the model.

3.3.2 Hybrid Regression-ARMA-GARCH Model

Motivated by [5], we investigate the impact of exogenous predictors through a time series regression model, and fit an ARMA model to the residuals. If heteroscedasticity (or another violation of residual assumptions) is detected within the ARMA model residuals, a GARCH model will be fit to the differenced log of the ARMA residuals (in order to Taylor approximate the return of the residuals); this model will then be transformed back so it can be summed with the above prediction results for final forecasting. We may also attempt to fit a GARCH model even if no issues are detected with the original ARMA model, since it is an interesting opportunity to apply our knowledge of material covered in the latter half of the course and is motivated by [2].

3.3.3 Neural Network Models

In addition to classical time series models, we will consider variations of recurrent neural networks (RNN) for forecasting as a separate benchmark to compare against the aforementioned approaches. In particular, we are interested in RNNs' abilities to capture long-term temporal dependencies and incorporate exogenous variables. Following the guidelines in Section 2.6 of our literature review [6], we will utilize a stacked architecture with LSTM (long short-term memory) cells with dropout. The learning rate will be tuned dynamically using the ADAM optimizer.

Before training, we apply several preprocessing and feature engineering steps to enhance model performance. The time series will first be deseasonalized based on insights gained from exploratory data analysis. To enable the network to learn temporal patterns, we will convert the data into a supervised learning format using a moving window approach, where each window consists of lagged sequences of the input variables.

Feature engineering plays a crucial role in improving the model's predictive capabilities. In addition to the raw time series and exogenous covariates, we construct lag features (e.g., previous values of the target variable), rolling statistics (such as moving averages and standard deviations), and differenced versions of the series to capture short-term trends and volatility. These engineered features are designed to provide the network with richer temporal context and help it learn complex interactions across time. All input features are normalized prior to training to ensure stable convergence and efficient learning.

Key hyperparameters, including the window size, network depth, and size of the hidden state are determined via grid search to optimize forecasting performance through a validation subset of the training data. On the following page, we present a diagram of the overall network architecture.

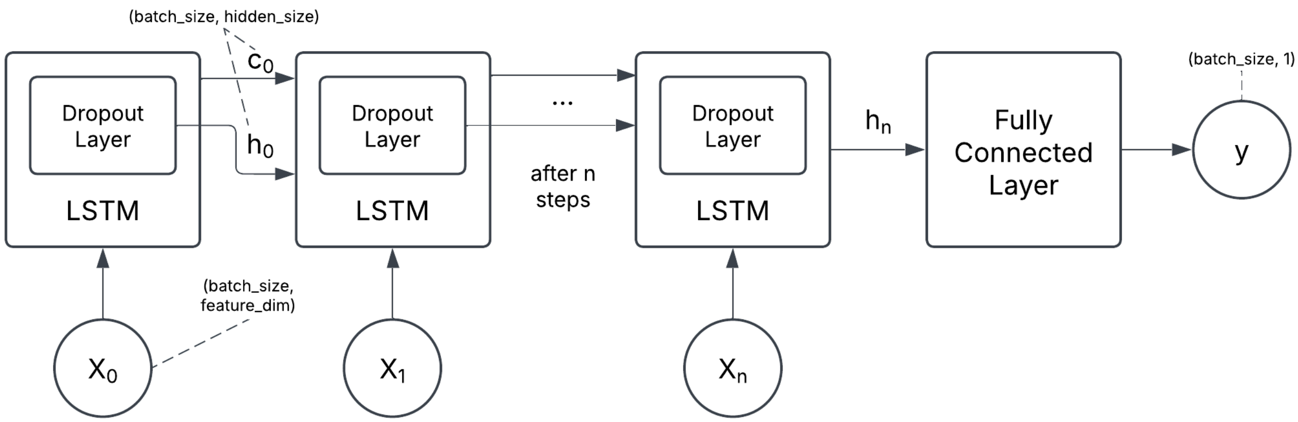

Figure 1: *Recurrent Neural Network Architecture.*

In the previous diagram, X_0 through X_n represents $n+1$ months of data, as determined by the window size. For example, suppose the window size was 12, then to predict the ICCO Cocoa furture price for April 2025, we would pass in data from April 2024 to March 2025. The diagram illustrates the case with a depth of 1, *i.e.,* only one LSTM cell; a larger depth means multiple LSTM cells are stacked on top of each other. Note that the network shown is *unrolled*, what actually happens is that (using the earlier example) we pass in the features for April 2024 (X_0), compute the hidden state h_0 and cell state c_0 using the trained weights. Next, these states are used when we pass in the features for May 2024 (X_1), and we update the states with the *same* weights. This is repeated until we pass in all 12 months of features, and the output hidden state is passed through a fully connected layer that gives us 1 output prediction. We update the weights using backpropagation through time (BPTT) with median absolute error (MAE) as the loss function. BPTT unfolds the network across all 12 time steps, computing gradients at each step and propagating the error backward through both time and layers to update the shared LSTM weights accordingly.

3.4 Model Validation

Model validation will be conducted to assess the forecasting performance. The dataset will be split into a training set (majority of the data) and a test set (the last few data points). The models will be trained on the training set and validated on the test set, ensuring a sequential split to preserve temporal order. If large systematic shifts in the series' behavior are observed after a certain time point, the dataset may need to be windowed to only train on the most recent, relevant behaviour as per the discussion on concept drifts in [5]. Forecast accuracy will be evaluated using metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). The model achieving the lowest error metrics on the test set will be selected as the final model.

4 Data

In this section, we describe the data used in our analysis and the method we used to split our dataset into training and testing sets. We then present the results of Exploratory Data Analysis (EDA) on the training dataset, including summary statistics, visualizations, and stationarity checks. Note that due to limited space, we cannot present the EDA results for every variable included in our model, and will therefore limit our discussion to those variables which we feel are most relevant for forecasting.

4.1 Description of Primary Data

4.1.1 Climate Data from Ghana & Cocoa Futures Prices

Our dataset contains monthly climate observations from Ghana from October 1994 up to February 2025. These data include several climate variables:

- `DATE`: The month and year of observation (MM-DD-YYYY format). Since our data is grouped by month, all these observations are of form MM-01-YYYY.
- `TAVG`: The average daily temperature across a given month (in Fahrenheit), observed at 2 meters above ground level.

- `PRCP`: The average daily precipitation across a given month (in millimeters).

In addition, our dataset includes the monthly average price of cocoa futures:

- `ICCO_price`: The ICCO (International Cocoa Organization) monthly average price for cocoa futures, in USD.

4.2 Additional Data and Rationale for Inclusion

4.2.1 Additional Macroeconomic Variables

In order to refine our predictive models and demonstrate creativity, we felt it was important to include economic indicators as a supplement to the data described above. As suggested in prior studies from our literature review (e.g. Section 2.7 [7]), macroeconomic variables can influence cocoa prices. Thus, we have integrated:

- **USD/Ghana Exchange Rates.** In our literature review, Alori & Katu [7] find that a depreciation in the USD/Nigerian exchange rate is positively correlated with the volume of cocoa exports. Therefore, we suspect that a similar relationship holds for the exchange rate between the USD and Ghanian Cedi, and have thus decided to include this rate in our analysis. The monthly values of these rates, from October 1994 to February 2025, are stored in the `Ghana_official` column.
- **Global Consumer Price Indices (CPIs):** Alori & Katu [7] state that CPIs, which are estimates of the average change over time in the prices paid by consumers for a representative "basket" of goods and services—– including food, housing, clothing, transportation, medical care, and education—are positively correlated with the nominal value of cocoa exports. Therefore, we collected global CPI data for October 1994 to February 2025 from the DataBank website [9], and stored it as a `CPI_world` column in our dataset.

4.3 Splitting the Data into Training and Testing Datasets

After combining all relevant climate and economic data, we performed a temporal split:

4.3.1 Training Dataset

The training dataset covers the period from the earliest available date (10-01-1994) up to a specific cutoff date (10-01-2024). We use these observations to fit, tune, and compare our time-series models.

4.3.2 Testing Dataset

The testing dataset contains more recent data, covering the period from 11-01-2024 to 02-01-2025. This 4-month horizon was chosen based on a recommendation from Prof. Selvaratnam during lecture, although we initially considered a longer horizon of 24 months. The short horizon allows us to focus on near-term forecasting accuracy, which is particularly relevant for commodity markets like cocoa.

