Forecasting Cocoa Prices with Daily SARIMAX Models: A Time Series Analysis of Weather, Currency, and COVID-19 Impacts

Evaluating Exogenous Variables in Commodity Price Modeling Using High-Frequency Data from Ghana and Côte d'Ivoire

Cristina Burca Quynh Do Karim Hijazi Kristin Mai

April 4, 2025

Table of contents

1 Introduction

Cocoa prices are influenced by a variety of global and local factors, including environmental volatility, underinvestment in agricultural operations, and economic uncertainty. These concerns contribute to irregular fluctuations in production and global supply, contributing to recurring cocoa shortages and price shocks (J.P. Morgan 2024).

Time series forecasts allow for the monitoring of prices and inform governments, farmers, and industries to manage risks and plan accordingly (Karishma Harrykissoon 2023). Given that many of these factors are recursive or follow seasonal patterns, it is important to build a model that can capture both autoregressive behaviour and the effects of external influences.

This study implements a Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors (SARIMAX) model to forecast daily cocoa prices, incorporating significant external factors of weather patterns, currency fluctuations, and global factors. The number of external factors incorporated into the model was restricted due to a lack of quantifiable data for other majors affecting cocoa prices such as the prevalence of pests and diseases on cocoa trees, the stability of politics and the supply chain, yields and production, as well as the global demand for cocoa from dependent industries.

The report is motivated by the need to develop predictions of future cocoa prices to help stakeholders in the cocoa industry, including farmers, traders, and policymakers, better manage risks and make informed decisions when faced with sudden price fluctuations. Through developing a model that accurately produces forecasts of cocoa prices, the study aims to provide information that could contribute to improved supply chain management through production and distribution adjustments, increase profits for producers through the ability to plan production to align with predicted peaks and slumps, and stabilise market conditions by reducing uncertainty.

2 Literature Review

There are many different time series models that exist which were considered when approaching the task of modelling future cocoa prices. The Exponential Smoothing model is a classic time series model which produces forecasts of data from the weighted averages of past data. "Exponential" refers to the weights associated with each observation decay at an exponential rate with less recent time (Hyndman and Athanasopoulos 2018). The Autoregressive Integrated Moving Average Model (ARIMA) is another classic time series model and is one of the most widely used time series models alongside the Exponential Smoothing model (Hyndman and Athanasopoulos 2018). Similar to the Exponential Smoothing model, forecasts are based on past observations. However, it forecasts future values using a linear combination of past values of the variable, focusing on the relationship between past observations and present values while also incorporating error terms (Hyndman and Athanasopoulos 2018). The Seasonal ARIMA (SARIMA) model is a subclass of the ARIMA model which takes seasonal fluctuations into account and has seen to be successful for commodity price modelling (Giri and Giri 2024; Theerthagiri and Ruby 2023).

Given studies that demonstrate that ARIMA models outperform Exponential smoothing in forecasting stock prices (Yogesh Funde 2023) and wood prices (Tetsuya Michinaka and Yamamoto 2016), and the relevance of considering seasonal fluctuations when modelling commodity prices, an approach using a variation of the SARIMA model was chosen to model future cocoa bean prices over an Exponential Smoothing model. However, the SARIMA model is limited in that it only considers the past observations of the data and does not incorporate external factors in the model. The Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors (SARIMAX) is an extension of the SARIMA model with the added ability to consider exogenous factors which may have an effect on the model– which overcomes this limitation– making it a suitable choice for our study. The consideration of exogenous factors in our forecasting model is supported by existing research which indicates improved accuracy in the forecasting of agricultural commodity prices when external factors such as weather (Haoyang Wu 2017).

3 Methodology

To model and forecast daily cocoa prices, we implemented a SARIMAX model. While traditional SARIMA models are effective in capturing internal time series dynamics and seasonal trends, they do not sufficiently account for external shocks that may greatly affect the variable of interest. SARIMAX addresses this by incorporating external variables, making it highly effective for real-world applications where internal and external factors shape market behaviour (Karishma Harrykissoon 2023).

In this case, cocoa prices are driven by weather, exchange rates, and the macroeconomic impact of the COVID-19 pandemic. The SARIMAX model explicitly models these external factors alongside the internal time series structure, resulting in a more flexible and interpretable forecasting process.

To improve forecast accuracy, daily temperature data is included from the two largest cocoa-producing countries: Ghana and Côte d'Ivoire— together responsible for approximately 76% of global cocoa production. Climate variability in these regions directly affects cocoa yields and export volumes, as cocoa production is highly sensitive to temperature stress and heat extremes.

Exchange rates are known to influence global commodity prices through trade competitiveness and producer margins, especially in developing economies where volatility can severely disrupt market participation (Adeoye, Babajide, and Folarin 2019). Thus, currency exchange rates for the Ghanaian Cedi (GHS) and West African CFA franc (XOF) were included.

Furthermore, the COVID-19 pandemic introduced a structural break in global markets, causing disruptions in labor, logistics, and international trade. During this period, cocoa was not classified as an essential good, resulting in supply chain delays, export restrictions, and demand declines, which led to a decline in prices. As economies began to recover, prices trended upward. To capture these structural shifts, indicators were introduced to capture the distinct pre and post COVID-19 trends and correctly account for them in the model.

3.1 Model Preparation

In preparation for the model, the date range was set to ensure consistent alignment across all variables, and to reserve recent cocoa prices for the comparison of the forecasted values with the most recent real market prices for a 120-day horizon. Data was restricted to the overlapping date range shared by cocoa prices, weather, and currency data to ensure regressors are available for the entire modelling period. Missing values were removed, and all variables were aligned and merged by date into one dataset with one observation per day.

A univariate time series object `(price_ts)` was created from daily cocoa prices, with a corresponding exogenous regressor matrix `xreg_vars` , including the following variables:

- **Daily Weather:** `TAVG`, `IC_TAVG` for Ghana and Cote d'Ivoire respectively,
- **Exchange Rates:** `GHS` for Ghanaian Cedi, `XOF` for West African CFA franc,
- **Structural Indicators:**
  - `Covid_Lag`: A dummy variable marking the prominent period of the COVID-19 pandemic, from March 2020 to December 2022,
  - `Post_2023`: A step indicator to capture level shifts in prices beginning January 1, 2023,
  - `Post_2023_Trend`: A continuous linear trend starting from Jan 1, 2023, to account for gradual post-COVID dynamics.

