



**HAL**  
open science

# Investigating and forecasting the impact of crop production shocks on global commodity prices

Rotem Zelingher, David Makowski

► **To cite this version:**

Rotem Zelingher, David Makowski. Investigating and forecasting the impact of crop production shocks on global commodity prices. *Environmental Research Letters*, 2023, 19 (1), pp.014026. 10.1088/1748-9326/ad0dda . hal-04341715

**HAL Id: hal-04341715**

**<https://hal.science/hal-04341715>**

Submitted on 30 Apr 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

LETTER • **OPEN ACCESS**

## Investigating and forecasting the impact of crop production shocks on global commodity prices

To cite this article: Rotem Zelingher and David Makowski 2024 *Environ. Res. Lett.* **19** 014026

View the [article online](#) for updates and enhancements.

You may also like

- [Is voluntary certification of tropical agricultural commodities achieving sustainability goals for small-scale producers? A review of the evidence](#)  
Ruth S DeFries, Jessica Fanzo, Pinki Mondal et al.
- [Development of Superior Plantation Commodities Based on Land Suitability in Pining Sub District Gayo Lues Regency Aceh Province](#)  
I Alfiansyah and Lindawati
- [Identifying the community structure of the food-trade international multi-network](#)  
S Torreggiani, G Mangioni, M J Puma et al.

The Breath Biopsy® Guide  
Fourth edition

FREE

DOWNLOAD THE FREE E-BOOK

BREATH BIOPSY

OWLSTONE MEDICAL

ENVIRONMENTAL RESEARCH  
LETTERS

## LETTER

## Investigating and forecasting the impact of crop production shocks on global commodity prices

## OPEN ACCESS

RECEIVED  
23 April 2023REVISED  
9 November 2023ACCEPTED FOR PUBLICATION  
17 November 2023PUBLISHED  
5 December 2023

Original Content from  
this work may be used  
under the terms of the  
[Creative Commons  
Attribution 4.0 licence](#).

Any further distribution  
of this work must  
maintain attribution to  
the author(s) and the title  
of the work, journal  
citation and DOI.

Rotem Zelingher<sup>1,2,\*</sup> and David Makowski<sup>3</sup> <sup>1</sup> Advancing Systems Analysis, International Institute for Applied Systems analysis, Schlossplatz 1, 2361 Laxenburg, Austria<sup>2</sup> School of Social and Policy Studies and The Steinhardt Museum of Natural History, Tel Aviv University, Tel Aviv, Israel<sup>3</sup> Université Paris-Saclay, INRAE, AgroParisTech, UMR MIA-PS, 91120 Palaiseau, France

\* Author to whom any correspondence should be addressed.

E-mail: [zelingher@iiasa.ac.at](mailto:zelingher@iiasa.ac.at)**Keywords:** food-security, agricultural commodities, price forecasting, agricultural production, interpretable machine learning**Abstract**

In this study, we investigate and forecast the impact of crop production shocks on the global prices of three major international agricultural commodities: maize, soybean, and cocoa. We perform a thorough assessment of the forecasting performances of five econometric and machine learning models using 60 years of data. First, we train the models on production and price data to forecast the monthly price variations for each crop separately considering different time horizons. Next, we implement a cross-validation procedure to identify the models with the most accurate forecasting ability for each crop. After choosing the best forecaster, we identify the most influential producing areas using several local and global model-agnostic interpretation tools. Our findings indicate significant differences among commodities in terms of prediction accuracy, with cocoa exhibiting a higher level of prediction error compared to less volatile markets like maize and soybean. Our results reveal a significant influence of Northern America's maize and soybean production on the global prices of these commodities. The effects of production on prices are asymmetrical: small decreases in US production lead to substantial price increases, while small increases in production do not systematically decrease prices. In contrast, cocoa price variations are influenced by production coming from several regions, not from a single one.

**1. Introduction**

This paper introduces a novel machine learning (ML) approach to forecast agricultural commodity (AC) prices and identify the most influential producing regions, addressing a critical gap in the existing literature and informing policies and strategies for global food security. Decades of AC market instability have posed persistent challenges, impacting farmers, traders, and policymakers worldwide. Influenced by diverse factors such as supply-demand dynamics, weather conditions, geopolitical events, and financial speculation, the fluctuation in prices, as documented by the Food and Agriculture Organization (FAO) dataset (FAO 2023), underscores a critical issue. Historical instances like the 2007–2008 food crisis, the 2020 COVID-19 pandemic and the Ukraine–Russia war demonstrate the profound implications of price shocks on food security, societal stability, and global

supply chains (Kalkuhl 2016, Schmidhuber *et al* 2020, Glauber and Laborde Debucquet 2023).

This study navigates the complex terrain of price dynamics, particularly focusing on three globally traded ACs, namely maize, soybean, and cocoa<sup>4</sup>, emphasizing their historical impact on food security, especially in developing nations (Headey 2011, Kalkuhl 2016). The COVID-19 pandemic and subsequent disruptions in 2020 revealed vulnerabilities in global food supply chains, leading to significant price spikes in ACs, including rice, wheat, and maize (Schmidhuber *et al* 2020). Even the cocoa market experienced dramatic price swings during the pandemic, driven by logistical challenges and supply chain disruptions (Organization 2021). More recently, the Russian–Ukrainian war has added

<sup>4</sup> for more information, see appendix A.

another layer of uncertainty and volatility to the global food market, as both countries are major producers and exporters of wheat and other grains (FAO 2023). The war has disrupted the Black Sea trade routes and caused prices to soar for staple foods that are essential for food security in many regions (Glauber and Laborde Debucquet 2023).

Price volatility in AC markets profoundly affects farmers, particularly smallholders who lack resources for effective price risk management (Wollni and Zeller 2007). Furthermore, price instability can strain food assistance programs, exacerbating food security challenges for vulnerable populations (Headey 2011). Anticipating price shocks is crucial for mitigating their impact on food security, and the literature emphasizes the importance of proactive measures, including effective risk management strategies for farmers and policies promoting market transparency (Wollni and Zeller 2007, Kalkuhl *et al* 2016).

ML, once considered complex, is now more easily accessible thanks to user-friendly methods and open-access databases (Henrique *et al* 2019, Zelingher and Makowski 2022), offering new opportunities for predicting commodity price variations. By using these methods and data to forecasting prices at a medium time horizon, our study thus addresses a critical gap in the existing literature.

This study evaluates five econometric and ML techniques to provide interpretable medium-term forecasts over different forecast horizons (1–12 months), with a focus on the accessibility of the methods and results to a wide audience. Notably, random forest (RF) and gradient boosting machine (GBM) emerge as top-performing models, showcasing significant variations in prediction accuracy across commodities. The study also unveils the distinctive influence of Northern American production changes, emphasizing their asymmetrical impact on maize and soybean prices.

In addition to model evaluation, the study explores the broader dynamics of agricultural commodity markets. It identifies influential producing regions, answering specific research questions about the regions or countries with the most substantial impact on selected commodity prices. The paper concludes with a nuanced discussion of findings, offering insights that extend beyond mere forecasting to inform policies and strategies in the intricate realm of global food markets.

The paper is structured as follows. Section 2 outlines the data used in the study. Section 3 presents the five models used in this study, including the open-source packages for their implementation. Furthermore, this section describes the method used to evaluate the forecasting performances of the models and the different model-agnostic techniques implemented to identify the most influential producing regions (or countries). Section 4 presents

the results for the three commodities. Section 5 discusses the findings and draws conclusions.

