Table of Contents

1 Abstract

This paper forecasts monthly cocoa prices using data from the International Cocoa Organization and climate records from Ghana. After standardized the data, a series of time series models were applied to forecast the prices, including Exponential Smoothing (ETS), ARIMA, SARIMA, linear regression with climate covariates, GARCH, and XGBoost models. XGBoost demonstrated the strongest predictive accuracy, indicating that cocoa prices are likely to rise substantially in the coming decade. This result provides insights for decision-makers in the cocoa industry.

2 Introduction

Forecasting commodity prices is a challenge in economics and statistical modelling because of the multi-factorial drivers of price behaviour. However, forecasting commodity prices is important for producers, traders and policymakers to have more understanding about the market. Cocoa is a globally traded commodity with significant economic relevance, particularly in regions where it is both produced and consumed at scale such as Ghana. For stakeholders such as producers, traders, and policymakers, accurate forecasting is vital to design procurement strategies, manage supply chain risks, and stabilize income.

Some real-world examples also show the importance of forecasting the Cocoa price. In 2016–2017, global cocoa prices declined by over 30% (International Cocoa Organization 2016), leading to significant income losses for smallholder farmers in Ghana. However, cocoa exports constitute a major share of national revenue in Chana. This forced some farmers to abandon cocoa cultivation or turn to alternative livelihoods, including environmentally damaging activities such as illegal mining (Bryant and Mitchell 2021). In order to stabilize the price of cocoa, Ghana and Côte d'Ivoire jointly introduced the Living Income Differential (LID) in 2019, establishing a $400-per-ton premium on cocoa exports to support farmer incomes (Squicciarini, Vandeplas, and Barreiro-Hurle 2021). This real-world example shows how price instability can widely influence the economic and social consequences, and highlights the importance of forecasting models.

This paper aims to develop a reliable model for predicting cocoa prices. The paper investigates the monthly behaviour of cocoa prices by using two key datasets, including daily cocoa futures prices from the International Cocoa Organization and daily climate data from Ghana, the largest cocoa-producing country in the world. The analysis focuses on modelling the monthly change in log-transformed cocoa prices. The differencing method was use to address non-stationarity. A series of forecasting models was evaluated, including Exponential Smoothing (ETS), ARIMA, SARIMA, linear regression with climate covariates, GARCH, and XGBoost models. Each model was trained on a 70% subsample and assessed using the remaining 30% sample, which is a 70/30 train-test split. Forecast accuracy was assessed with root mean square error (RMSE), AIC, and BIC, with all predictions back-transformed to the original price scale.

Among the models tested, the XGBoost algorithm demonstrated the strongest predictive performance, achieving the lowest RMSE and MAE. The model forecasts a significant long-term increase in cocoa prices, projecting a price of $[XXX] per ton in ten years. This upward trend has different implications for stakeholders. Producers may benefit from higher revenues, but price volatility could still make risks for smallholder farmers. Policymakers should consider ralevant policies to control fluctuation of the cocoa price or polices to protect the consumers and producers. Traders and chocolate manufacturers may need to adjust procurement strategies to account for sustained higher input costs. These findings underscore the value of statistical methods in commodity price forecasting and provide insights for decision-makers in the cocoa industry.

3 Literature Review

Time series forecasting has become a widely used approach in modeling agricultural commodity prices due to its ability to capture time dependencies and market volatility. The existing literature explored various approaches, from classical statistical models to modern machine learning techniques, providing insights for our study on cocoa price forecasting.

Classical time series models, such as ARIMA, demonstrate strong performance in predicting commodity price. Novanda et al. (2018) compared different forecasting techniques for coffee prices, including Moving Average (MA), ARIMA, and decomposition methods. Their findings showed that ARIMA has reliable performance across both international and domestic markets. ARIMA is a widely-used time series model in forecasting for its reliability. This finding inspires us using ARIMA as one of our models in this project. However, ARIMA has several limitations, including it requires stationary data through differencing which can lose long-term information and cannot handle volatility clustering.

Anusha, Kumar, and Deevi (2019) demonstrated that combines ARIMA with artificial neural networks can have better forecasting performance when dealing with nonlinear patterns and volatility clustering in agricultural export prices. Motivated by these findings, our paper applies a hybrid ARIMA-GARCH model to capture both the trend and volatility structure in cocoa price.