Initial exploration of cocoa prices in Figure 1 revealed clear non-stationarity, confirmed by the upward trend in prices, and the slowly decaying ACF. To address this, first-order differencing was applied to stabilize the mean, shown in Figure 2. ACF/PACF plots for the difference series were used to identify relevant AR and MA terms.

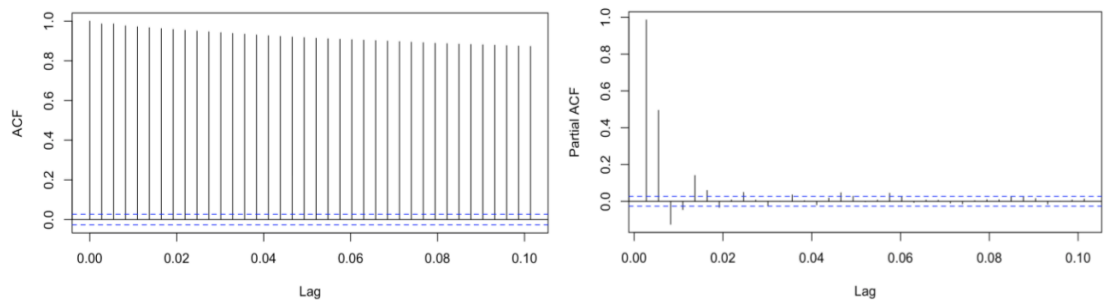

Figure 1: ACF and PACF of cocoa prices time series prices_ts, showing non-stationarity.

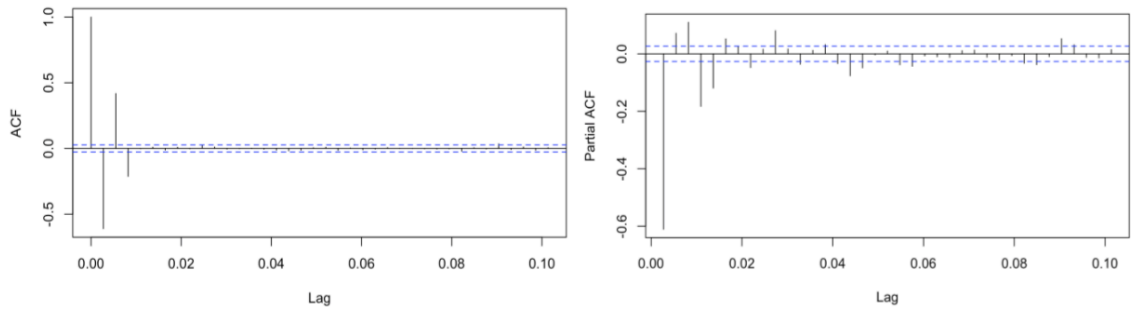

Figure 2: ACF and PACF of differenced cocoa prices time series prices_ts, showing stationarity.

To account for seasonal structure, a weekly seasonal component is specified to capture weekly cycles. Due to software and convergence limitations, a fully annual cycle (`period = 365`) is not feasible, as it would introduce an unmanageable high number of parameters, and is likely to overfit. A weekly period offers a realistic and computationally feasible approximation for seasonality in daily economic data. The following SARIMAX models were fitted with varying autoregressive and moving average terms, increasing in complexity:

Table 1: Comparison of SARIMAX Models Using AIC and BIC

Model	AIC	BIC
SARIMAX(2,1,2)	9550.478	9614.353
SARIMAX(3,1,3)	9500.398	9573.397
SARIMAX(4,1,4)	9498.467	9580.591
SARIMAX(5,1,3)	9473.034	9555.158

The final model was chosen based on lowest AIC and BIC, indicating the best trade-off between model fit and complexity. The selected model structure is:

$$\text{SARIMAX}(5,1,3) \times (1,0,1)_7,$$

- `p=5`: 5 AR lags to capture long-term price movements,
- `d=1`: first differencing to address non-stationarity,
- `q=3`: 3 MA terms to account for short-term shocks,
- `P=1, Q=1`: seasonal AR and MA lags to capture weekly cyclical effects,
- `D=0`: seasonal differencing not required,
- `period = 7`: weekly seasonality to reflect common volatility.

This structure allows the model to flexibly capture both short-term and long-run trends, while including external drivers and accounting for cyclical behaviour in the data.

4 Data

The following datasets were gathered and processed between February 1, 1999, and November 28, 2024:

- **Cocoa Prices:**
  - Source: International Cocoa Organization (ICCO) (2025)
  - Variable: Prices (USD/tonne)

Figure 3: Daily international cocoa prices from the International Cocoa Organization (ICCO) from 1999 and 2024. A dramatic price surge is observed in late 2023 into 2024, following a period of relative stability.

Figure 3 shows the global cocoa price in USD/tonne. The plot exhibits a plateaued upward trend with a dramatic spike in late 2024, indicating recent market shocks likely due to production deficits from COVID-19. This justifies the structure of the COVID-19 indicators in the model.

- **Weather Data:**
  - Source: National Centers for Environmental Information (NCEI) (2025)
  - Variable: `TAVG` daily mean temperature in degrees F°
  - Countries: Ghana, Côte d'Ivoire
  - Perception data was excluded due to its 43% missingness. Data were aggregated from multiple weather stations per country using daily means

Figure 4 show daily average temperatures from 1999 to 2024, plotted through smoothing. Strong expected seasonality of temperatures are depicted, with heavier oscillations in recent years. These patterns validate the use of seasonal components in the model.