## 2. Data and methodology

### 2.1. Data

#### 2.1.1. Model output—monthly global price variation

We extracted global monthly price data for maize, soybean, and cocoa from the World Bank's commodity market database (World-Bank 2023) between January 1960 and December 2020 (732 values). To remove the effect of inflation, all three price time series were deflated into real 2010 USD values, using the agricultural price index of the corresponding period. Let us define  $p_{m,y}^n$  as a nominal price relative to a month  $m$  in a year  $y$ ,  $p_{m,y}^d$  as the deflated prices and  $In_{m,y}$  as the price index, both relative to the same period. Setting 2010 as the year of basis ( $In_{m,2010} \approx 100$ ) the deflation was as follows:

$$p_{m,y}^d = \frac{p_{m,y}^n \times In_{m,2010}}{In_{m,y}}. \quad (1)$$

Typically, globally traded crops are harvested in a main harvest season, according to the climate conditions of the area. Consequently, changes in production levels can potentially affect yearly price changes of that crop. On account of this, we define the dependent variable in the analysis as the proportion of price change relative to the same month  $m$  of the previous year such as

$$p_{m,y} = \frac{p_{m,y}^d - p_{m,y-1}^d}{p_{m,y-1}^d} \quad (2)$$

for any year  $y$ .

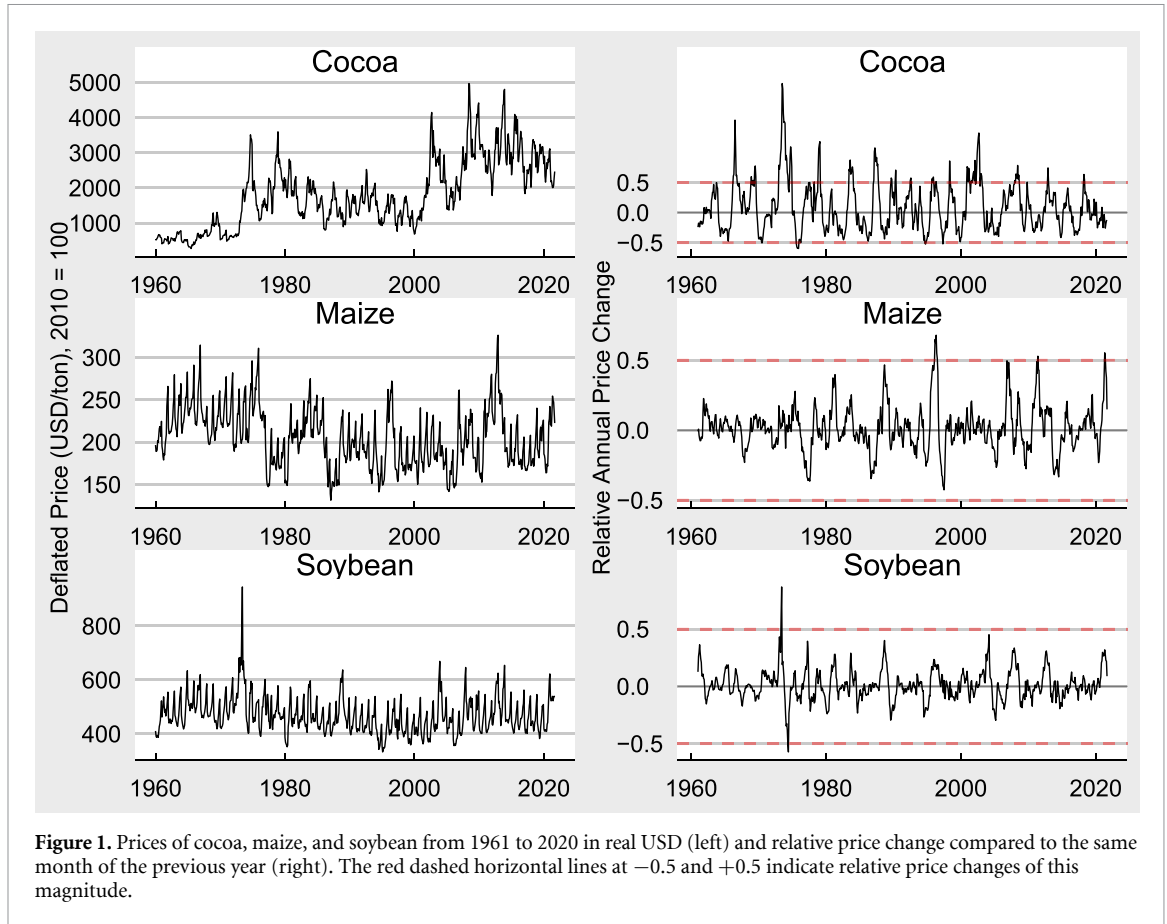
Figure 1(left) presents the global monthly prices of the three AC's over the past six decades in real 2010 USD.

#### 2.1.2. Model predictors

Our dataset comprises annual crop yields and production. This information was sourced from the FAOstat database (FAO 2023) (1960–2020) and encompassed two geographical scales: national (individual countries) and regional (clusters of countries defined by the FAO). To illustrate, a representative regional cluster is Western Africa, spanning 16 countries including Nigeria, Ghana, and Côte d'Ivoire. The adopted methodology involved (i) adhering to FAOstat's region delineations, (ii) calculating regional production as the summation of national yields multiplied by respective national cultivated areas, and (iii) determining regional yield by dividing regional production by regional cultivated area.

To continue along the same line as the relative price change function (equation (2)), we transformed





the national and regional yield and production data into relative annual changes, as follows:

$$x_{k,y} = \frac{q_{k,y} - q_{k,y-1}}{q_{k,y-1}} \quad (3)$$

where  $q_{k,y}$  is the production (or yield) in a geographical unit (country or region)  $k$  ( $k = 1, \dots, K$ ) and year  $y$ , and  $x_{k,y}$  is the relative production (or yield) change in the same year and area.

The values of  $x_{k,y}$  obtained for different regions were used as predictors to forecast global price changes  $p_{m,y}$ . Appendix B features tables 7–12, presenting an inclusive inventory of the production (or yield) locations for each crop along with accompanying descriptive statistics for production (and yield). Annual yield and production changes were computed at the national and regional scales, leading to four different sets of predictors. These sets were considered separately because of their strong correlations. Note that the price change obtained in a given year and a given month was forecast using yield or production data available prior to that month. To do this, we used local crop calendars (ITC and UNCTAD/WTO 2001, FAS-USDA 2023) to determine the harvest periods for each crop in all the countries and regions considered.

Figure 2 shows the production data for the leading producers of each of the considered commodities, in

terms of total quantity produced between 1961 and 2020.

Of the three commodities, cocoa is the least stable in terms of price fluctuations (figure 1) and soybean has the highest change in terms of regional production share in the global soybean production (figure 2). Despite several important shocks over the years, the maize market is stable relative to soybean and cocoa, both in terms of price fluctuations and production trends.