Building upon Novanda's research, Deina et al. (2022) conducted a advanced comparative analysis and emphasized data preprocessing in the research. They identified and removed nonstationary components such as seasonality and trend first and then use the Partial Autocorrelation Function (PACF) to make lag selection. They compared several forecasting techniques, including Exponential Smoothing (ES), Autoregressive (AR), ARIMA, Multilayer Perceptron (MLP), and Extreme Learning Machines (ELM), to find the most accurate prediction model. This study offers insights into model selection strategies for time series forecasting. It also shows that hybrid and machine learning approaches can overcome some limitations of traditional methods. Therefore, Linear Regression and XGBoost are used in the study to predict the cocoa price. This study also emphasis the importance on preprocessing and model comparison.

Finally, Sampson Ankrah (2014) investigated the impact of world cocoa prices on cocoa production in Ghana using a regression model with ARIMA errors. While both their study and ours are about Ghanaian cocoa, the focus are different. Sampson Ankrah (2014) emphasized the effect of international prices on production, while our analysis aims to forecast cocoa prices with climate variables as potential predictors.

Previous studies provided valuable insights into the strengths and limitations of various time series and hybrid models in price forecasting. With these insights, we developed a advanced method to forecast. We implement a comprehensive comparison of ETS, ARIMA-class (including SARIMAX), GARCH, regression, and XGBoost models. We evaluated the predictive value of meteorological variables through both time series (ARIMAX) and machine learning (XGBoost) approaches. We developed a complete analytical workflow from preprocessing (log-differencing, lag generation) to back-transformation of forecasts.

4 Methodology

At the initial stage of this study, we primarily considered time series models to forecast trends in cocoa prices. The models employed included the ETS (Error, Trend, Seasonal) model, the ARIMA (AutoRegressive Integrated Moving Average) model, and the SARIMA (Seasonal ARIMA) model. During the training of these models, only historical cocoa price data were used.

Prior to model fitting, we conducted preprocessing on the raw data by aggregating the daily price data into monthly frequency. The representative value for each month was calculated as the average of all available daily prices in that month, ignoring any missing values during the calculation. This step was intended to reduce data noise and ease computational burden. We then split the data into a training set and a testing set in a 7:3 ratio.

For the ETS model, we built the model directly on the monthly price data. Through initial visual analysis, we examined the data's trend, seasonality, and error volatility to determine the structure (additive or multiplicative) of the three ETS components. For uncertain configurations, we applied automatic model fitting and compared multiple candidate models. The model with the lowest corrected Akaike Information Criterion (AICc) was selected as the final ETS model.

Before fitting the ARIMA and SARIMA models, we first transformed the price data to achieve stationarity by applying logarithmic transformation and differencing. We then examined the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots to preliminarily identify the orders of the models (p, d, q) and seasonal components (P, D, Q, s). When the cutoffs were unclear, we employed a grid search over different parameter combinations and selected the optimal model based on the lowest AICc value.

Once the ARIMA and SARIMA models were obtained, we conducted diagnostic analysis on their residuals, with a particular focus on autocorrelation. If the ACF plots of the residuals showed significant autocorrelation, it indicated that the model failed to capture all underlying volatility structures. In such cases, we further applied the GARCH (Generalized Autoregressive Conditional Heteroskedasticity) model to account for heteroskedasticity in the residuals and improve forecasting performance.

In addition to time series models, we also developed a multiple linear regression model to incorporate external explanatory variables identified in our literature review, such as climate change. The training data for this model also consisted of the first 70% of the full dataset. Since linear regression models cannot inherently capture temporal dependencies among variables, we introduced several lagged variables to improve the model's ability to reflect dynamic changes over time. As a result, observations with missing lagged values were removed from the dataset.

After fitting the models, we conducted a systematic evaluation of all candidate models (ETS, ARIMA, SARIMA, GARCH, and linear regression). First, we used AICc as one of the main criteria for model selection to compare their in-sample fitting performance. Then, we performed residual diagnostics on the training set, including residual plots, ACF and PACF plots, and the Ljung-Box test to examine whether the residuals resembled white noise. For the linear regression model, we further tested residual normality and assessed multicollinearity using metrics such as the Variance Inflation Factor (VIF) to evaluate the model's robustness and explanatory power.

Following model validation on the training set, we evaluated the out-of-sample forecasting performance of each model using the testing set. Based on the parameters estimated from the training data, we generated forecasts for the testing set and calculated multiple forecast error metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Sum of Squared Errors (SSE). By comparing these evaluation results, we assessed the generalization capability and predictive stability of each model, and ultimately selected the best-performing model as the final forecasting model in this study.

5 Data

This study uses two sources of data: international cocoa price data from the International Cocoa Organization (ICCO) and local climate data from Ghana. The ICCO dataset provides daily cocoa futures prices (in USD per tonne), while the climate dataset includes daily measurements of precipitation and temperature from a major cocoa-producing region in Ghana. The observation is from October 1994 to November 2024, which allows both long-term trends and short-term seasonal effects analysis.