- **Currency Exchange Rates:**
  - Source: USD/XOF Exchange Rates (Investing.com 2025), Interbank FX Rates (Ghana 2025)
  - Variable: `GHS` (Ghanaian Cedi), `XOF` (West African CFA franc)
  - Mid-rate values were interpolated for daily frequency. Pre-July 2007 Cedi values were scaled down to account for redenomination (Dzokoto, Young, and Mensah 2010)

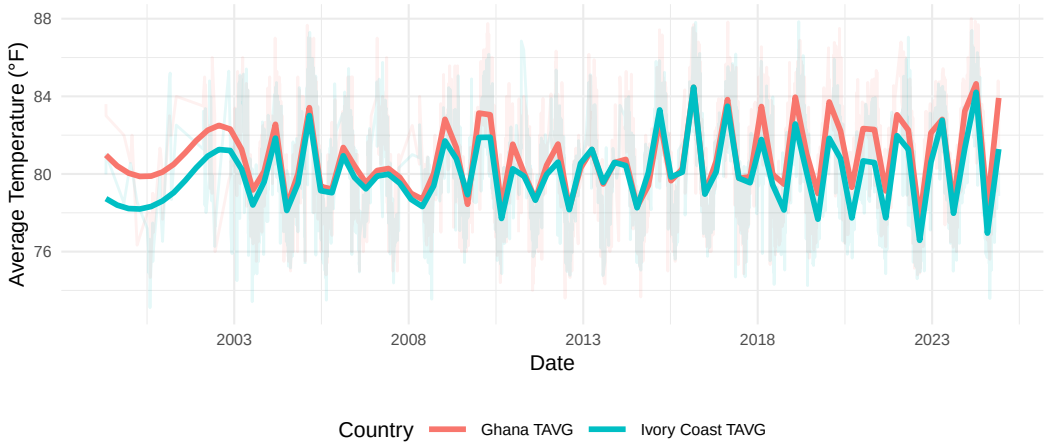

Figure 4: Daily average temperatures (°F) from 1999 to 2024 in Ghana and Côte d'Ivoire. Strong annual seasonality is visible in both countries, and no major structural climate shocks during the period.

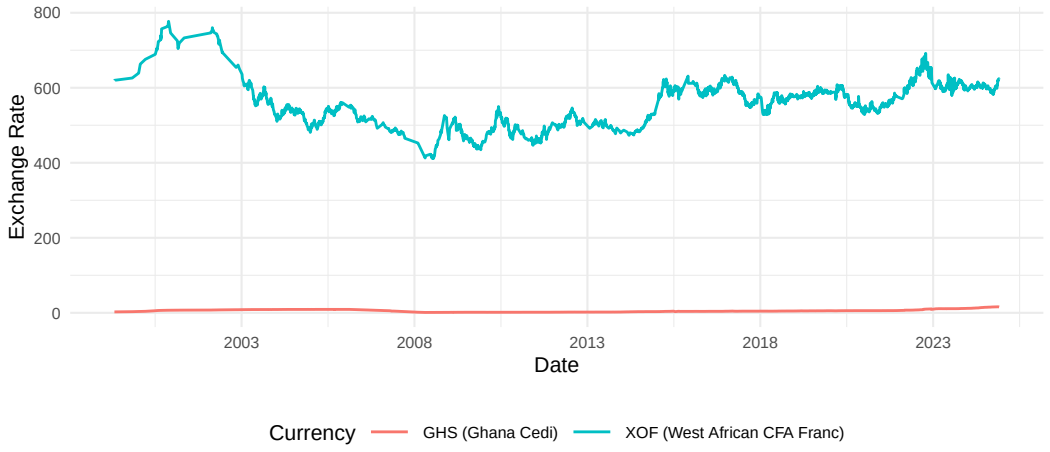

Figure 5: Exchange rates of Ghanaian Cedi (GHS) and West African CFA Franc (XOF) from 1999 to 2024. The Cedi experienced redenomination in 2007, followed by gradual devaluation.

Figure 5 shows daily mid-rate exchange value of the Ghanaian Cedi (GHS) from 1999 to 2024. A redomination occurred in July 2007, causing a visible reset in the exchange rate. After stabilization, the GHS experienced a steady depreciation, with an accelerated decline starting around 2022—likely influenced by post-pandemic economic pressures.

- **Structural Indicators:** `Covid_Lag`, `Post_2023`, `Post_2023_Trend`, as described above.

For the goal of a short term 120-day forecast, frequent data is prioritized to assure accuracy in daily predictions. While additional data sources—such as production volumes, export statistics, and input costs—could offer valuable context, these were ultimately excluded from the model due to their low temporal resolution and limited recent availability – this is further discussed in Section 6.1. Similarly, although initial ideas explored incorporating temperature data from multiple major cocoa-producing countries, exploratory analysis revealed high multicollinearity. As shown in the correlation matrix Figure 6, average daily temperatures across Ghana, Côte d'Ivoire, Nigeria, and Cameroon were strongly correlated, with Pearson correlation coefficients above 0.70. Including all of them in the same model introduced redundancy leading to overfitting, and thus only the two most influential producers—Ghana and Côte d'Ivoire— were considered for their climate influence.

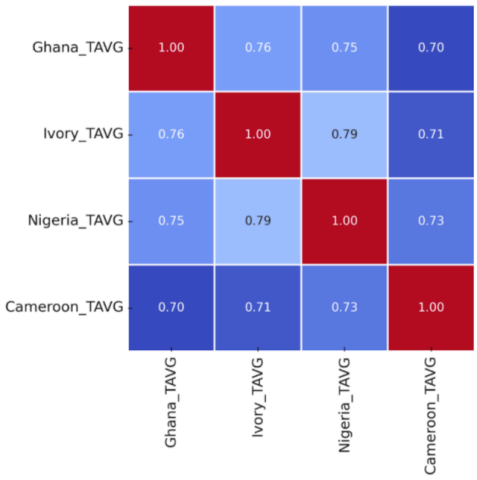

Figure 6: Correlation Matrix of Average Daily Temperatures Across Cocoa-Producing Countries (1999–2024).

5 Forecasting & Results

5.1 Model Training & Validation

The SARIMAX model was trained using daily data from February 1, 1999, to November 28, 2024. After merging cocoa price data with aligned weather and exchange rate variables, the dataset included over 8,000 complete daily observations. Several SARIMAX specifications were tested to determine the optimal structure, all incorporating a weekly seasonal component to capture short-run cycles. Models were evaluated using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) to balance model fit and complexity. As shown in Table 1, SARIMAX(5,1,3) outperformed alternative specifications with the lowest AIC (9473.03) and BIC (9555.16), and was therefore selected for forecasting. To validate model performance, the final 120 days of observed data were withheld from training and used as a holdout test set. This approach mimicked real-world forecasting conditions and allowed for a robust comparison between predicted and actual prices.