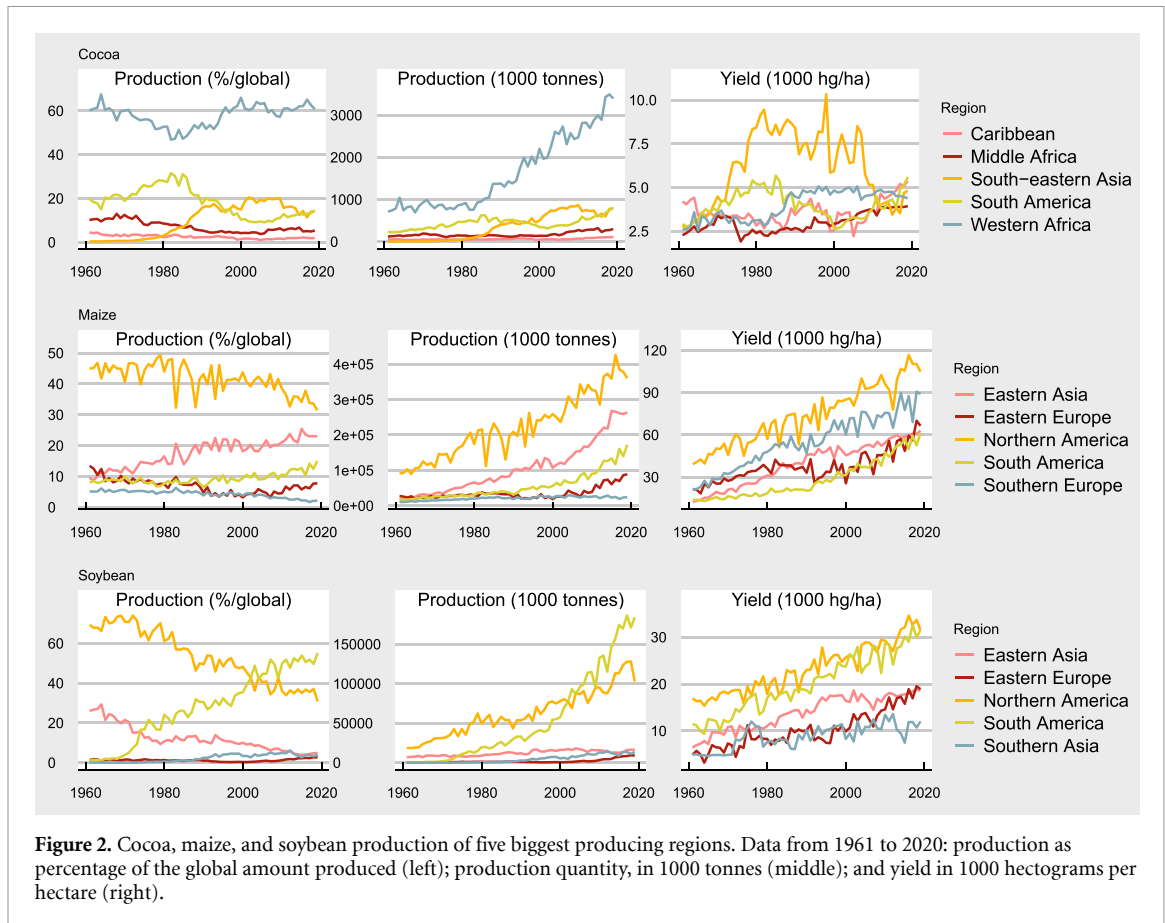
A more complete description of the data is supplied in table 5 and data sources are provided in table 6, both in appendix B.

In addition to yield and production changes, we considered the price changes observed the previous years ( $p_{m,y-1}, p_{m,y-2}, \dots$ ) as additional predictors.

## 2.2. Models

### 2.2.1. Machine-learning models

This study examines the forecasting accuracy of three ML models, namely CART, RF, and GBM, to predict price changes  $p_{m,y}$  for each month  $m$  as a function of  $x_{k,y}$ ,  $k = 1, \dots, K$ , using each set of predictors in turn (i.e. yield or production changes, at the national or regional scales). CART (Breiman *et al* 1984) builds an individual recursive tree. At each step, CART splits the data sample into two homogeneous groups relative to the highest ranked predictor, in a way that minimizes



the forecasting error. CART continues the splitting process until reaching the point where the data partition does not reduce the forecasting error. The resulting set of splitting rules defines a tree that can be used to predict the target variable, here price change. We fit CART using the `rpart` package (Therneau *et al* 2019) in a regression form, in which the algorithm aims to reduce the residual sum of squares.

Although CART has the advantage of being simple to interpret, it is considered as a ‘weak learner’. Its high sensitivity to minor changes in the data makes it a rather unstable, thus not always reliable forecaster. To overcome this problem, we test another two ML techniques, namely RF and GBM.

RF (Hastie *et al* 2009) is a bagging algorithm processing a collection of random sub-datasets, training a single decision tree with each one, and combining all trees derived from all sub-datasets. RF predictions are usually smoother and more stable than CART. Here, we applied RF to forecast price changes using the R package `randomForest` (Liaw and Wiener 2002). To maximize the forecast ability of the model, we tuned RF by testing 14 different numbers of trees (hyperparameter `nree`, whose values were set equal to 10, 20, 30, 40, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, successively). Additionally, we searched for the optimal number of variables to sample randomly at each split (hyperparameter `mtry`). Values of `nree`

and `mtry` were optimized to minimize the out-of-bag mean square error.

The third ML algorithm, GBM, relies on a boosting technique. As RF, GBM generates sub-datasets and fits trees to each of them but, unlike RF, the trees defined by GBM are not independent; each tree is designed to predict the errors of the previous tree. GBM was applied here using the `gbm` package (Greenwell *et al* 2020). We tuned two hyperparameters of GBM, namely the number of trees (same numbers as those considered for RF) and the maximal depth of each tree (1, 2, . . . , or 9). Both hyperparameters were optimized to minimize the out-of-bag mean square error.

### 2.2.2. Econometric models

Within the class of econometric models, we consider two types of models: a standard multivariate LM and the Trigonometric Seasonal Box Transformation with ARMA residuals Trend and Seasonal Components (TBATSS).

The model LM describes the impact of  $x_{k,y}$  on the  $p_{m,y}$  through linear relationships, as follows:

$$p_{m,y} = \alpha_0 + \alpha_1 p_{m,y-1} + \sum_{k=1}^K \beta_k x_{k,y} + \epsilon_{m,y} \quad (4)$$

where  $\alpha_0$  is the intercept;  $\alpha_1$  and  $\beta_k$  are regression parameters and  $\epsilon_{m,y}$  are the residuals, all relative to a month  $m$ . Note that this model includes two types of predictors; the price change in year  $y - 1$  and the production (or yield) changes in the  $K$  regions. A separate model was fitted for each set of predictors (yield vs. production, national vs. regional scale) and each month (one set of regression parameters was estimated for each type of predictors and each month). For each model, the final set of predictors is selected using a stepwise selection based on the Akaike (1974) information criterion (AIC).