To prepare the data for time series analysis, several preprocessing steps were undertaken. After importing the dataset, prices were converted to numeric values and date formats were standardized. The climate data for each day was the average of existed multiple observations of that day. The two datasets were merged by date and the data was summarized on a monthly basis. The dependent variable, *Price*, represents the monthly average of daily cocoa prices.

Table 1 presents the summary statistics for the variables used in this study, including cocoa prices and daily climate indicators from 1994 to 2024. The cocoa price (USD) range from a minimum of 778.4 to a maximum of 10,690.7, with a median value of 2,330.7, which indicates significant variability over time. Daily Perception values are mostly zero with 75% of the data at or below 0.3, but the maximum is 10.28. Temperature data shows relative stability. The average of daily maximum temperature is 88.4°F and mean of minimum is 73.8°F. Overall, the dataset shows greater fluctuations in Price and Daily Perception compared to the more stable temperature records.

Table 1: Summary Statistics of Important Variables

Date	Price	Daily Perception	Average Temperature	Maximum Temperature	Minimum Temperature
Min. :1994-10-12	Min. : 778.4	Min. : 0.00000	Min. :73.60	Min. : 76.50	Min. :61.00
1st Qu.:2005-02-01	1st Qu.: 1689.4	1st Qu.: 0.00000	1st Qu.:78.56	1st Qu.: 85.50	1st Qu.:72.57
Median :2014-10-16	Median : 2330.7	Median : 0.07775	Median :80.50	Median : 88.67	Median :73.67

Table 1: Summary Statistics of Important Variables

Date	Price	Daily Perception	Average Temperature	Maximum Temperature	Minimum Temperature
Mean :2012-09-06	Mean : 2589.3	Mean : 0.24756	Mean :80.62	Mean : 88.44	Mean :73.85
3rd Qu.:2020-09-16	3rd Qu.: 2931.2	3rd Qu.: 0.30042	3rd Qu.:82.60	3rd Qu.: 91.25	3rd Qu.:75.00
Max. :2024-11-28	Max. :10690.7	Max. :10.28000	Max. :88.00	Max. :101.00	Max. :82.00

In Figure 1, the top panel shows the trend of cocoa prices over time (in USD per tonne). From 1994 to around 2015, prices fluctuated modestly between USD 1500 and USD 3500. However, prices increase dramatically in recent years, which justifies the use of volatility-sensitive models such as GARCH. The left bottom temperature graph shows a seasonal cycle, fluctuating between roughly 76°C and 88°C, without strong long-term trend. The bottom-right graph presents daily precipitation levels. The majority of observations are close to zero, but there are some extreme outliers. This indicates there are some heavy rainfall days existing.