5.2 Performance & Evaluation

To evaluate the performance of the final SARIMAX$(5,1,3)(1,0,1)_7$ model, we used three standard error metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) displayed in Table 2. The model achieved an RMSE of 1524.13 and a MAE of 1324.79, indicating that the average prediction error in USD/tonne remained within a reasonably tight margin given the highly volatile price series. The MAPE of 13.31% suggests moderate forecasting accuracy relative to actual values. While performance was acceptable overall, visual inspection of the forecast reveals that the model failed to anticipate extreme deviations, highlighting limitations in capturing recent shocks and nonlinear behavior.

Table 2: Comparison of SARIMAX Models Using AIC and BIC

Table 2: Performance statistics of SARIMAX(5,1,3)(1,0,1)[7] model

Metirc	Value
RMSE	1524.13
MAE	1324.79
MAPE (%)	13.31

5.3 Forecasted Values & Observed Patterns

The SARIMAX model was used to generate a 120-day out-of-sample forecast of daily cocoa prices. As shown in Figure 7, the model predicts a modest but steady increase in cocoa prices, aligning with recent structural shifts in the market. However, actual prices displayed significantly more volatility than expected, rising rapidly through December before sharply declining in early January. This divergence highlights the challenge of forecasting in highly volatile commodity markets, particularly in periods of structural change. Nevertheless, the model successfully captured the general directional trend of prices and signaled a potential post-COVID upward shift—evident in its anticipation of a rise in November, which materialized in the observed data.

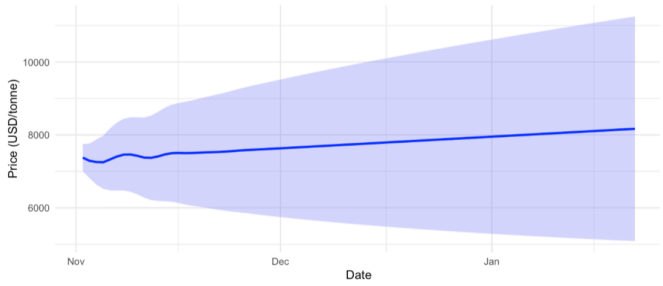

Figure 7: Forecast of Cocoa Prices (120-Day Horizon) from November 2024 to February 2025 with 95% confidence intervals.

5.4 Graphical Representations of Forecast vs. Observed

Figure 8 presents the side-by-side forecasted and actual cocoa prices over the 120-day horizon. The solid blue line shows the predicted values, while the red dashed line overlays actual market prices. The shaded region represents the 95% prediction interval. While the model underestimated the magnitude of price swings, it captured the timing of the late-November spike, providing valuable foresight. Figure 8 displays the same forecast window without the actual values to isolate the projected trend and uncertainty range.

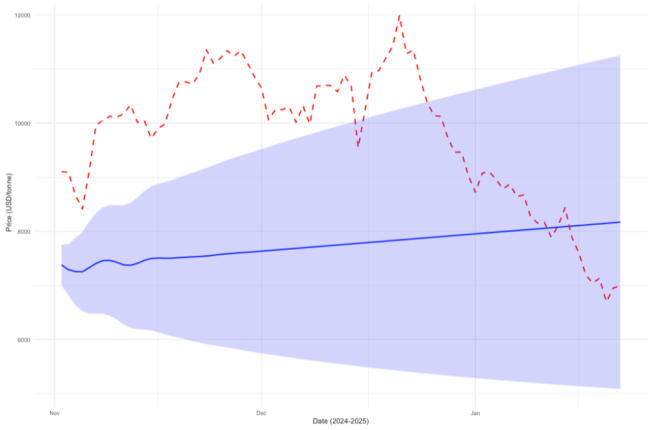

Figure 8: Forecast of Cocoa Prices of 120-Day Horizon (blue solid line) with observed market values (red-dashed line) from November 2024 to February 2025. 95% confidence intervals (shaded blue region) included.

6 Discussion

The SARIMAX(5,1,3) model developed and used in this study is seen to be effective in forecasting daily cocoa prices by capturing trends and dynamics.

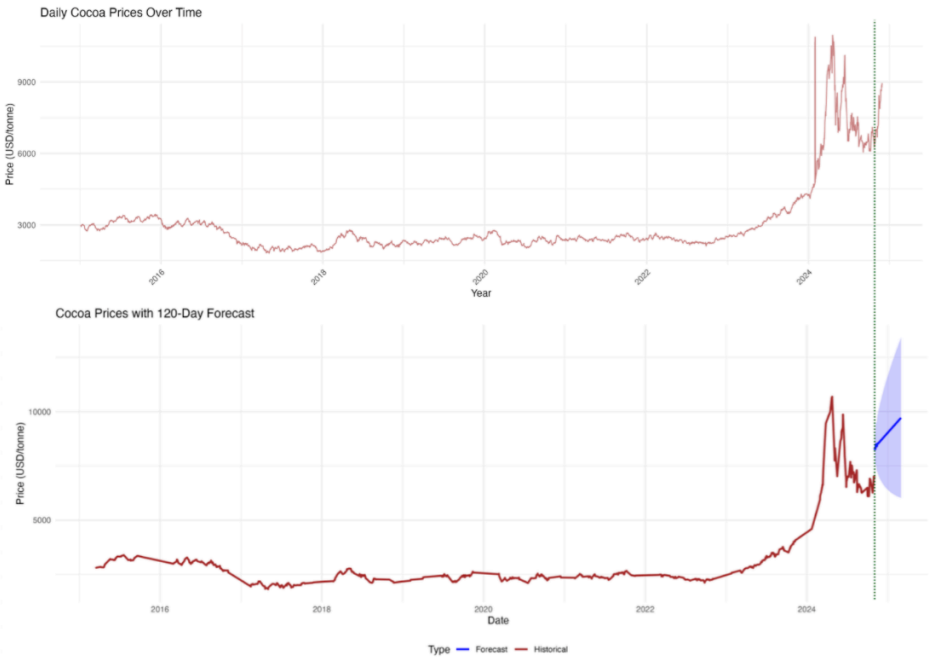

Figure 9: Forecast and observed market values of cocoa prices long-term, with forecast compared to observed values from 2024-11-08 to 2025-02-27.