TBATS (De Livera *et al* 2011) is a more sophisticated time-series model used here to forecast price changes from a combination of trend, seasonal, and residual components, such as:

$$p_{m,y} = \text{Tr}_{m,y} + S_{m,y} + \text{ARMA}_{m,y} \quad (5)$$

where  $\text{Tr}_{m,y}$  is the trend component (including a generalization of the Holt–Winters algorithm),  $S_{m,y}$  is the seasonal component based on Fourier series, and  $\text{ARMA}_{m,y}$  is the residual component expressed as an auto-regressive moving average (ARMA) model. The model can also use a Box–Cox transformation of the price data, if required. The TBATS model does not rely on production data but on past price data only. It is a powerful time series model for short-term forecasting (Kyriazi *et al* 2019, Chrulski 2021, Crespo Cuaresma *et al* 2021, Perone 2022). It includes an internal model selection procedure to automatically evaluate and select the best model among a large number of time series models (with/without seasonality, with/without trends, with/without Box–Cox transformation, with/without AR/MA components) based on the AIC. Here, we included the TBATS model as a benchmark in our evaluation to determine whether models using production data as predictors were able to outperform a powerful forecasting tool based solely on price data. TBATS was applied using the R package `forecast` (Hyndman *et al* 2020), and was used to predict price changes at a 1 to 12 month forecasting horizons.

### 2.3. Model training and selection

The process of model training and selection involved several crucial steps to build an effective predictive model. Firstly, the dataset was divided into a training set (consisting of the first  $T$  years of observations, with  $T \geq 44$ ) and a testing set (consisting of the next observation following the year  $T$ ). In our study, we aim to forecast the global prices of ACs for the year following the last year of price observation. Our evaluation method was chosen to reflect this type of real-life application. To achieve this, we employ a rolling cross-validation approach (Hyndman and Athanasopoulos 2018) where we iteratively expand the training set by adding one year of data in each iteration. This approach allows us to forecast the price of

the following year using the data available up to the previous year. By selecting  $T = 44$  as the minimum training period, we ensure that our training sets encompass a substantial historical data range, providing a robust basis for model training. As we progress through each iteration, the training set expands (i.e. it includes  $T = 45, 46$ , etc observations), capturing the temporal dynamics and fluctuations in the global food market over time. All models were trained using different combinations of hyperparameters as explained in the previous section, resulting in several variants for each model type. All model variants were then used to predict the observation (price change) available at the year  $T + 1$ . This process is repeated using rolling cross-validation in order to predict all price changes (for the years  $> 44$ ), with each model variant in turn.

The performances of the different models were evaluated month by month by comparing the RMSE of each model to the standard deviation of the observed price changes over the same time period. This comparison was done by computing a relative advantage (RA) index for each model, defined as

$$RA_m = 1 - \frac{\text{RMSE}_m}{\text{SD}(p_{m,y})} \quad (6)$$

<sup>5</sup>Finally, mean absolute error (MAE) is also used as an additional evaluation criterion to check the robustness of the model selection. Further details and explanations regarding the model comparison and the rationale behind it are available in appendix A.1.

In the last phase of our analysis, we used a one-way ANOVA and  $t$ -test to determine if there are any significant differences between the commodities in terms of mean values of RMSE and RA. To compare the performance of different forecasting methods, we select the top three forecasters that exhibit the highest accuracy among 16 possibilities ((3 ML models + 1 LM model)  $\times$  2 types of inputs (production/yield)  $\times$  2 geographic scales (regional/national)).

### 2.4. Model interpretation

Understanding the origins of the variability of the model predictions is essential to increase the confidence of the model users (Spavound and Kourentzes 2022). We implemented several model-agnostic techniques (Molnar 2022) that provide detailed and straightforward explanations of the model outcomes. These techniques were implemented with the most

<sup>5</sup> By definition, the value of  $\text{SD}(p_{m,y})$  is equal to the RMSE of the historical data average. Thus, a value of  $RA$  higher than one indicates that the model considered performs better (i.e. has a lower RMSE) than the historical average. On the opposite, a value equal to or lower than one indicates that the model is not more accurate or is even less accurate than a constant prediction equal to the historical average.

accurate models selected at each month<sup>6</sup>. The selected models were trained again using the whole set of observations, and several global and local agnostic methods were then implemented to shed light on the nature of the forecasting results and the underlying mechanism.

First, we analyzed and ranked the relative importance of the predictors for predicting price changes (Fisher *et al* 2018). We assessed the importance of each  $x_{k,y}$ ,  $k = 1, \dots, K$  by calculating its contribution to the prediction errors of the model using a permutation method. More precisely, we computed the increase of mean square errors resulting from a random permutation of each predictor in turn using the *FeatureImp* R function. A large increase of mean square errors indicates that the corresponding predictor is important. A low or absence of increase indicates that the predictor is not important or even useless. The results of this analysis are reported in relative importance plots for each crop in figures 4, 6, and 8. On these plots, the ranking of the predictors is indicated by the locations of the predictors along the  $x$ -axis and by their colors, with 1 (dark purple) representing maximum importance and 0 (bright yellow) representing no importance to the model. Error bars are used to 95% confidence intervals of importance values of the predictors.

Second, after measuring the impact of each predictor on the average prediction errors, we assessed the contribution of each of the  $K$  predictors to each individual price change prediction. To achieve this, we implemented the game-theory approach, Shapley (Shapley 1952), which measures the contribution of a given model predictor to a given model prediction. Here, Shapley values describe the contribution of the regional/national yields/production to predicted price variations. These values can be used to identifying the origins of the most extreme predicted price variations corresponding to major price shocks. Shapley values provide information on both the directions and the magnitudes of the effects of the predictors. When plotted as a function of the values of  $x_{k,y}$ , they provide a visualization tool to assess the risk of price shocks as a function of the levels of variation in regional production. The average absolute Shapley values (computed by region and country) can also serve for ranking the producing regions or countries according to their influence on predicted price variations. Here, this approach was implemented with the *iml* R package (Molnar *et al* 2018). The robustness of the results obtained with the Shapley values is assessed using an alternative approach based on local models (LIME) implemented with the *iml* R package (Molnar *et al* 2018).

<sup>6</sup> TBATS is not concerned by the interpretation stage as it does not take crop production into account and is considered to be a black box.

## 3. Results

### 3.1. Accuracy of model forecasts

Table 1 presents the best forecasting options for different months and different crops. The names of the ML models correspond to the models showing the highest RA for predicting price change at each month. The name between brackets indicates the type of input, which was found to have the most substantial impact on price. The numerical values represent the values of RA of the best ML models ( $RA_{ML}$ ) and of TBATS for the lags (months).

For maize, the one-month-ahead predictions made with TBATS tend to be higher in accuracy than those obtained using ML techniques. The best ML tool (GBM with national production inputs for predicting price change in May) achieved a RA of 60%, while the highest RA for TBATS was 80% in March, when considering one-month-ahead forecasts. The good performances of TBATS are not sustained for extended time-horizon forecasts. Indeed, the RA of TBATS decreases rapidly with each increment of time horizon (table 1). Thus, when considering longer term projections, the model comparison no longer favors TBATS and, compared to other alternatives, GBM shows higher RA levels when using regional production as predictors. In general, for maize, production-based forecasts tend to be more accurate than yield-based forecasts. In comparison to GBM, the models CART and LM tend to have lower RA values, especially when using regional predictors.

Results obtained for soybean, as shown in table 1, are rather different. First, the average performance of TBATS was lower, with average RA values of 58% for one-month-ahead forecasts and 47% for two-months-ahead forecasts. Second, there was a relatively high variability in the forecasting accuracy of the ML methods. Here, the highest RA reached 90% in March with GBM and regional yields as predictors, but was only equal to 29% in August. Overall, for soybean, higher RA levels were achieved during the first six months of the year, especially for price changes in February ( $RA_2 = 87\%$ ), March and April ( $RA_4 = 78\%$ ). Lastly, while GBM performed systematically better for maize, the RF method performed better in forecasting soybean price variations for four months of the year only.