4.4 Results of Exploratory Data Analysis (EDA) on the Training Data

4.4.1 Initial Data Inspection and Summary Statistics

We began by reading in the data, converting the `DATE` column to an actual date object in R, and sorting it chronologically. Table 1 below shows a list of summary statistics for a subset of the variables from the training dataset.

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max	NA%
`CPI_world`	64.99	80.89	99.32	114.39	153.98	226.57	0.00
`TAVG`	76.27	78.92	81.41	81.09	82.83	86.90	0.83
`PRCP`	0.0000	0.1321	0.2357	0.2638	0.3295	2.1460	12.74
`Ghana_official`	0.1010	0.7695	1.4291	2.8772	4.3905	16.0085	0.00
`ICCO_price`	800.90	1559.50	2197.70	2271.40	2679.00	9876.60	0.00

Table 1: *Summary statistics for selected variables in the training dataset.*

Missing values were identified in both of the `TAVG` (average temperature) and `PRCP` (precipitation) columns, though the other variables had no missing values for all months. The percentage of missing data for temperature was quite small ($< 1\%$), while the percentage of missing data for precipitation was more significant (12.74%).

4.4.2 Monthly-Climatology Imputation for Temperature and Precipitation

To handle missing values for `TAVG` and `PRCP`, we adopted a monthly-climatology approach. Specifically, we:

1. Computed the average temperature (and precipitation) for each calendar month (January, February, ..., December) using all available data across all years.
2. Whenever an observation had a missing temperature or precipitation value, we replaced it with the corresponding month's mean (e.g., for a missing TAVG in October 1998, we used the average of all TAVG values recorded in October across the dataset).

This approach is grounded in meteorological practice, where the typical climate of a given month is considered relatively stable across years (one source for this claim can be found in Section 2.5 of our literature review: [4]).

4.4.3 Time Series Plots of Primary Climate Variables

The following figure shows time series plots of `TAVG` and `PRCP` values, after sorting chronologically.

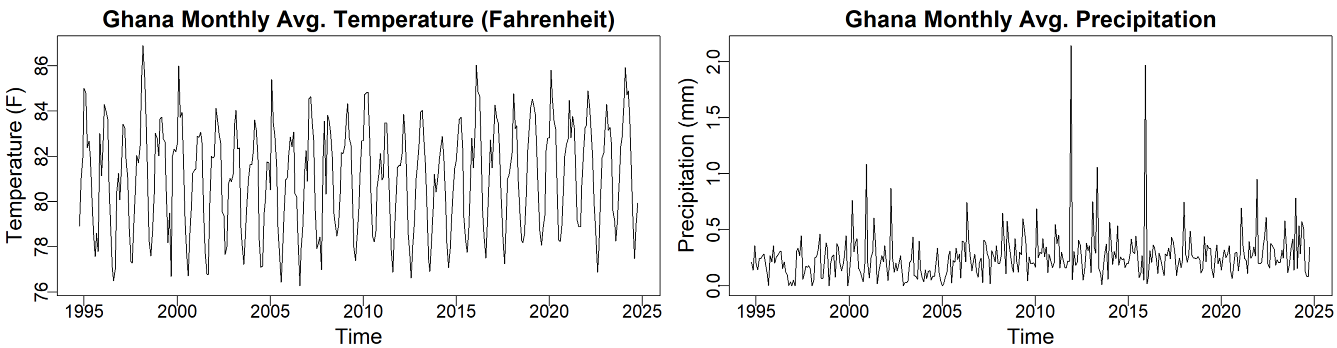

Figure 2: *Time Series of Monthly Average Temperature & Monthly Precipitation.*

From the above, we see a clear wave-like pattern in the `TAVG` plot which appears to repeat yearly (this could be indicative of seasonality). The `PRCP` plot also appears to exhibit periodic motion, though it is not as easy to determine if seasonality is present merely by looking at this plot. There is no long-term upward or downward trend in either plot, suggesting that both of these series could be stationary (although further investigation of the ACFs is necessary to make firm conclusions).

4.4.4 Time Series Plot of Cocoa Futures Prices

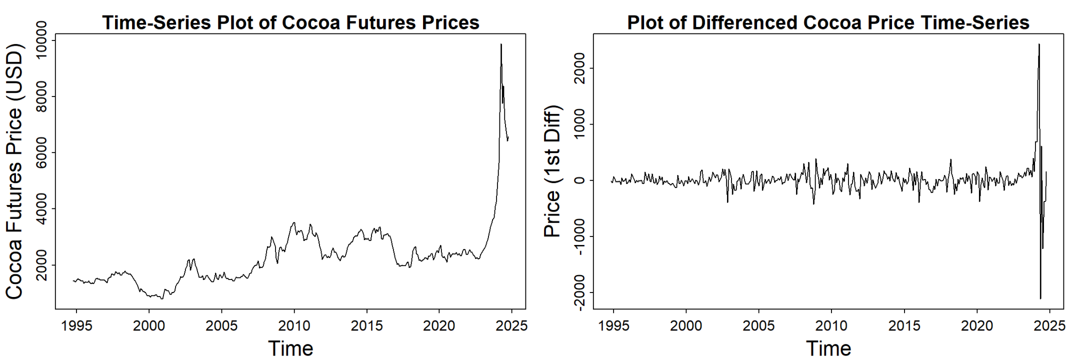

Figure 3: *Time Series Plots of ICCO Cocoa Futures Prices (left) & Differenced Cocoa Prices (right)*

From the above, we note that this time series is non-stationary, since futures prices have generally increased over time at a non-constant rate, with a more drastic rate of increase beginning sometime around 2022-2023.

On the other hand, the differenced plot generally appears more stationary, although not completely so (since in the most recent observations beginning in around 2023, the differenced price starts to fluctuate drastically).

4.4.5 Time Series Plots of Additional Economic Variables

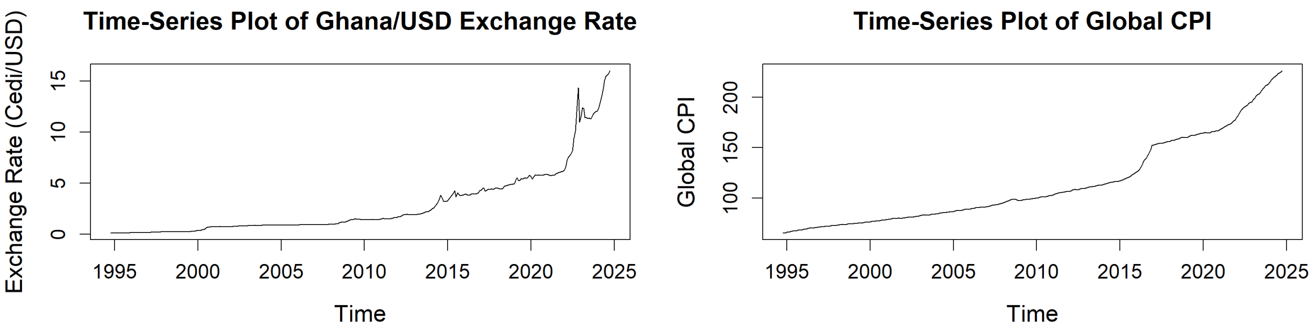

Figure 4: *Time Series Plots of Ghanaian Cedi/USD & Global CPI.*

From the above, we see that both of these variables exhibit a notable upward trend as time progresses, indicating non-stationarity. For the Ghana/USD Exchange rate plot, we note a drastic increase which seems to have begun in 2022-2023. We expect other included currencies to follow a similar trend.

4.4.6 STL Decomposition of Time Series

After the above plots were created, we used STL (Seasonal and Trend decomposition using Loess) to decompose each time series into trend, seasonal, and remainder components:

- **Cocoa Price (ICCO_price)**: The decomposition revealed an upward/downward long-term trend and a moderate seasonal influence (though the seasonality is less pronounced than in climate variables).
- **Temperature (TAVG)**: Strong seasonality was evident, with a well-defined annual cycle.
- **Precipitation (PRCP)**: Also showed strong seasonality with a well-defined annual cycle.
- **Ghana/USD exchange rate (Ghana_official)**: Exhibited a clear upward trend over time, reflecting local currency depreciation.

Our STL plots can be found in the Appendix.