The forecast figure (Figure 9) showcases the full historical trend of cocoa prices alongside the 120-day forecast window. Notably, the model correctly anticipates the sharp upward price movement following the training period. While the forecasted values slightly overestimate the spike's magnitude, the model successfully captures both the timing and direction of the surge, an impressive outcome given the inherent volatility of commodity markets and structural changes in the cocoa sector after 2023. This demonstrates the model's ability to generalize well beyond the training data and its robustness in identifying key directional shifts

during turbulent market conditions. The results align with findings in the literature that support the effectiveness of variations of ARIMA models in forecasting prices (Giri and Giri 2024; Theerthagiri and Ruby 2023).

The integration of exogenous variables such as temperature, exchange rates, and the dummy variables `Covid_Lag`,`Post_2023` and `Post_2023_Trend` for the COVID-19 pandemic and post-2023 shifts, strengthened the model's explanatory power for structural changes as it was able to reflect key external factors affecting cocoa prices. Through the model's consideration of key external shocks and features, the SARIMAX model was able to produce a more nuanced analysis over simpler models such as ARIMA and ETS which solely consider past observations, reflecting the findings of previous research in agricultural price forecasting (Haoyang Wu 2017).

6.1 Limitations & Next Steps

This study aimed to forecast cocoa prices using high-frequency daily data, but several limitations affected model development and scope. A key issue was multicollinearity in the weather data. Daily temperatures from multiple cocoa-producing countries were highly correlated, leading to unstable estimates when included together. To reduce this, only Ghana and Côte d'Ivoire were used in the model.

The model's reliance on historical data restricts its capacity to predict extreme volatility triggered by unforeseen events, like supply chain disruptions, climatic disasters and worker strikes. Importantly, it excludes qualitative factors such as pest/disease outbreaks, political instability, and speculative market dynamics due to the lack of quantifiable data, potentially limiting its accuracy during socio-economic or agricultural crises. Other useful variables, such as production, export volumes, demand, and input costs were also excluded due to their low frequency (monthly or yearly), and limited coverage beyond 2022. Including them would have reduced the forecast window and made it difficult to align with daily cocoa prices. Future work could focus on integrating low-frequency economic data with daily time series to capture broader but significant structural factors.

Data gaps further constrain performance: precipitation data (PRCP) with a missing rate of up to 43% had to be excluded from the study, which led to possible omitting rainfall's impact on yields, while other meteorological variables were absent.

To address these gaps and enhance future analysis, it is recommended that future research should:

1. Develop more advanced missing data handling methods, such as Mean substitution, regression imputation and last observation carried forward (Kang 2013), as an alternative to simple variable culling
2. Establish more powerful outlier detection mechanisms to cope with extreme market fluctuations and shocks (Raymond 2018)
3. Utilise dimensionality reduction techniques like PCA or regularization methods like LASSO to address multicollinearity in weather variables
4. Combine SARIMAX with machine learning approaches such as random forests and neural networks to better capture nonlinearities and complex interactions in the data through hybrid modeling relationships and external shock effects, aligning with recent literature (Chalermrat Nontapa 2021; Wenjuan Liang 2023)

Addressing such limitations would likely lead to an improvement in the predictive accuracy and reliability of the model.

7 Conclusion

This study validates the significant value of time series forecasting in helping farmers, policy makers and industry participants to anticipate price movements and manage risks. The results produced by the SARIMAX(5,1,3) model proved to be effective in forecasting daily cocoa prices: by systematically integrating exogenous variables such as temperature, exchange rate, and the COVID-19 pandemic. The model not only efficiently captures long-term trends and seasonal fluctuations, but also accurately reflects key external factors affecting cocoa prices. In particular, by introducing dummy variables, the model strengthens its explanatory power for structural changes in the market, providing a reliable basis for decisions on production planning and pricing strategies. These findings underscore the model's ability to generalize beyond training data and its robustness in identifying critical directional shifts during turbulent conditions.

To improve predictive accuracy, it is advised that future research should address data limitations, utilize dimensionality reduction techniques or regularization techniques, and explore hybrid modeling techniques that combine SARIMAX with machine learning methods.