For cocoa, TBATS has an average RA of 62% for a one-month-ahead forecasting period and 52% for a two-month-ahead forecasting horizon. On the other hand, the use of machine learning techniques, particularly RF and GBM, shows more accurate price forecasts, reaching a maximum value of 76% for the August price change (regional yields as predictors). The results also highlight the weakness of the linear regression model, especially when relying on national predictors.

**Table 1.** Relative advantage (RA) of the best machine learning models and TBATS for different months and different crops. The names and values reported for each month correspond to the models showing the highest RA for predicting price change at this period. NA's are reported for TBATS models with RAs values lower than the RA value of the best machine learning model.

a. Maize		ML model (input)	TBATS (by lag)				
Month		RA <sub>ML</sub>	1	2	3	4	5
1	GBM (regional production)	0.57	0.74	0.71	0.61	NA	NA
2	GBM (regional production)	0.50	0.73	0.67	0.64	0.56	NA
3	GBM (regional production)	0.50	0.80	0.70	0.64	0.60	0.54
4	GBM (countries' production)	0.54	0.59	NA	NA	NA	NA
5	GBM (countries' production)	0.60	0.70	0.68	0.62	0.55	0.51
6	GBM (countries' production)	0.48	0.58	0.56	0.50	NA	NA
7	GBM (regional production)	0.46	NA	NA	NA	NA	NA
8	GBM (regional production)	0.57	NA	NA	NA	NA	NA
9	GBM (regional yield)	0.47	0.54	NA	NA	NA	NA
10	GBM (regional production)	0.58	0.60	NA	NA	NA	NA
11	GBM (regional production)	0.45	0.56	NA	NA	NA	NA
12	GBM (regional production)	0.52	0.70	0.62	0.53	NA	NA
b. Soybean		ML model (input)	TBATS (by lag)				
Month		RA <sub>ML</sub>	1	2	3	4	5
1	GBM (countries' yield)	0.85	NA	NA	NA	NA	NA
2	GBM (regional yield)	0.87	NA	NA	NA	NA	NA
3	GBM (regional yield)	0.90	NA	NA	NA	NA	NA
4	RF (countries' production)	0.78	NA	NA	NA	NA	NA
5	RF (countries' yield)	0.74	NA	NA	NA	NA	NA
6	GBM (countries' yield)	0.87	NA	NA	NA	NA	NA
7	RF (regional yield)	0.33	0.52	0.42	0.36	NA	NA
8	RF (countries' yield)	0.44	0.54	NA	NA	NA	NA
9	RF (countries' yield)	0.44	0.47	NA	NA	NA	NA
10	GBM (regional yield)	0.61	NA	NA	NA	NA	NA
11	GBM (countries' yield)	0.65	0.66	NA	NA	NA	NA
12	GBM (countries' yield)	0.68	0.76	NA	NA	NA	NA
c. Cocoa		ML model (input)	TBATS (by lag)				
Month		RA <sub>ML</sub>	1	2	3	4	5
1	RF (countries' production)	0.66	NA	NA	NA	NA	NA
2	RF (countries' yield)	0.68	NA	NA	NA	NA	NA
3	RF (countries' production)	0.61	NA	NA	NA	NA	NA
4	RF (countries' production)	0.55	0.75	0.57	NA	NA	NA
5	RF (countries' production)	0.58	0.75	0.68	NA	NA	NA
6	GBM (countries' yield)	0.65	NA	NA	NA	NA	NA
7	RF (regional yield)	0.60	0.63	0.62	NA	NA	NA
8	RF (regional yield)	0.76	NA	NA	NA	NA	NA
9	RF (countries' yield)	0.58	NA	NA	NA	NA	NA
10	GBM (countries' production)	0.73	NA	NA	NA	NA	NA
11	GBM (countries' production)	0.62	NA	NA	NA	NA	NA
12	RF (countries' production)	0.68	NA	NA	NA	NA	NA

### 3.2. Differences in performances between commodities

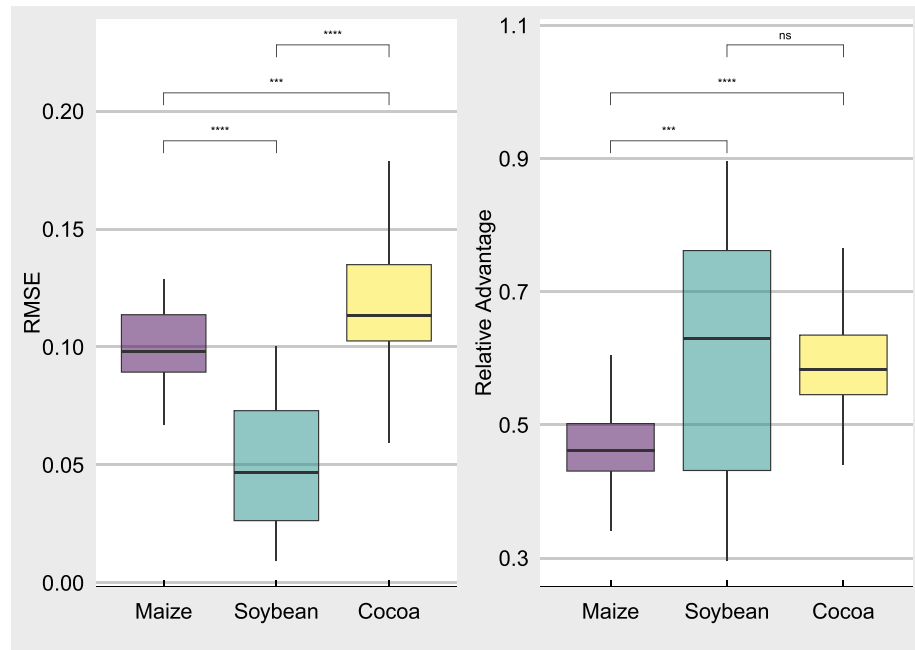
The differences in RMSE and RA between commodities are evaluated in figure 3. Regarding the RA, all three commodities have a mean RA greater than 0.5, indicating an overall advantage of model-based price forecasting compared to constant predictions. Soybean has the highest average RA (0.68), suggesting a stronger benefit of model-based forecasting for this commodity. The mean RA of cocoa is very similar (0.64). In terms of RMSE, soybean shows the

lowest mean RMSE (0.04) and cocoa has the highest mean (0.10). The three RMSE values are significantly different ( $p < 0.05$ ) according on a two-sided  $t$ -test.