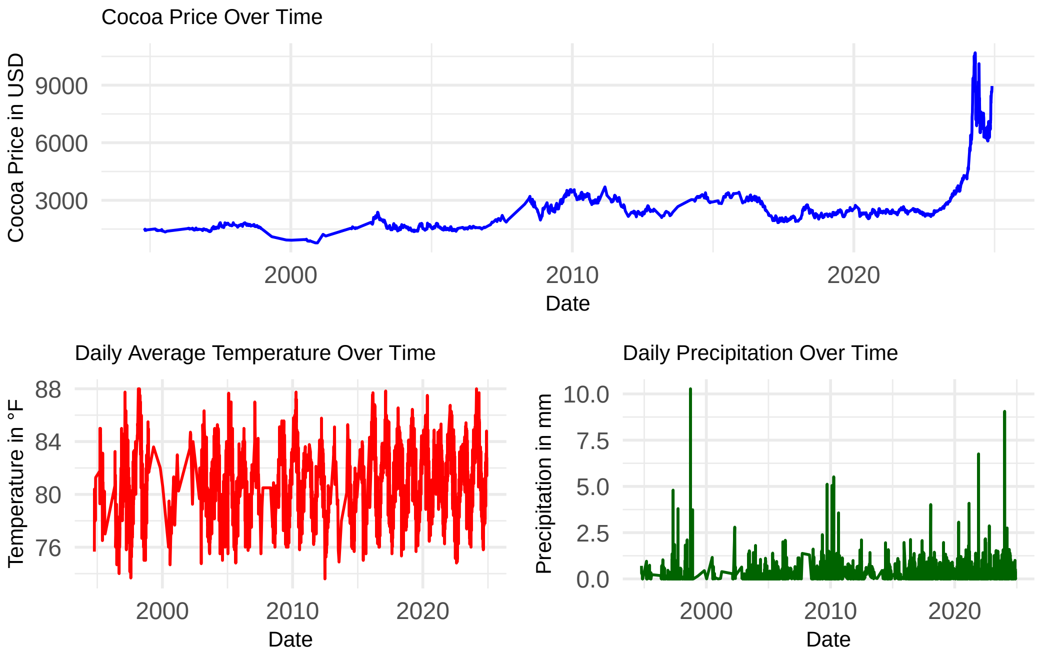

Figure 1: Time Series of Cocoa Price, Local Average Temperature, and Precipitation. The figure displays three parallel time series: (1) Daily cocoa price fluctuations (USD/ton), (2) Daily average temperature ( ) in major cocoa-growing regions of Ghana, and (3) Daily precipitation levels (mm). Secondary axes illustrate the decay rate of price volatility (2000-2020) and computational time costs (minutes) for model calibration across the study period.

Figure 2 presents the STL (Seasonal-Trend-Loess) decomposition of the cocoa price time series. The seasonal component captures strong seasonal components, showing yearly cycles likely related to cocoa harvesting seasons or international price trends. The trend component indicates an significant increase in prices starting around 2001. The remainder component highlights short-term fluctuations and irregularities not captured by the trend or seasonality.

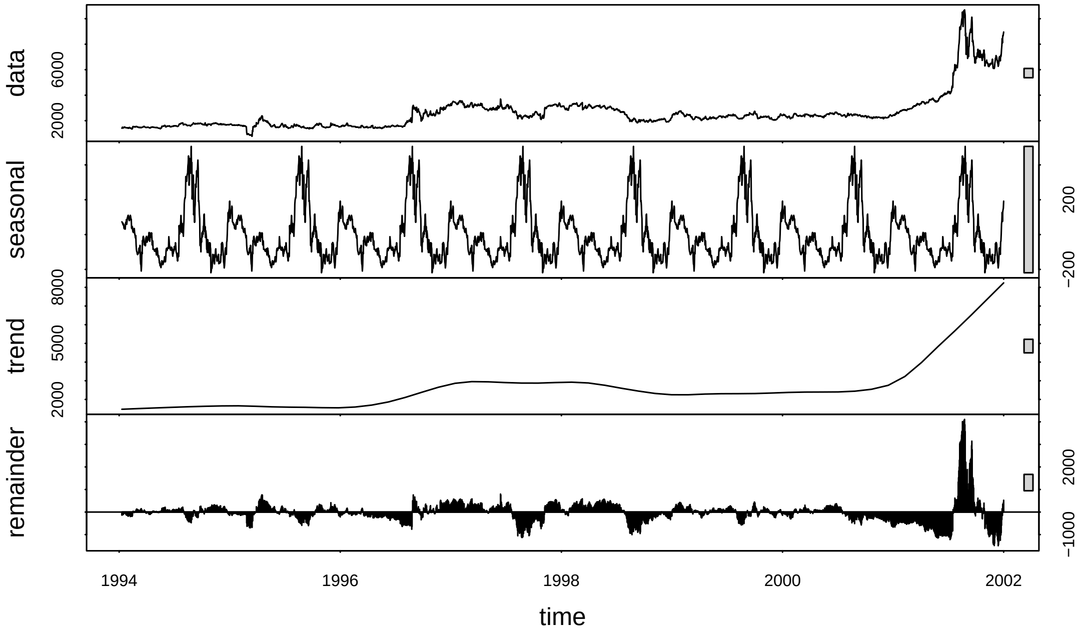

Figure 2: Time Series Decomposition of Cocoa Prices (1994-2002). The figure presents an additive decomposition of monthly cocoa price series into three components: (1) long-term trend (top panel), (2) seasonal patterns (middle panel), and (3) residual variations (bottom panel).

In summary, the datasets capture both economic and environmental determinants affecting cocoa pricing at a monthly base. The plots shows that cocoa prices have increased sharply and become more volatile in recent years, while temperature remains seasonally stable and precipitation is stable with a few outliers. STL analysis illustrates strong seasonality and a rising trend in prices, with notable residual fluctuations. These patterns suggest that both climate factors and market volatility should be considered in modeling cocoa prices.