8 Appendix

```r
# Set your file path (adjust if needed)
file_path <- "/Users/karimhijazi/Desktop/coco457/Daily Prices_ICCO.csv"  # Cocoa price data

# Load libraries
library(readr)
library(ggplot2)
library(scales)
library(dplyr)
library(forecast)
library(zoo)

# ------------------------------------------------------------ #
# Load cocoa data
cocoa_data <- read_csv(file_path)
cocoa_data$Date <- as.Date(cocoa_data$Date, format = "%d/%m/%Y")

# Plot cocoa prices
ggplot(cocoa_data, aes(x = Date, y = `ICCO daily price (US$/tonne)`)) +
  geom_line(color = "brown", alpha = 0.6) +
  labs(title = "Daily Cocoa Prices Over Time",
       x = "Year",
       y = "Price (USD/tonne)") +
  scale_x_date(date_breaks = "2 years", date_labels = "%Y") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))


# ------------------------------------------------------------ #
# Load and prepare Ghana weather
ghana_data <- read_csv("/Users/karimhijazi/Desktop/coco457/Ghana_weather.csv")
names(ghana_data)[names(ghana_data) == "DATE"] <- "Date"
ghana_data$Date <- as.Date(ghana_data$Date, format = "%d/%m/%Y")
ghana_weather <- ghana_data %>%
  group_by(Date) %>%
  summarise(
    TAVG = mean(as.numeric(TAVG), na.rm = TRUE),
    PRCP = mean(as.numeric(PRCP), na.rm = TRUE)
  )


# Load and prepare Ivory Coast weather
ivory_data <- read_csv("/Users/karimhijazi/Desktop/coco457/Ivory_weather.csv")

# Convert Date column from character to Date
ivory_data$Date <- as.Date(ivory_data$Date, format = "%Y/%m/%d")


# Group by date and calculate daily means
ivory_weather <- ivory_data %>%
  group_by(Date) %>%
  summarise(
```

```
    IC_TAVG = mean(as.numeric(TAVG), na.rm = TRUE),
    IC_PRCP = mean(as.numeric(PRCP), na.rm = TRUE)
  )

# Load cleaned Ghana and Ivory Coast currency data
ghana_curr <- read_csv("/Users/karimhijazi/Desktop/coco457/cleaned_ghana_cedi.csv")
ivory_curr <- read_csv("/Users/karimhijazi/Desktop/coco457/cleaned_xof_cfa.csv")


#Load and clean Nigera, Cameroon, Ecuador weather
# Load
ecuador <- read_csv("/Users/karimhijazi/Desktop/coco457/Ecuador_data.csv")
nigeria <- read_csv("/Users/karimhijazi/Desktop/coco457/Nigeria_data.csv")
cameroon <- read_csv("/Users/karimhijazi/Desktop/coco457/Cameroon_data.csv")

#rename DATE
ecuador <- ecuador %>%
  rename(Date = DATE,
         TAVG = TEMP)

nigeria <- nigeria %>%
  rename(Date = DATE,
         TAVG = TEMP)

cameroon <- cameroon %>%
  rename(Date = DATE,
         TAVG = TEMP)

ecuador <- ecuador %>%
  group_by(Date) %>%
  summarise(TAVG_EC = mean(TAVG, na.rm = FALSE))

nigeria <- nigeria %>%
  group_by(Date) %>%
  summarise(TAVG_NI = mean(TAVG, na.rm = FALSE))

cameroon <- cameroon %>%
  group_by(Date) %>%
  summarise(TAVG_CAM = mean(TAVG, na.rm = FALSE))

# ------------------------------------------------------------- #
#check the available dates
range(cocoa_data$Date)
range(ghana_weather$Date)
range(ivory_weather$Date)
range(ghana_curr$Date)
range(ivory_curr$Date)
range(ecuador$Date)
range(nigeria$Date)
range(cameroon$Date)

start_date <- as.Date("1999-02-01")
end_date <- as.Date("2024-11-28")
```

```r
ghana_weather <- filter(ghana_weather, Date >= start_date & Date <= end_date)
ivory_weather <- filter(ivory_weather, Date >= start_date & Date <= end_date)
ghana_curr <- filter(ghana_curr, Date >= start_date & Date <= end_date)
ivory_curr <- filter(ivory_curr, Date >= start_date & Date <= end_date)
cocoa_data <- filter(cocoa_data, Date >= start_date & Date <= end_date)
cameroon <- filter(cameroon, Date >= start_date & Date <= end_date)
nigeria <- filter(nigeria, Date >= start_date & Date <= end_date)
ecuador <- filter(ecuador, Date >= start_date & Date <= end_date)


# ------------------------------------------------------------ #
# Merge all datasets
merged_data <- cocoa_data %>%
  left_join(ghana_weather, by = "Date") %>%
  left_join(ivory_weather, by = "Date") %>%
  left_join(ghana_curr, by = "Date") %>%
  left_join(ivory_curr, by = "Date") %>%
  left_join(cameroon, by = "Date") %>%
  left_join(nigeria, by = "Date") %>%
  left_join(ecuador, by = "Date")

# Rename currency columns
colnames(merged_data)[colnames(merged_data) == "Mid Rate"] <- "GHS"
colnames(merged_data)[colnames(merged_data) == "Price"] <- "XOF"


# ------------------------------------------------------------ #
# Create dummy variables
merged_data$Covid_Lag <- ifelse(merged_data$Date >= as.Date("2020-03-01") &
                                  merged_data$Date <= as.Date("2022-12-31"), 1, 0)
merged_data$Post_2023 <- ifelse(merged_data$Date >= as.Date("2023-01-01"), 1, 0)
merged_data$Post_2023_Trend <- ifelse(merged_data$Date >= as.Date("2023-01-01"),
                                      as.numeric(merged_data$Date - as.Date("2023-01-01")),
                                      0)


# ------------------------------------------------------------ #
# Create time series and exogenous variables
merged_data <- na.omit(merged_data)
price_ts <- ts(merged_data$`ICCO daily price (US$/tonne)`, frequency = 365)

xreg_vars <- cbind(
  merged_data$TAVG,
  merged_data$IC_TAVG,
  # remove TAVG_EC, TAVG_NI, TAVG_CAM
  merged_data$GHS,
  merged_data$XOF,
  merged_data$Covid_Lag,
  merged_data$Post_2023,
  merged_data$Post_2023_Trend
)

colnames(xreg_vars) <- c("Ghana_TAVG", "Ivory_TAVG",
```

```
                        "GHS", "XOF", "Covid_Lag", "Post_2023", "Post_2023_Trend")


# ------------------------------------------------------------ #
# Fit SARIMAX models
model_212 <- Arima(price_ts, order = c(2,1,2), seasonal = list(order = c(1,0,1),
                                                period = 7), xreg = xreg_vars)
model_313 <- Arima(price_ts, order = c(3,1,3), seasonal = list(order = c(1,0,1),
                                                period = 7), xreg = xreg_vars)
model_414 <- Arima(price_ts, order = c(4,1,4), seasonal = list(order = c(1,0,1),
                                                period = 7), xreg = xreg_vars)
model_513 <- Arima(price_ts, order = c(5,1,3), seasonal = list(order = c(1,0,1),
                                                period = 7), xreg = xreg_vars)

# Compare models
comparison <- data.frame(
  Model = c("SARIMAX(2,1,2)", "SARIMAX(3,1,3)", "SARIMAX(4,1,4)", "SARIMAX(5,1,3)"),
  AIC = c(AIC(model_212), AIC(model_313), AIC(model_414), AIC(model_513)),
  BIC = c(BIC(model_212), BIC(model_313), BIC(model_414), BIC(model_513))
)
print(comparison)


# ------------------------------------------------------------ #
# ------------------------------------------------------------ #
# Choose best model (example: SARIMAX(4,1,4), update if different)
best_model <- model_513

# Create future exogenous variables for 120 days
last_date <- max(merged_data$Date)
future_dates <- seq(from = last_date + 1, by = "day", length.out = 120)

future_xreg <- as.matrix(data.frame(