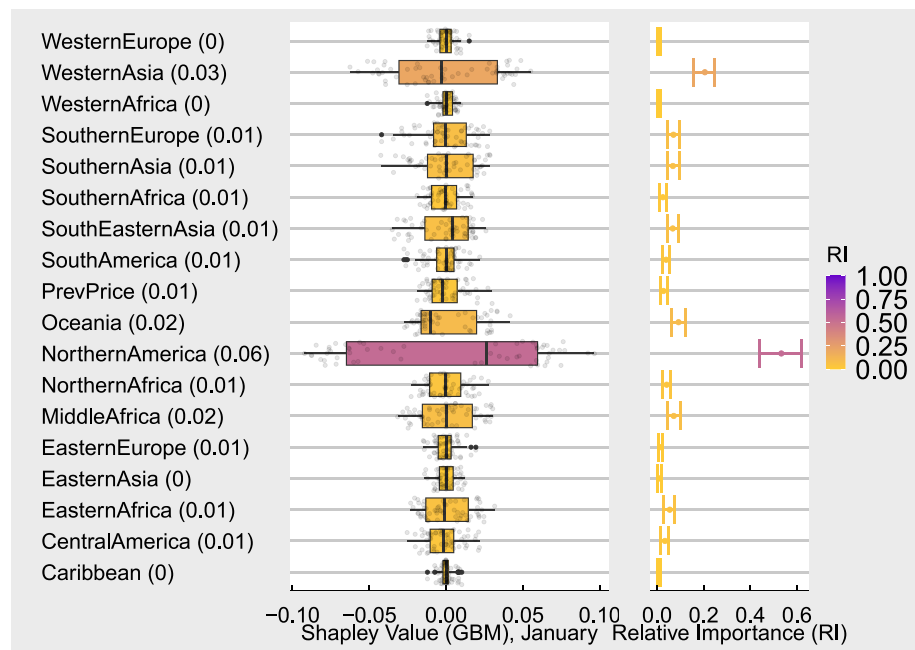
### 3.3. Effect of crop production on prices

In this section, we present the results of the analysis of the importance of the predictors on the prediction accuracy of the models (figures 4, 6, and 8). Next, we show the partial dependence of price changes on the most important predictors (figures 5, 7, and 9).

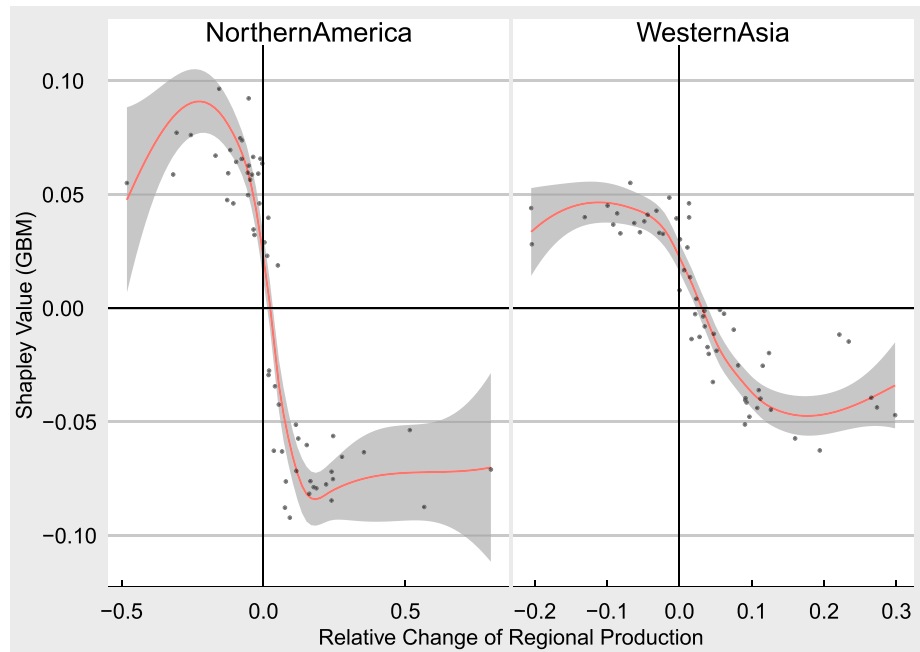




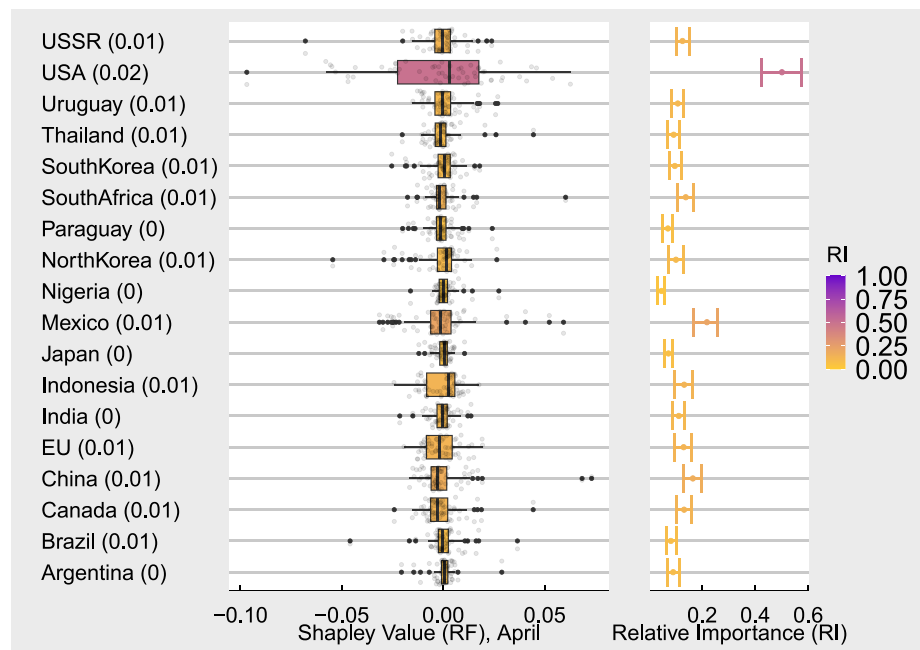
**Figure 3.** Model performances for maize, soybean, and cocoa according to the RMSE (left) and relative advantage (RA = 1 – RMSE/Standard deviation of the price data) of model-based predictions versus constant predictions (right). The distributions of RMSE and RA across forecasting scenarios are presented in box plots, and the results of statistical tests comparing the mean values of RMSE and RA are presented in black. Signif. codes: ns :  $p > 0.05$ , not – significant; \* :  $p \leq 0.05$ ; \*\* :  $p \leq 0.01$ ; \*\*\* :  $p \leq 0.001$ ; \*\*\*\* :  $p \leq 0.0001$ .



**Figure 4.** Assessment of the importance of regional production changes for maize price forecasting in January with the GBM model, using regional production changes as predictors. The left-side box-plot displays the distribution of Shapley values, indicating the impact of each predictor (regional production change) on the model output (price change). The gray points within each box represent the individual Shapley values attributed to specific predictor variables. The right-side feature importance plot illustrates the contribution of each predictor to the RMSE resulting from a random permutation, with dark purple indicating highly influential predictors and orange representing low-impact features.