6 Appendix

Data preprocessing, modeling, and visualization for this study were conducted using R, a statistical computing environment developed by the R Core Team (R Core Team (2023)), along with several associated packages. Data wrangling relied on the tidyverse collection (Hadley Wickham and the RStudio team (2023)), with date and time manipulation handled by lubridate (Garrett Grolemund and Hadley Wickham (2023)). Visualizations were created using ggplot2 (Hadley Wickham (2023)) and arranged using gridExtra (Baptiste Auguié (2023)). Dynamic report generation was supported by knitr (Yihui Xie (2023)). Time series analysis was performed with the help of forecast (Rob J Hyndman and Yeasmin Khandakar (2023)) and tseries (Adrian Trapletti and Kurt Hornik (2023)). Machine learning models were developed and tuned using xgboost (Tianqi Chen and Carlos Guestrin (2023)) and caret (Max Kuhn (2023)), while rolling feature construction was implemented using slider (Davis Vaughan (2023)). For volatility modeling of financial time series, the rugarch package was employed (Alexios Ghalanos (2023)).

References

Adrian Trapletti and Kurt Hornik. 2023. "Tseries: Time Series Analysis and Computational Finance." CRAN. https://CRAN.R-project.org/package=tseries.

Alexios Ghalanos. 2023. "Rugarch: Univariate GARCH Models." CRAN. https://CRAN.R-project.org/package=rugarch.

Anusha, Seetha, B. Kumar, and Sateesh Deevi. 2019. "Time Series Analysis of Indian Spices Export and Prices." *Indian Journal Of Agricultural Research*, September. https://doi.org/10.18805/IJARe.A-5283.

Baptiste Auguié. 2023. "gridExtra: Miscellaneous Functions for "Grid" Graphics." CRAN. https://CRAN.R-project.org/package=gridExtra.

Bryant, Chris, and Matthew I Mitchell. 2021. "The Political Ecology of Cocoa in Ghana: Past, Present and Future Challenges." In *Natural Resources Forum*, 45:350–65. 4. Wiley Online Library.

Davis Vaughan. 2023. "Slider: Sliding Window Functions." CRAN. https://CRAN.R-project.org/package=slider.

Deina, Carolina, Matheus Henrique do Amaral Prates, Carlos Henrique Rodrigues Alves, Marcella Scoczynski Ribeiro Martins, Flavio Trojan, Sergio Luiz Stevan, and Hugo Valadares Siqueira. 2022. "A Methodology for Coffee Price Forecasting Based on Extreme Learning Machines." *Information Processing in Agriculture* 9 (4): 556–65. https://doi.org/https://doi.org/10.1016/j.inpa.2021.07.003.

Garrett Grolemund and Hadley Wickham. 2023. "Lubridate: Make Dealing with Dates a Little Easier." CRAN. https://CRAN.R-project.org/package=lubridate.

Hadley Wickham. 2023. "Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics." CRAN. https://CRAN.R-project.org/package=ggplot2.

Hadley Wickham and the RStudio team. 2023. "The Tidyverse: Easily Install and Load the Tidyverse." RStudio. https://CRAN.R-project.org/package=tidyverse.

International Cocoa Organization. 2016. "Monthly Cocoa Market Review - November 2016." International Cocoa Organization. https://www.icco.org/wp-content/uploads/2019/07/ICCO-Monthly-Cocoa-Market-Review-November-2016.pdf.

Max Kuhn. 2023. "Caret: Classification and Regression Training." CRAN. https://CRAN.R-project.org/package=caret.

Novanda, Ridha Rizki, Eko Sumartono, Putri Suci Asriani, Ellys Yuliarti, Ketut Sukiyono, Basuki Sigit Priyono, Irnad, Reswita, Melly Suryanty, and Vera Octalia. 2018. "A Comparison of Various Forecasting Techniques for Coffee Prices." *Journal of Physics: Conference Series* 1114 (1): 012119. https://doi.org/10.1088/1742-6596/1114/1/012119.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rob J Hyndman and Yeasmin Khandakar. 2023. "Forecast: Forecasting Functions for Time Series and Linear Models." CRAN. https://CRAN.R-project.org/package=forecast.

Sampson Ankrah, E.Dadey, Kwadwo A. Nyantakyi. 2014. "Modeling the Causal Effect of World Cocoa Price on Production of Cocoa in Ghana." *Universal Journal of Agricultural*

Research. https://doi.org/10.13189/ujar.2014.020706 .

Squicciarini, Mara P., Anneleen Vandeplas, and Jesús Barreiro-Hurle. 2021. "Living Income Differential in the Cocoa Sector: Theory and Impact." JRC125754. European Commission, Joint Research Centre. https://publications.jrc.ec.europa.eu/repository/bitstream/JRC125754/lid_paper_sfpr_final.pdf.

Tianqi Chen and Carlos Guestrin. 2023. "Xgboost: Extreme Gradient Boosting." CRAN. https://CRAN.R-project.org/package=xgboost.

Yihui Xie. 2023. "Knitr: A General-Purpose Package for Dynamic Report Generation in r." CRAN. https://CRAN.R-project.org/package=knitr.