  TAVG = mean(merged_data$TAVG, na.rm = TRUE),
  IC_TAVG = mean(merged_data$IC_TAVG, na.rm = TRUE),
  # remove TAVG_EC, TAVG_NI, TAVG_CAM
  GHS = mean(merged_data$GHS, na.rm = TRUE),
  XOF = mean(merged_data$XOF, na.rm = TRUE),
  Covid_Lag = 0,
  Post_2023 = 1,
  Post_2023_Trend = as.numeric(future_dates - as.Date("2023-01-01"))
))


# Forecast 120 days ahead
forecast_120 <- forecast(best_model, xreg = future_xreg, h = 120)


#----------------------------
# Rebuild plot_df for 120-day forecast
full_price <- c(as.numeric(price_ts), as.numeric(forecast_120$mean))
upper_95 <- c(rep(NA, length(price_ts)), forecast_120$upper[,2])
lower_95 <- c(rep(NA, length(price_ts)), forecast_120$lower[,2])
```

```r
all_dates <- c(merged_data$Date, future_dates)

plot_df <- data.frame(
  Date = all_dates,
  Price = full_price,
  Upper = upper_95,
  Lower = lower_95,
  Type = c(rep("Historical", length(price_ts)), rep("Forecast", 120))
)

# Plot using ggplot with dates and CI
ggplot(plot_df, aes(x = Date)) +
  geom_line(aes(y = Price, color = Type), linewidth = 1) +
  geom_ribbon(aes(ymin = Lower, ymax = Upper), fill = "blue", alpha = 0.2) +
  labs(title = "Cocoa Prices with 120-Day Forecast",
       x = "Date", y = "Price (USD/tonne)") +
  scale_color_manual(values = c("Historical" = "brown", "Forecast" = "blue")) +
  theme_minimal()


#---------------------------
# Metrics:

# Reload the full cocoa data (with extended date range)
full_cocoa_data <- read_csv(file_path)
full_cocoa_data$Date <- as.Date(full_cocoa_data$Date, format = "%d/%m/%Y")

# Extract actual 120 prices after the last date in merged_data
real_prices_120 <- full_cocoa_data %>%
  filter(Date > last_date & Date <= last_date + 120) %>%
  pull(`ICCO daily price (US$/tonne)`)

# Compare forecast to actual values
predicted <- as.numeric(forecast_120$mean)
actual <- real_prices_120  # make sure this is a numeric vector of length 120

# Calculate metrics
rmse <- sqrt(mean((actual - predicted)^2))
mae <- mean(abs(actual - predicted))
mape <- mean(abs((actual - predicted) / actual)) * 100

# Print results
cat("RMSE:", round(rmse, 2), "\n")
cat("MAE:", round(mae, 2), "\n")
cat("MAPE:", round(mape, 2), "%\n")


#-----------------------------------------#
#FIGURES:

# === (1) ACF/PACF plots ===
acf(prices_ts)
pacf(prices_ts)
```

```r
diff_prices_ts <- diff(prices_ts)

acf(diff_prices_ts)
pacf(diff_prices_ts)


# === (2) Cocoa prices ===
ggplot(merged_data, aes(x = Date)) +
  geom_line(aes(y = TAVG, color = "Ghana TAVG"), alpha = 0.1) +
  geom_line(aes(y = IC_TAVG, color = "Ivory Coast TAVG"), alpha = 0.1) +
  geom_smooth(aes(y = TAVG, color = "Ghana TAVG"), method = "loess",
              span = 0.05, se = FALSE, linewidth = 1) +
  geom_smooth(aes(y = IC_TAVG, color = "Ivory Coast TAVG"), method = "loess",
              span = 0.05, se = FALSE, linewidth = 1) +
  labs(
    x = "Date",
    y = "Average Temperature (°F)",
    color = "Country"
  ) +
  theme_minimal(base_size = 8) +
  scale_x_date(date_breaks = "5 years", date_labels = "%Y") +
  scale_y_continuous(limits = c(73, 88)) +
  theme(legend.position = "bottom")


# === (3) Weather ===
ggplot(merged_data, aes(x = Date)) +
  geom_line(aes(y = TAVG, color = "Ghana TAVG"), alpha = 0.1) +
  geom_line(aes(y = IC_TAVG, color = "Ivory Coast TAVG"), alpha = 0.1) +
  geom_smooth(aes(y = TAVG, color = "Ghana TAVG"), method = "loess",
              span = 0.05, se = FALSE, linewidth = 1) +
  geom_smooth(aes(y = IC_TAVG, color = "Ivory Coast TAVG"), method = "loess",
              span = 0.05, se = FALSE, linewidth = 1) +
  labs(
    x = "Date",
    y = "Average Temperature (°F)",
    color = "Country"
  ) +
  theme_minimal(base_size = 8) +
  scale_x_date(date_breaks = "5 years", date_labels = "%Y") +
  scale_y_continuous(limits = c(73, 88)) +
  theme(legend.position = "bottom")


# === (4) Weather ===
ggplot(merged_data, aes(x = Date)) +
  geom_line(aes(y = GHS, color = "GHS (Ghana Cedi)")) +
  geom_line(aes(y = XOF, color = "XOF (West African CFA Franc)")) +
  labs(
       x = "Date", y = "Exchange Rate",
       color = "Currency") +
  theme_minimal(base_size = 8) +
  scale_x_date(date_breaks = "5 years", date_labels = "%Y") +
```

```
  theme(legend.position = "bottom")


# === (4) Model Training and Validation ===
# Comparison table of all trained models
print(comparison)  # Already contains Model, AIC, and BIC


# === (5) Performance Evaluation Using Metrics ===



# Create metrics table
metrics_df <- data.frame(
  Metric = c("RMSE", "MAE", "MAPE (%)"),
  Value = round(c(rmse, mae, mape), 2)
)

# Print in console
print(metrics_df)

# Save metrics table as PNG to Downloads
library(gridExtra)
png("~/Downloads/forecast_metrics_table.png", width = 600, height = 300, res = 150)
grid.table(metrics_df)
dev.off()

# === Forecast vs Actual Only (no historical) ===

# Trim forecast and dates to match actual
n_actual <- length(actual)
forecast_only_df <- data.frame(
  Date = future_dates[1:n_actual],
  Forecast = as.numeric(forecast_120$mean[1:n_actual]),
  Actual = actual[1:n_actual],
  Lower = forecast_120$lower[1:n_actual, 2],
  Upper = forecast_120$upper[1:n_actual, 2]
)

# Plot
ggplot(forecast_only_df, aes(x = Date)) +
  geom_line(aes(y = Forecast), color = "blue", linewidth = 1) +
  geom_ribbon(aes(ymin = Lower, ymax = Upper), fill = "blue", alpha = 0.2) +
  labs(title = "Forecast of Cocoa Prices (120-Day Horizon)",
       subtitle = "Blue = Forecast | Red dashed = Actual",
       x = "Date", y = "Price (USD/tonne)") +
  theme_minimal()

# SIde by side:
library(ggplot2)
library(gridExtra)
library(ggplot2)
library(cowplot)
```

```r
# Plot 1: Daily Cocoa Prices (from 2015)
p1 <- ggplot(cocoa_data %>% filter(Date >= as.Date("2015-01-01")),
             aes(x = Date, y = `ICCO daily price (US$/tonne)`)) +
  geom_line(color = "brown", alpha = 0.6) +
  labs(title = "Daily Cocoa Prices Over Time", x = "Year", y = "Price (USD/tonne)") +
  scale_x_date(date_breaks = "2 years", date_labels = "%Y") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Plot 2: Forecast with legend (from 2015)
p2 <- ggplot(plot_df %>% filter(Date >= as.Date("2015-01-01")), aes(x = Date)) +
  geom_line(aes(y = Price, color = Type), linewidth = 1) +
  geom_ribbon(aes(ymin = Lower, ymax = Upper), fill = "blue", alpha = 0.2) +
  labs(title = "Cocoa Prices with 120-Day Forecast", x = "Date", y = "Price (USD/tonne)") +
  scale_color_manual(values = c("Historical" = "brown", "Forecast" = "blue")) +
  theme_minimal() +
  theme(legend.position = "bottom")

# Stack vertically, legend included
stacked_plot <- plot_grid(
  p1,
  p2,
  ncol = 1,
  rel_heights = c(1, 1.2)
)

# Save to Downloads
ggsave("~/Downloads/cocoa_forecast_stacked.png", stacked_plot,
       width = 14, height = 10, dpi = 300, bg = "white")
```