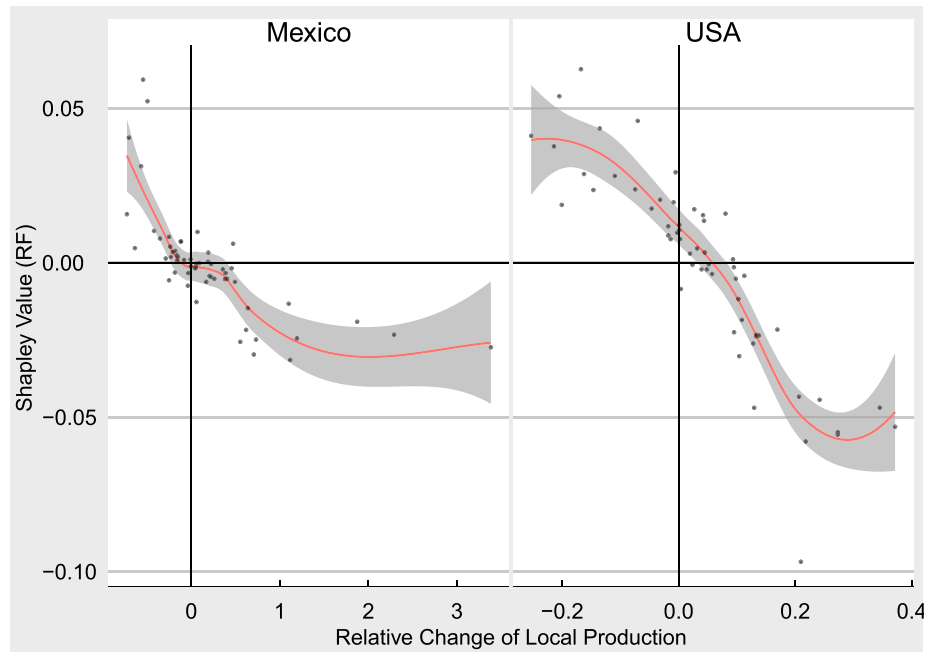


**Figure 5.** Shapley-based partial dependence plot (PDP) for maize. The black points scattered along the x-axis represent individual feature values, while their corresponding Shapley values are depicted on the y-axis. The smoothing curve derived from these points (in red) forms the PDP, which offers insights into the connection between relative production changes in Northern America (on the left) and Western Asia (on the right) and the projected relative maize price changes in January. The predictions are based on the GBM model, chosen for its high forecasting accuracy. The gray bands indicate the 95% confidence intervals.

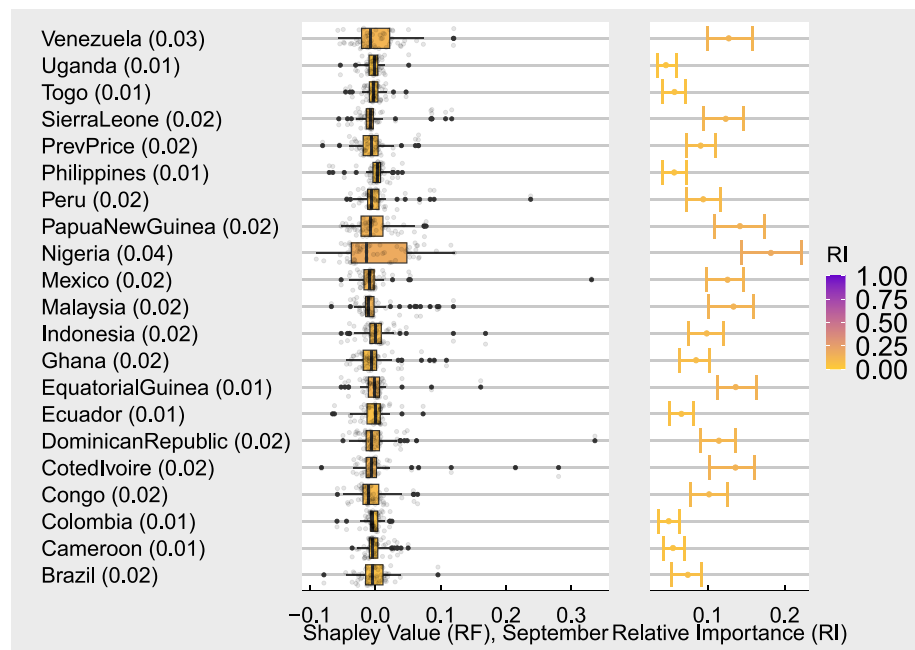


**Figure 6.** Assessment of the importance of national production changes for soybean price forecasting in April with the RF model using national production changes as predictors. The left-side box-plot displays the distribution of Shapley values, indicating the impact of each predictor (country’s production change) on the model output (price change). The gray points within each box represent the individual Shapley values attributed to specific predictor variables. The right-side feature importance plot illustrates the contribution of each predictor to the RMSE resulting from a random permutation, with dark purple indicating highly influential features and orange representing low-impact predictors.

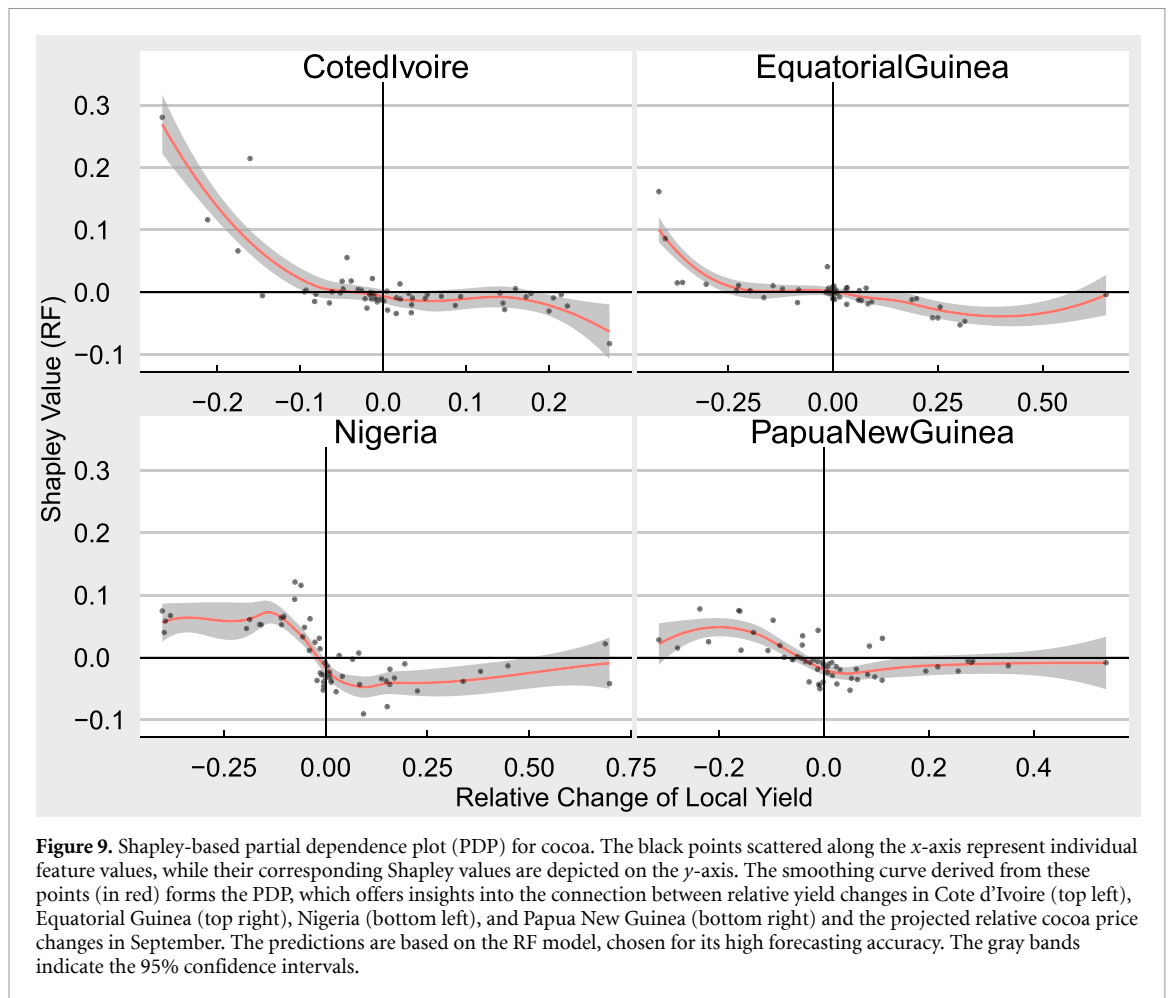




**Figure 7.** Shapley-based partial dependence plot (PDP) for soybean. The black points scattered along the x-axis represent individual feature values, while their corresponding Shapley values are depicted on the y-axis. The smoothing curve derived from these points (in red) forms the PDP, which offers insights into the connection between relative production changes in Mexico (on the left) and the USA (on the right) and the projected relative soybean price changes in April. The predictions are based on the RF model, chosen for its high forecasting accuracy. The gray bands indicate the 95% confidence intervals.