4.4.7 ACF/PACF Plots, Augmented Dickey-Fuller Test, and Differencing Investigation for Primary Climate Variables

Next, we performed more thorough stationarity checks of the primary climate variables by analyzing their ACFs/PACFs, conducting the Augmented Dickey-Fuller (ADF) test, and employing the `ndiffs` function in R to see how many differences were suggested to achieve stationarity. We performed additional seasonality checks by inspecting ACFs of key variables for indicative patterns. To save space, we relegate plots of the majority of the ACFs/PACFs to the Appendix.

- **Precipitation**: The ADF test returned a p-value of 0.01 (which suggests stationarity), but `ndiffs` suggested one difference was required. Overall, plotting the differenced series showed a more stable mean/variance (see Figure 17 in the Appendix).
- **Average Monthly Temperature**: `ndiffs` suggested zero differences for stationarity, which is corroborated by the ADF returning a p-value of 0.01. The ACF strongly indicated seasonality, through a sinusoidal pattern with peak ACF values nearing 1 when lags were exactly 1 year apart. A figure is shown below.

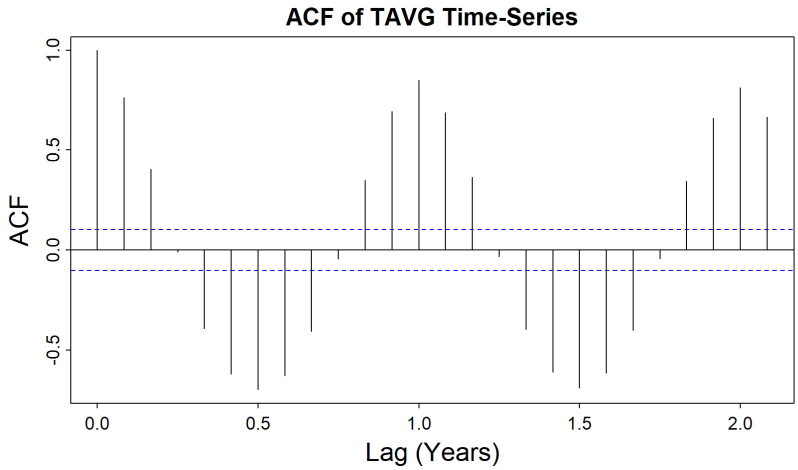

Figure 5: *ACF Plot of Ghanaian Monthly Average Temperature Time Series*

4.4.8 ACF/PACF Plots, Augmented Dickey-Fuller Test, and Differencing Investigation for Cocoa Futures Prices & Additional Economic Variables

- **Cocoa Futures Prices:** The ACF had values well above the significance band at high lags, and the ADF test returned a p-value of 0.322; these results both indicate non-stationarity. The ACF of the first-order differenced series had values well below the significance band for all positive lag values, indicating that a first-order differencing transformation could lead to stationarity; this is further supported by `ndiffs` suggesting a first-order differencing. An image of the `ICCO_price` ACF and PACF is shown below, as well as an image of the ACF/PACF of the differenced series.

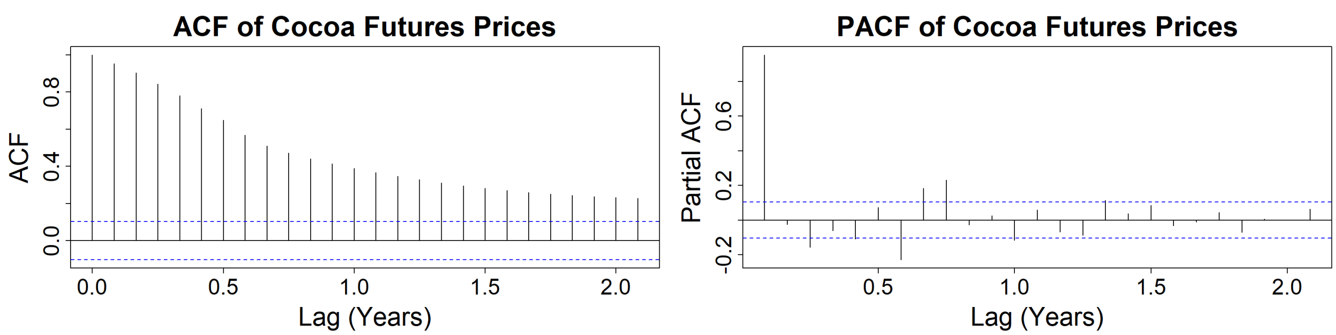

Figure 6: *ACF/PACF Plots of Cocoa Futures Prices*

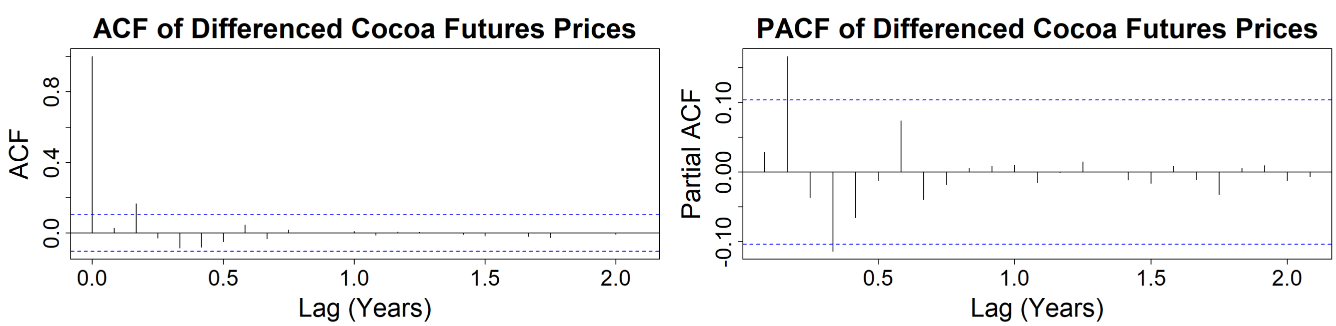

Figure 7: *ACF/PACF Plots of Differenced Cocoa Futures Prices*

- **USD/Ghanian Cedi Exchange Rate:** The p-value returned by this test was 0.99, indicating non-stationarity of cocoa prices. This was corroborated by large values of the ACF occurring at high lags. `ndiffs` suggested 2 differences to achieve stationarity.

- **Global CPI:** The ACF had values well above the significance band at high lags, indicating non-stationarity. This was supported by the ADF test statistic returning

4.5 Cross-Correlation Analysis of the Training Data

To gain further insights on the relations between important variables in our dataset, we examined cross-correlations (via `ccf` in R) between the following variables.

- **Price vs. Temperature**: Some correlation when lags ranged between 2 and 5 months, suggesting that temperature changes might lead cocoa price by a few months.
- **Price vs. Precipitation**: Large positive correlation at most lags indicates that higher precipitation might generally be associated with higher prices, though we should be careful about jumping to this conclusion.

The cross-correlation plots we analyzed can be found in the Appendix (Figure 21).

5 Forecasting and Results

5.1 Results of Hybrid Regression-ARMA-GARCH Model

This analysis forecasted ICCO cocoa prices using a hybrid regression-ARMA-GARCH model trained on post-2020 data to prioritize recent trends. The test set is chosen as the last 4 months of the dataset to preserve temporal order for forecasting purposes. The key predictors included are the exchange rates of the leading cocoa producers (Ghana, Côte d'Ivoire, Nigeria), the minimum, maximum and average temperatures of Ghana, and the inflation indices (CPI), along with the lagged versions, square and cubic transformations of the aforementioned. When constructing lagged predictors, we only analyze predictors with a lag of $h = 1$ month, as the EDA indicated that correlations were found only for small lag values.

Beginning with a robust full model that included all transformed exogenous regressors, a backwards stepwise-selected linear regression formed the base regression model using 30 degrees of freedom (all coefficients with p-value < 0.01), with an AIC value of 746.12.

Its residuals were further modeled with ARMA for temporal patterns and GARCH for volatility to improve prediction intervals. After initial regression fitting, an AR(1) model is fit to the residuals. The residual diagnostics of this AR(1) model are shown in Figure 8.