References

Adeoye, I. A., A. A. Babajide, and O. E. Folarin. 2019. "Export Function of Cocoa: Production, Exchange Rate Volatility and Prices in Nigeria." *Journal of Economics and Sustainable Development* 10 (8): 21–31. https://www.researchgate.net/publication/333101292_Export_Function_of_Cocoa_Production_Exchange_Rate_Volatility_and_Prices_in_Nigeria.

Chalermrat Nontapa, Nicha Kaewhawong, Chainarong Kesamoon. 2021. "A New Hybrid Forecasting Using Decomposition Method with SARIMAX Model and Artificial Neural Network." *International Journal of Mathematics and Computer Science* 16 (4): 1341–54. https://future-in-tech.net/16.4/R-Nontpa.pdf.

Dzokoto, Vivian Afi Abui, Jessica Young, and Clifford Edwin Mensah. 2010. "A Tale of Two Cedis: Making Sense of a New Currency in Ghana." *Journal of Economic Psychology* 31 (4): 520–26. https://doi.org//10.1016/j.joep.2010.03.014.

Ghana, Bank of. 2025. "Historical Interbank FX Rates." https://www.bog.gov.gh/treasury-and-the-markets/historical-interbank-fx-rates/.

Giri, Anisha, and Vijay Ray Giri. 2024. "Forecasting Cauliflower Prices in Nepal: A Comparative Analysis Using Seasonal Time Series and Nonlinear Models." *Cogent Food & Agriculture* 10 (January): 2340155. https://doi.org/10.1080/23311932.2024.2340155.

Haoyang Wu, Minfeng Zhu, Huaili Wu. 2017. "A New Method of Large-Scale Short-Term Forecasting of Agricultural Commodity Prices: Illustrated by the Case of Agricultural Markets in Beijing." *Journal of Big Data* 4: 1. https://doi.org/10.1186/s40537-016-0062-3.

Hyndman, Rob J., and George Athanasopoulos. 2018. *Forecasting: Principles and Practice*. 2nd ed. OTexts. https://otexts.com/fpp2/.

International Cocoa Organization (ICCO). 2025. "ICCO Daily Cocoa Prices (USD/Tonne)." https://www.icco.org/statistics/daily-prices/.

Investing.com. 2025. "Historical Data: USD/XOF Exchange Rates." https://ca.investing.com/currencies/usd-xof-historical-data.

J.P. Morgan. 2024. "Cocoa Prices Hit Record Highs: What's Next?" J.P. Morgan Global Research. https://www.jpmorgan.com/insights/global-research/commodities/cocoa-prices.

Kang, H. 2013. "The Prevention and Handling of the Missing Data." *Korean Journal of Anesthesiology* 64 (5): 402–6. https://doi.org/10.4097/kjae.2013.64.5.402.

Karishma Harrykissoon, Patrick Hosein. 2023. "Crop Price Prediction: A Comparison of the Recursive and Direct Forecasting Strategies on SARIMAX Models." The University of the West Indies. https://lab.tt/wp-content/uploads/2023/05/Crop-Price-Prediction.pdf.

National Centers for Environmental Information (NCEI). 2025. "Global Summary of the Day (GSOD) Weather Data." https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00516/html.

Raymond, Anjanette. 2018. "Information and the Regulatory Landscape: A Growing Need to Reconsider Existing Legal Frameworks." Social Science Research Network. https://doi.org/10.2139/ssrn.3349737.

Tetsuya Michinaka, Kazuya Tamura, Hirofumi Kuboyama, and Nobuyuki Yamamoto. 2016. "Forecasting Monthly Prices of Japanese Logs." *Forests* 7 (5): 94. https://doi.org/10.3390/f7050094.

Theerthagiri, Prasannavenkatesan, and A. Usha Ruby. 2023. "Seasonal Learning Based ARIMA Algorithm for Prediction of Brent Oil Price Trends." *Multimedia Tools and Applications* 82: 24485–504. https://doi.org/10.1007/s11042-023-14819-x.

Wenjuan Liang, Pan Hu, Ailing Hu. 2023. "Estimating the Tuberculosis Incidence Using a SARIMAX-NNARX Hybrid Model by Integrating Meteorological Factors in Qinghai Province, China." *International Journal of Biometeorology* 67: 55–65. https://doi.org/10.1007/s00484-022-02385-0.

Yogesh Funde, Akshay Damani. 2023. "Comparison of ARIMA and Exponential Smoothing Models in Prediction of Stock Prices." *Journal of Project Management* 17. https://doi.org/10.5750/jpm.v17i1.2017.