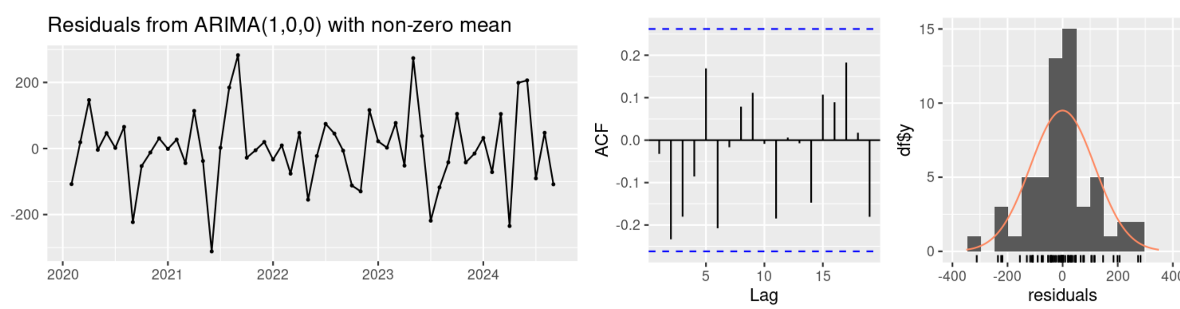

Figure 8: *Residual diagnostics of the AR(1) model fit to the linear regression.*

Observing the residual diagnostics, they appear to approximately resemble white noise, although we note that our choice to window the data past 2020 may have masked the true price volatility present in the full training data. As such, a GARCH(1,1) model is used on the ARMA residuals to account for unobserved volatility.

The final forecasts with volatility adjusted prediction intervals are shown in Figure 9, where the last 4 months of the ICCO price data are used for model validation. The performance metrics of this model's forecasting results are

$$\text{RMSE} = 1308.1860, \qquad \text{MAPE} = 13.1604\%. \tag{1}$$

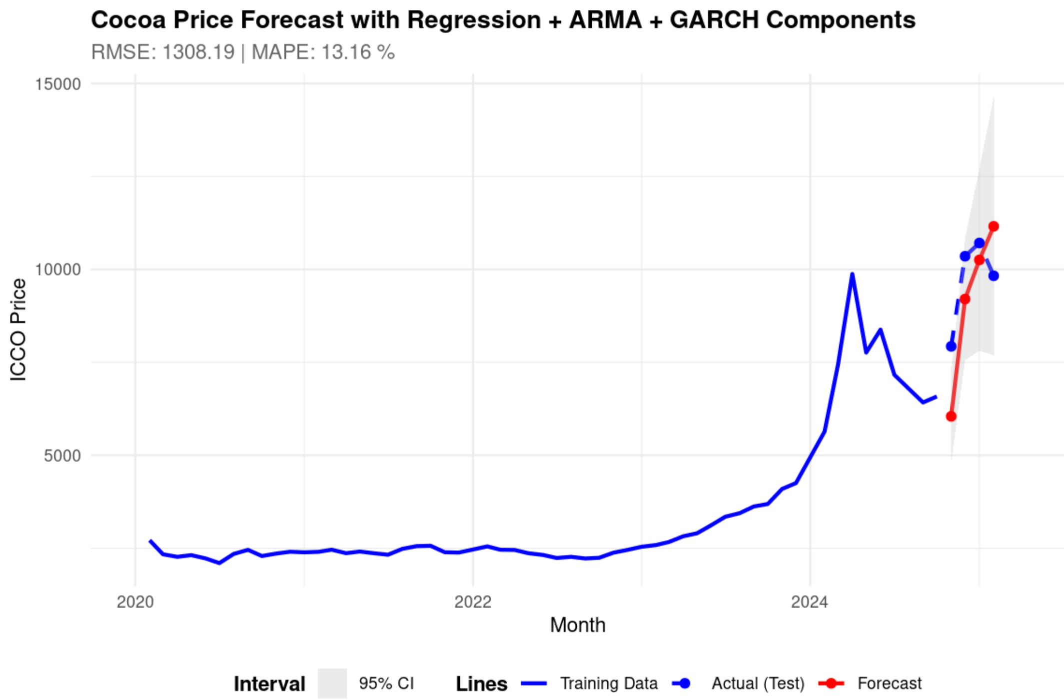

Figure 9: *Forecasted values for the ICCO prices (red) against the test set (dashed blue) for validation. The training set is also plotted in solid blue. The gray shaded area is the 95% prediction interval for forecasted values.*

5.2 Results of SARIMAX Model Forecasting

In this section, we experiment with two modeling approaches:

- Seasonal ARIMA (SARIMA) without exogenous predictors,
- Seasonal ARIMA with exogenous predictors (SARIMAX).

Note that when constructing these models, we decided to window the training data so that only observations from January 2020 to October 2024 were considered, for the same reasons as discussed in Section 5.1 (i.e., we wished to capture the larger amount of volatility present in recent observations more effectively). Below, we discuss the process through which we arrived at the final SARIMA and SARIMAX models, as well as the final prediction results we obtained.

After examining the ACF and PACF plots of the differenced series, we determined that the seasonal order of any SARIMA model would almost certainly be 12, since the ACF clearly displayed a yearly repeating pattern with maximal correlations nearing 1 when lags were separated 12 months apart (see Figure 5). Further, we observed that $d = 1$ would be an adequate choice for the differencing order since the ADF test returned a p-value of 0.01 for the differenced series.

Therefore, we decide to consider all SARIMA models with seasonal order 12 and $d = 1$ as possible candidates, provided that they satisfied $0 \leq p, q \leq 5$ and $0 \leq P, Q \leq 1$. We then iterated through all of these combinations using a backward selection strategy, eliminating models with higher AIC values and statistically insignificant coefficients. Ultimately, we found that SARIMA$(3, 1, 0)[1, 1, 0]_{12}$ had the lowest AIC value (870.4589) out of all SARIMA models that we considered. The 4-month forecast given by this SARIMA model is shown below, together with a visual comparison of this forecast to the actual values from the testing dataset.

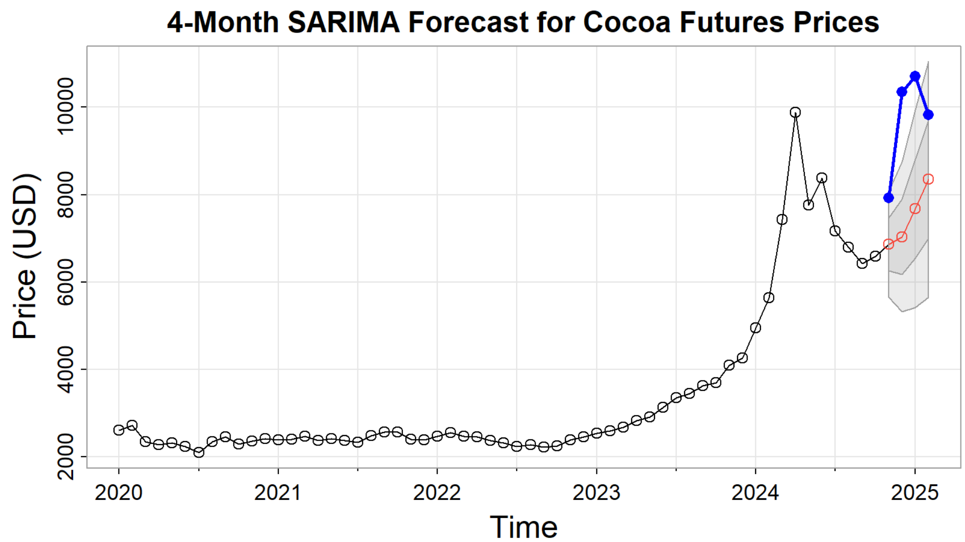

Figure 10: *4-month SARIMA forecast for the ICCO prices (red), plotted against the test set (solid blue) for validation. The training set is also plotted in black. The light gray shaded area is the 95% prediction interval for forecasted values, and the dark gray shaded area is the 80% prediction interval.*

To explore beyond the STA457 course curriculum and demonstrate creativity in our model construction process, we decided to experiment with incorporating exogenous predictors into the above SARIMA model. Since the ARIMAX/SARIMAX literature (e.g. Section 2.8 of our Literature Review) generally recommends choosing exogenous predictors whose futures values *will be known*, we elected to choose the *month of the year* as an exogenous predictor, in alignment with the method followed by Lestari & Dini [8].

The forecast of the SARIMAX model we constructed using this predictor is shown below, together with a visual comparison of this forecast to the actual values from the testing dataset.

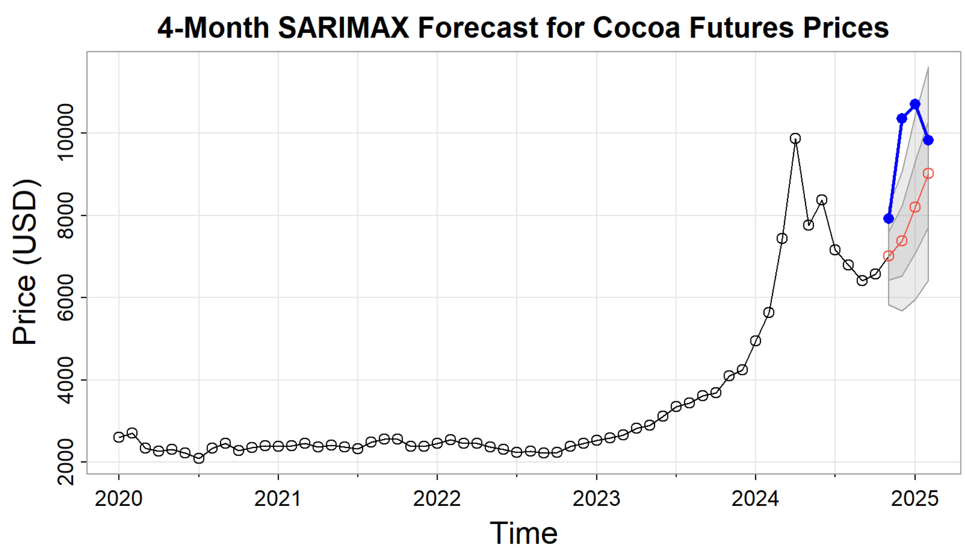

Figure 11: *Forecasted values for the ICCO prices (red) against the test set (blue) for validation. The training set is also plotted in black. The light gray shaded area is the 95% prediction interval for forecasted values, while the dark gray shaded area is the 80% prediction interval.*

We omit an extensive discussion of the residuals, as per Prof. Selvaratnam's advice in lecture (since this is not the model that we ended up deciding was the best choice). However, we will elect to mention that we observed a weak violation of normality in the QQ-plot, but the rest of the residual assumptions were all satisfied.

5.2.1 RMSE and MAPE Values for SARIMA & SARIMAX

Ultimately, we found that our SARIMA$(3,1,0)[1,1,0]_{12}$ model returned values of

$$\text{RMSE} = 1527.4780, \qquad \text{MAPE} = 15.296\%, \tag{2}$$

while our SARIMAX$(3,1,0)[1,1,0]_{12}$ model returned values of

$$\text{RMSE} = 1432.316, \qquad \text{MAPE} = 13.967\%. \tag{3}$$

This confirms that introducing the month of the year as an exogenous predictor yielded superior prediction results relative to ordinary SARIMA.

5.3 Results of Neural Network Model Forecasting

To demonstrate further creativity and extend our investigations even further beyond the STA457 curriculum, we decided to implement a recurrent neural network using the long-short-term memory (LSTM) architecture as described in the methodology (Section 3.3.2). Staying consistent with the models outlined in Sections 5.1 and 5.2, we have elected to only train on data after 2020 given the increased volatility and higher relevance of this data for futures predictions.

5.3.1 Feature Engineering

The feature engineering step constructs five additional variables based on the original `ICCO_price` series to provide richer temporal context for modeling. `ICCO_lag1` and `ICCO_lag2` represent the ICCO price values from one and two months prior, respectively, allowing the model to learn short-term dependencies. `ICCO_roll_mean` captures the three-month rolling average of ICCO prices, reflecting recent trends, while `ICCO_roll_std` provides the corresponding rolling standard deviation to quantify short-term volatility. Finally, `ICCO_diff` is the first-order difference of ICCO price, highlighting monthly price changes and helping the model detect upward or downward momentum.

5.3.2 Deseasonalization

Based on the analysis performed in Section 4.4.6, we have elected to manually deseasonalize the `TAVG` and `PRCP` data using STL (Seasonal and Trend decomposition using Loess) based on a 12 month cycle, we will refer to the deseasonalized variables as just `TAVG` and `PRCP` moving forward in this section.

5.3.3 Feature Selection

We selected 9 features for the neural network training, these are `CPI_world`, `Ghana_official`, `TAVG`, `PRCP`, `ICCO_lag1`, `ICCO_lag2`, `ICCO_roll_mean`, `ICCO_roll_std`, and `ICCO_diff`. We did not include all of the exchange rate and CPI variables to avoid over-complicating the model: `Ghana_official` was selected for consistency, as the other location-specific variables also pertain to Ghana, while `CPI_world` was included to provide a general macroeconomic context.

5.3.4 Training Setup

Before training, we applied standardization to both the input features and the target variable. This step transforms the data to have zero mean and unit variance, which is important because neural networks are sensitive to the scale of input values. By scaling the data, we ensure that each feature contributes proportionally to the learning process. Finally, the features were structured into a moving window format to prepare the data for sequential modeling with LSTMs. In this setup, a fixed-length window of past observations is used to predict the target value at the next time step. For example, with a window size of 12, each training sample consists of 12 consecutive months of the 9 selected features and the corresponding label is the ICCO price in the following month.

5.3.5 Hyperparameter Selection

To identify the optimal LSTM architecture for forecasting, we conducted a grid search over key hyperparameters: the window size (6, 12, 24 months), hidden layer size (50, 100, 150 units), network depth (1–3 stacked LSTM layers), and number of training epochs (10, 25, 50). For each configuration, we split the training set into a training subset and a validation subset using a chronological 80/20 split. We found that the best combination of hyperparameters contained a window size of 24 months, a hidden layer size of 150 units, a network depth of 1, and 25 training epochs.

5.3.6 Results

Below, we discuss the overall performance of the LSTM model, let us first consider the performance metrics:

Set	MSE	MAE	RMSE	MAPE (%)
Training	900,469.48	482.77	948.93	7.60
Test	441,268.49	611.16	664.28	6.54

Table 2: *Performance metrics for training and test sets.*

The performance metrics indicate that the model performs reasonably well on both the training and test sets, with some expected variance. Notably, the MSE values appear large at 900,469 on the training set and 441,268 on the test set, but this is expected given that ICCO prices are in the thousands, so squared errors naturally become large in magnitude. Notably, the RMSE and MAPE of the test set was lower than that of the training set, suggesting that the model generalized well. Consider the forecasted prices below:

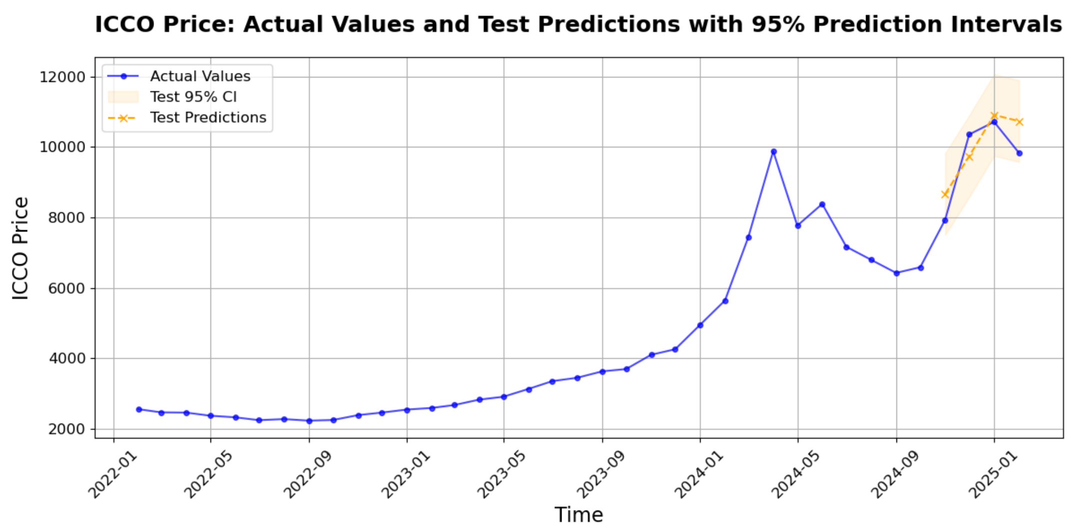

Figure 12: *Forecasted values for the ICCO prices (orange) on the test set against the actual prices (blue) for validation. The light orange shaded area is the 95% prediction interval for forecasted values.*

As seen in the figure, the predicted values for the test set were quite accurate with respect to the large amount of volatility found in the data that were not present in the training set. In particular, the 95% prediction interval covers the actual values for all points in the testing dataset, again confirming that the model generalized well and did not overfit. Note that we are not performing residual diagnostics because unlike classical linear time series models (e.g., ARIMA), neural networks like LSTMs do not rely on strict assumptions about the residuals being white noise or normally distributed. As per Section 2.6 of our literature review [6], neural networks can model any form of unknown relationships with minimum assumptions, and are considered universal approximators as such.

6 Discussion and Conclusion

6.1 Interpretation of Results

Below, we display a table summarizing each model's performance on the test set by evaluating the RMSE and MAPE.

Model	RMSE	MAPE (%)
`ARCH/GARCH`	1308.19	13.16
`SARIMA`$(3,1,0)[1,1,0]_{12}$	1527.48	15.30
`SARIMAX`$(3,1,0)[1,1,0]_{12}$	1432.32	13.97
`RNN (LSTM)`	664.28	6.54

Table 3: *RMSE and MAPE values of each of the models discussed in this report.*

We conclude that the SARIMA model performed the worst, with the ARCH/GARCH model slightly outperforming the SARIMAX model for both RMSE and MAPE. Thus for the **classic statistical time series models only**, the ARCH/GARCH model *was the best-performing model.* These results make sense because the SARIMA model fails to leverage the exogenous variables that were provided, and we can see how even with a very similar model, SARIMAX performed moderately better. However, both the SARIMA (Figure 10) and SARIMAX (Figure 11) models overestimated the forecasted cocoa futures prices on the test set. Without assessing the exogenous variables, it is possible that the SARIMA model could not capture the complexities of the cocoa futures prices, which is likely why we observed a weak violation of the normality assumption when performing residual diagnostics. On the other hand, the ARCH/GARCH hybrid model combined the exogenous variables through a base regression model and fitted the residuals using ARMA and GARCH; this produced better forecasting predictions (Figure 9) with the 95% prediction intervals capturing the true test set cocoa futures prices. Additionally, residual diagnostics of this model indicated that the residuals resemble white noise, satisfying the model assumptions and demonstrating a well-fitted model.

Finally, the recurrent neural network model using LSTMs performed *the best overall out of all the models we considered*, with a significantly lower RMSE and MAPE than the other 3 aforementioned models. The prediction intervals (Figure 12) reflected the accuracy: the values for both of these error metrics reflect the model's strong ability to generalize to unseen data and the expressive power of neural networks. Neural networks can achieve this performance through incorporating non-linearity and a large number of parameters. Our model trained $96,151$ parameters with just 1 LSTM layer. Nevertheless, we did observe some of the same challenges noted in Hewamalage's study on RNNs [6], such as being more computationally intensive and harder to train.

6.2 Limitations and Challenges of our Approach

Firstly, our analysis focused primarily on recent trends in the cocoa futures market, which has become much more volatile over the past 2 years. As such, it is possible that if the market returns to pre-2023 levels, our models could over-predict the level of volatility in the market, thus giving less accurate predictions. Secondly, the models that were evaluated considered a subset of possible exogenous variables, a more robust approach could be taken to identify other important covariates such as related commodity prices. Thirdly, we converted all data into monthly averages, this brings a tradeoff between how much volatility our data captures versus the amount of noise. Lastly, despite performing the best, the neural network model is not very interpretable (96,151 parameters) compared to the classical time-series models, creating challenges for understanding the actual market drivers behind cocoa futures prices.

6.3 Next Steps for Improvement

To address the first limitation, we can follow the discussion in Section 2.5 of our literature review [5] and attempt to determine whether concept drift, which consists of unforeseen changes in the underlying factors driving cocoa futures prices, is present in our neural network model. We can incorporate more flexibility in the model parameters by using an approach such as the Self-Exciting Threshold AutoRegressive (SETAR) model, which models the series as multiple autoregressive parts based on threshold conditions.

For the second limitation, we may note that we only included relevant data from the West Africa region (with a focus on Ghana) in our analysis; this region only accounts for around 70% of global cocoa production. Thus, a next step could be to account for the remaining 30% of cocoa production by introducing climate data or economic metrics pertaining to other regions (e.g. Latin America, or Southeast Asia). Additionally, the prices of other commodities such as sugar and palm oil could be considered, as well as metrics for other societal/environmental factors that could impact cocoa production (e.g. disasters or disease outbreaks). Given all these possible variables, ANOVA should be used for feature selection across all models. Incorporating these changes could enhance the scope of our analysis, leading to a more refined prediction model.

For the third limitation, it is worth considering the daily data that was originally provided as an option for analysis. This approach increases the sample size drastically, improving model training for the LSTM model and allows us to identify day-of-week effects or within-month patterns that might be predictive of the cocoa futures price.

Lastly, we should consider incorporating other models in our analysis, especially hybrid models that utilize neural networks. For example, we can use SARIMA to capture the overall trend of the data, and then train a recurrent neural network on the residuals. This approach leverages the interpretability of SARIMA and expressive power of neural networks. Other methods such as assembling multiple models into a single, overarching model can also be used to achieve a better forecast.

7 Appendix

7.1 Link to Code Used For EDA & Model Construction

https://drive.google.com/drive/folders/1GNQIDI7mt-AGGw3uzueiHA1OyghmPb3y?usp=sharing

7.2 Additional Plots & Figures

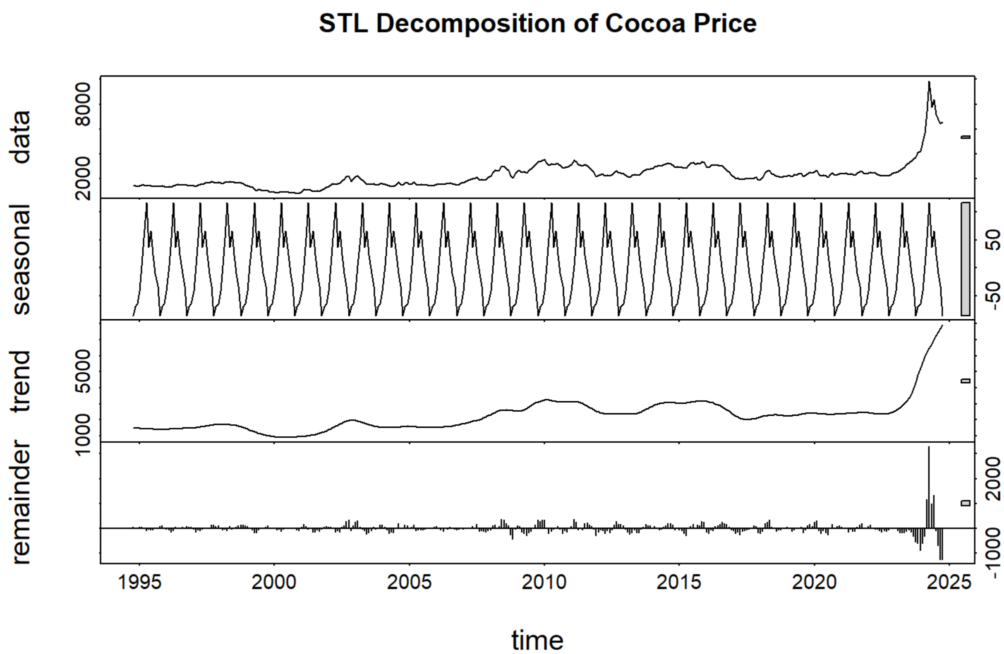

Figure 13: *STL Decomposition of Cocoa Futures Prices*

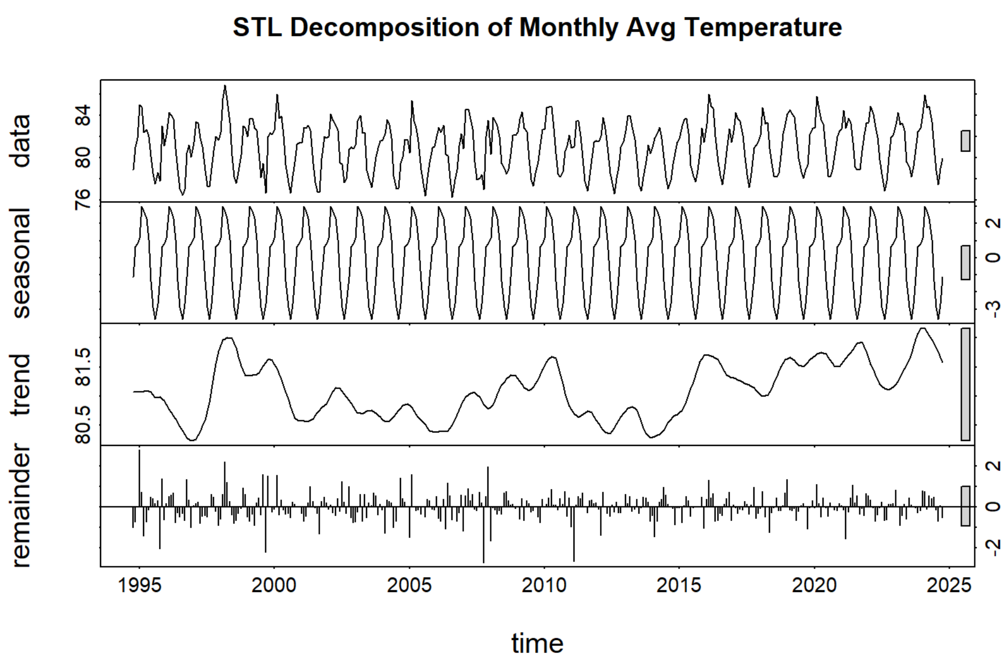

Figure 14: *STL Decomposition of Average Ghanaian Monthly Temperature*

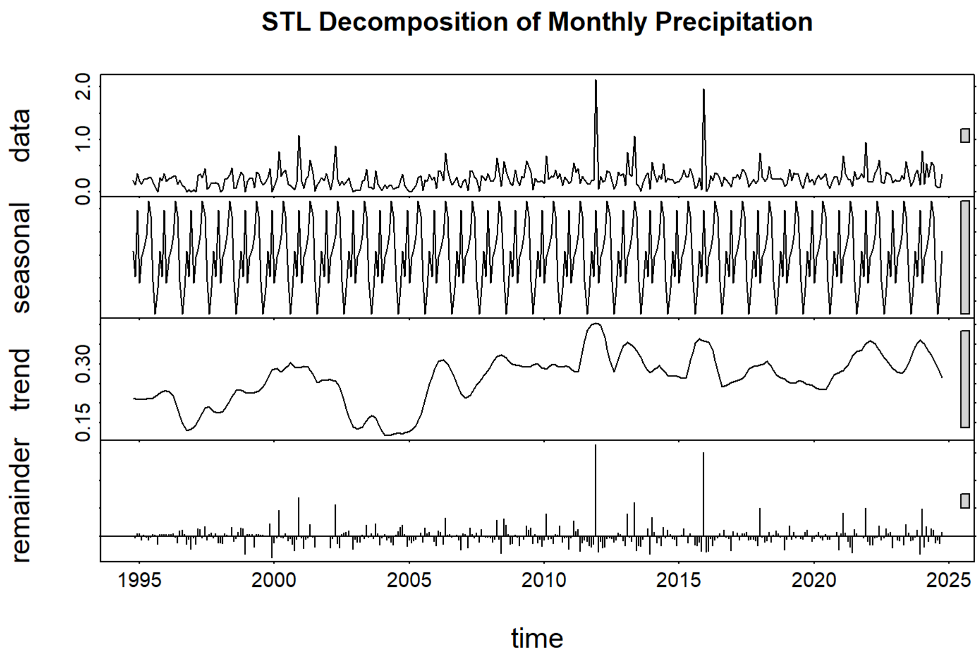

Figure 15: *STL Decomposition of Cocoa Futures Prices*

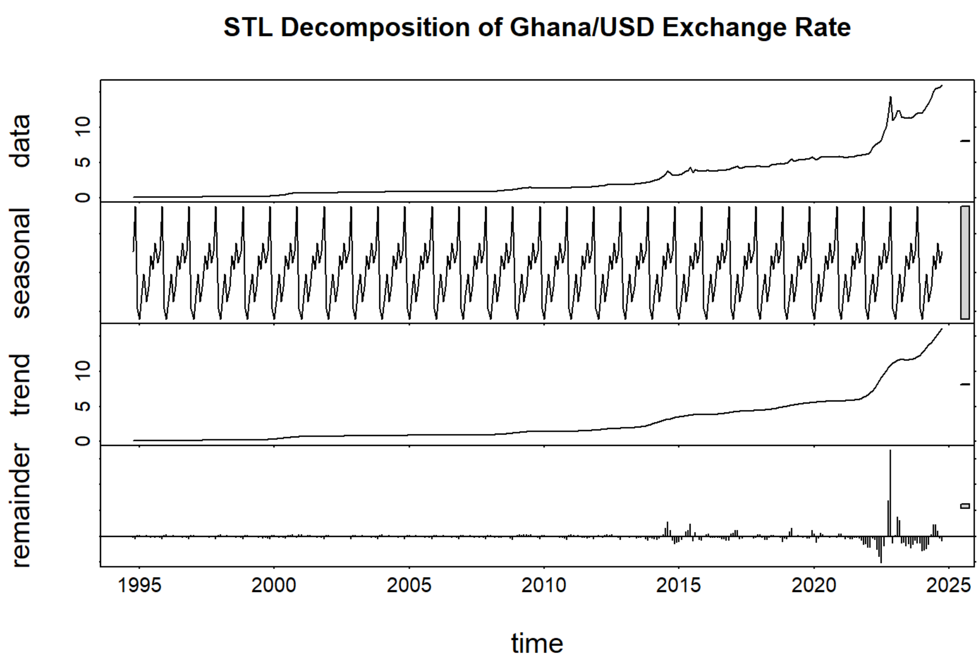

Figure 16: *STL Decomposition of Ghanaian Cedi/USD Exchange Rate*

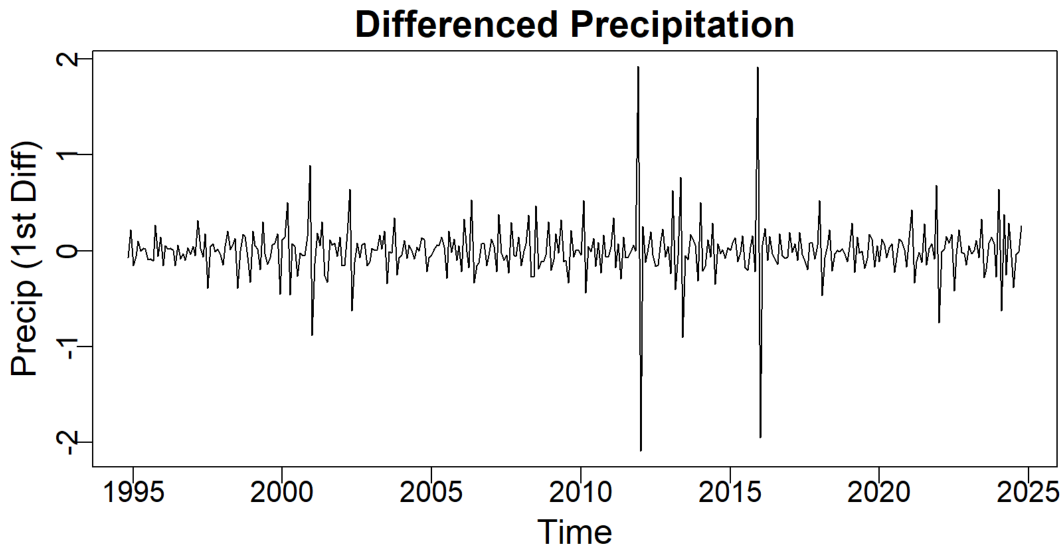

Figure 17: *Differenced Precipitation Time Series*

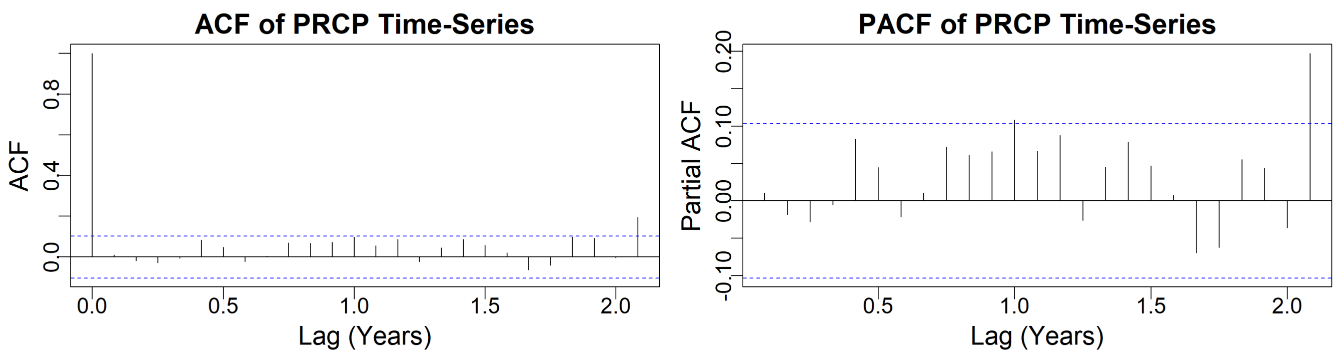

Figure 18: *ACF/PACF Plots of PRCP Time Series*

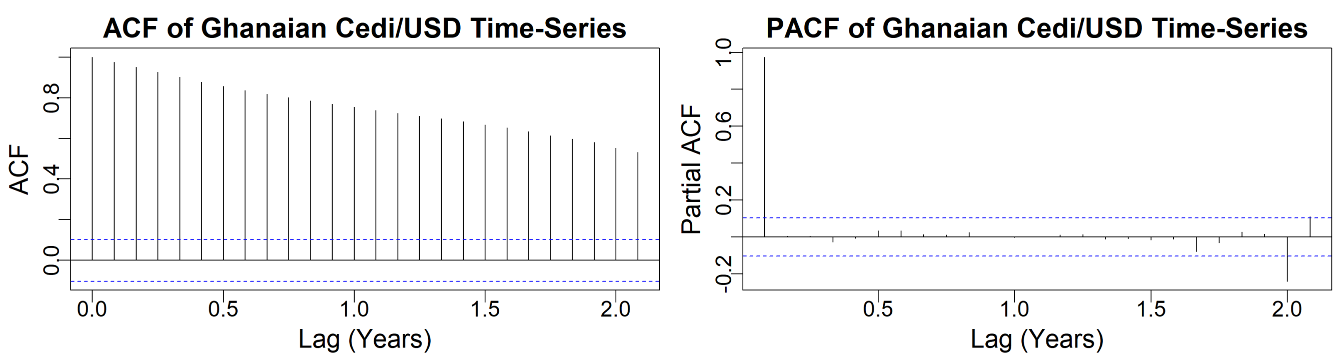

Figure 19: *ACF/PACF Plots of Ghanaian Cedi/USD Exchange Rate Time Series*

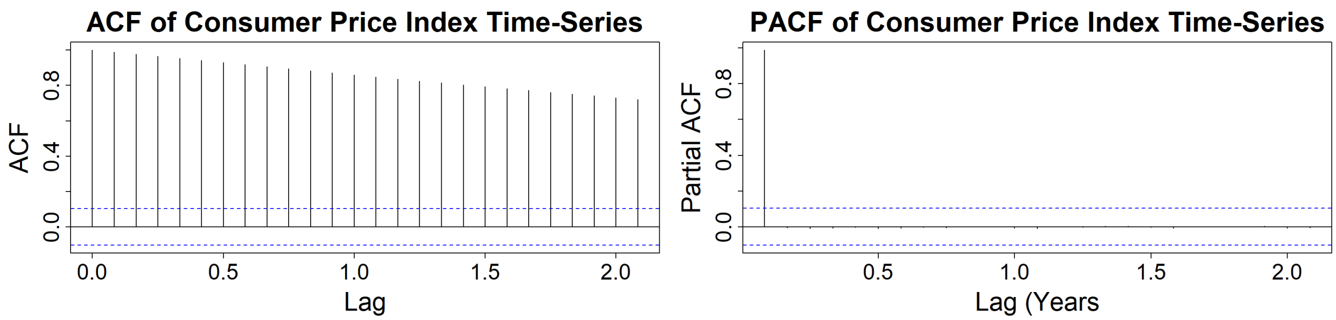

Figure 20: *ACF/PACF Plots of Global CPI Time Series*

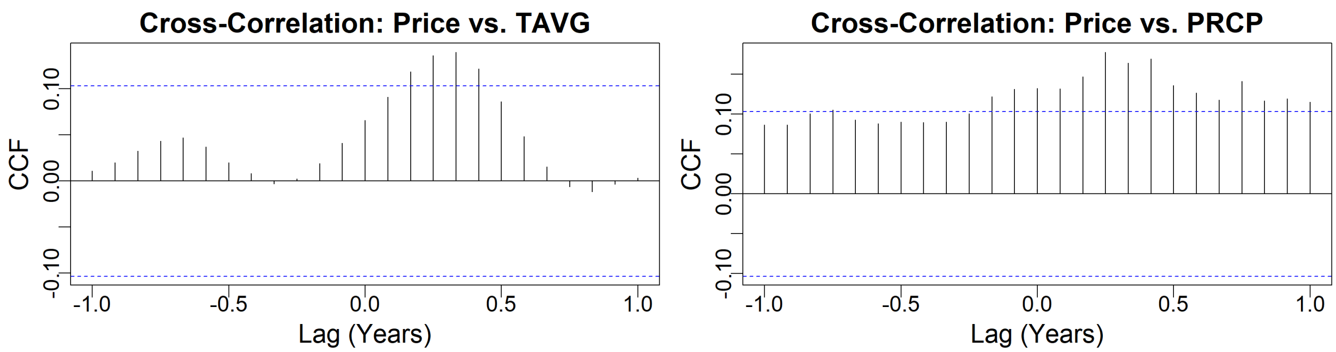

Figure 21: *CCF Plots of Price vs. TAVG and Price vs. PRCP*

8 References

[1] J. O. Ajetomobi and A. O. Olaleye, "Auto-regressive integrated moving average (arima) modeling of cocoa production in nigeria: 1900-2025," *Journal of Crop Improvement*, vol. 33, no. 4, pp. 445–455, 2019. DOI: `10.1080/15427528.2019.1610534`.

[2] L. Aklimawati and T. Wahyudi, "Estimating the volatility of cocoa price return with arch and garch models," *Pelita Perkebunan (a Coffee and Cocoa Research Journal)*, vol. 29, no. 2, 2013. DOI: `https://doi.org/10.22302/iccri.jur.pelitaperkebunan.v29i2.61`. [Online]. Available: `https://www.ccrjournal.com/index.php/ccrj/article/view/61`.

[3] M. P. J. Tabe-Ojong, O. T. A. Guedegbe, and J. Glauber, "Soaring cocoa prices: Diverse impacts and implications for key west african producers," *IFPRI Blog: Issue Post Markets, Trade, and Institutions*, 2024. [Online]. Available: `https://www.ifpri.org/blog/soaring-cocoa-prices-diverse-impacts-and-implications-key-west-african-producers`.

[4] C. I. Anderson and W. A. Gough, "Assessment of missing data treatment on monthly air temperature averages in climatological analyses," *International Journal of Climatology*, vol. 38, no. 3, pp. 1455–1465, 2018. DOI: `10.1002/joc.5801`.

[5] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," *Machine Learning*, vol. 23, pp. 69–101, 1996. DOI: `https://doi.org/10.1007/BF00116900`. [Online]. Available: `https://link.springer.com/article/10.1007/BF00116900`.

[6] H. Hewamalage, C. Bergmeir, and K. Bandara, "Recurrent neural networks for time series forecasting: Current status and future directions," *International Journal of Forecasting*, vol. 37, no. 1, pp. 388–427, Jan. 2021, ISSN: 0169-2070. DOI: `10.1016/j.ijforecast.2020.06.008`. [Online]. Available: `http://dx.doi.org/10.1016/j.ijforecast.2020.06.008`.

[7] A. D. Alori and A. A. Kutu, "Export function of cocoa production, exchange rate volatility and prices in nigeria," *Journal of Economics and Behavioral Studies*, vol. 11, no. 2(J), pp. 1–14, May 2019. DOI: `10.22610/jebs.v11i2(J).2813`. [Online]. Available: `https://ojs.amhinternational.com/index.php/jebs/article/view/2813`.

[8] D. Lestari and S. Dini, "Using ARIMAX with monthly dummies to forecast spice prices," *EKSAKTA Journal of Sciences and Data Analysis*, vol. 10, no. 2, pp. 45–58, 2024. DOI: `10.20885/EKSAKTA.vol5.iss1.art5`. [Online]. Available: `https://journal.uii.ac.id/Eksakta/article/view/28992`.

[9] DataBank, *Global CPI & Exchange Rate Data*, https://databank.worldbank.org/source/global-economic-monitor-(gem), 2025.