**Figure 8.** Assessment of the importance of national yield changes for cocoa price forecasting in September with the RF model using national yield changes as predictors. The left-side box-plot displays the distribution of Shapley values, indicating the impact of each predictor (national yield change) on the model output (price change). The gray points within each box represent the individual Shapley values attributed to specific predictor variables. The right-side feature importance plot illustrates the contribution of each feature to the RMSE resulting from a random permutation, with dark purple indicating highly influential features and orange representing low-impact predictors.



### 3.3.1. Maize

The model with the highest level of accuracy for predicting changes in January's maize price is GBM, using regional production data as predictors. Figure 4 indicates the most and least influential variables, and reveals that the production changes in Northern America have the strongest influence on price change predictions (mean absolute Shapley value = 0.062, relative importance = 0.53). The next most influential region, with a significant influence gap, is Western Asia (mean absolute Shapley value = 0.03, relative importance = 0.25). The influence of the other variables is considerably lower, and often close to zero. Similar results were obtained for most of the year. The variables are colored according to their level of relative importance<sup>7</sup> Overall, these findings underscore the significant impact of Northern America on world maize prices and highlight the

importance of considering regional factors when analyzing commodity price dynamics.

Figure 5 focuses on the two predictors with the highest relative importance: Northern America (left) and Western Asia (right). The partial dependence plots (PDPs) displayed on this figure reveal that production changes in Northern America have a stronger impact on world maize prices compared to Western Asia. The PDPs further indicated that an increase (decrease) of production in North America and Western Asia tends to decrease (increase) the maize price. The Shapley values computed for Western Asia are relatively evenly distributed along the regression line, with respect to the  $Y$ -axis, indicating a more consistent marginal effect. In contrast, the values obtained for Northern America are more extreme, with a higher density toward the top and bottom of the chart. The effects of an increase and decrease in production are asymmetrical. Small decreases in US production lead to substantial price increases, while small increases in production tend to not always decrease maize prices. An increase of US production needs to reach a certain level before inducing a decline of price. Inherent uncertainty arises when examining the Shapley value trends for the most extreme production changes in Northern America.

<sup>7</sup> The results obtained when using LIME tend to be more extreme. It seems that LIME, due to its simplicity and being based on a linear model, yields unstable results in the presence of extreme values. However, we find LIME very useful for understanding the model, and therefore we choose to use it, but place the results of the analysis in the appendices (See figure 11).

The limited number of observations below  $-45\%$  of production change induces wider confidence intervals, while greater confidence is attributed to the predicted outcomes for production changes within the  $-30\%$  to  $+30\%$  range due to a larger number of data.

These findings highlight the importance of considering both the magnitude and distribution of marginal effects when interpreting the results of the models. Furthermore, the results suggest that regional factors, such as production changes in Northern America and Western Asia, play a critical role in shaping global commodity prices.

### 3.3.2. Soybean

For soybean price variation in April, the most reliable predictions were made through RF, with countries' production as input. Figure 6 shows that the USA is the most influential producing country for soybean price changes, similar to maize price changes, with a negative correlation between production and soybean price changes. The relative influence of the other explanatory variables was lower and often closer to zero, as also reflected in figure 12. Nevertheless, some of the results obtained for soybean are different from those obtained for maize, such as the lower Shapley values obtained for soybean production in the USA compared to those for maize. The role of the USA in driving positive changes in the global soybean prices was found to be weaker compared to its effect in the maize market, as revealed by the lower Shapley values obtained for Soybean.

Figure 7 shows the partial dependence between price change and the two most influential predictors, namely soybean production in Mexico, on the left, and the USA, on the right. Both PDPs reveal a negative correlation with soybean price variations in April, with the primary effect occurring during the narrow transition phase from slightly negative to positive production changes. Production variations in the USA have a stronger effect on price change in April than production variation in Mexico. Interestingly, small increases in soybean production in the USA do not systematically lead to lower prices. Thus, a price decrease is achieved only for annual production increases close to  $+5\%$  in the USA.

### 3.3.3. Cocoa

Compared with the results obtained for maize and soybean, the results of the cocoa model appear to be less conclusive. The most precise predictions were usually obtained with the RF model using national yield changes as predictors. For cocoa, prices seem to be influenced by several countries. Surprisingly, Cote d'Ivoire, the main cocoa producer (accounting for approximately 31% of the total cocoa production since 1960), is not ranked first according to its relative

importance<sup>8</sup>. Figure 8 indicates that several countries have similar importance, such as Nigeria (0.18, [0.14, 0.22]), Papua New Guinea (0.14, [0.11,0.17]) and Cote d'Ivoire (0.14, [0.10,0.16]). To further elucidate the relationship between yield and cocoa price changes, we examine PDPs for the four countries with the highest relative importance. As depicted in figure 9, these countries display a weak-negative relationship between yield variation and price variation. Overall, our results demonstrate the inherent price variability within the cocoa market and the significant impact of domestic yield changes on international prices.

## 4. Discussion

This study introduces an innovative methodology for medium-term AC price analysis and forecasting, employing ML tools and data in open-access. We show the practical interest of this methodological framework through three case studies on maize, soybean and cocoa. Our findings highlight the importance of a rigorous model selection procedure, considering specific commodities, trading periods, and market characteristics. In particular, our results highlight the advantage of considering multiple forecasting methods in order to find the most accurate one for a given crop and forecast period (Kourentzes *et al* 2019, Wang *et al* 2022). The study underscores the superiority of ML over linear regression models, emphasizing the need for modeling complex nonlinear relationships. Nevertheless, distinct results emerge across the three commodities considered: TBATS and GBM excel for maize, RF performs better for cocoa, especially with national production data, while different ML methods provide good results for soybean depending on the month of forecast. Clearly, our results show that the best model should be selected on a case-by-case basis and that it is unlikely that a single model will be the best in all situations.

Our results also reveal that the model accuracy depends on the commodity, as model performance was found to be lower for cocoa due to its higher price instability compared to maize and soybean. More precise results were obtained for maize and soybeans, whose price variations were found to be closely linked to the annual variation in agricultural production among one or two major producers.

While acknowledging the impact of AC storage on price dynamics (Wright 2009, von Braun and Torero 2009, Bobenrieth *et al* 2013), we focused on production and yield data due to practical challenges

<sup>8</sup> Nigeria is ranked significantly high to predict September prices, which corresponds to the beginning of the cocoa trade year. Brazil is the only country who's cocoa trade year starts in May.

in obtaining comprehensive stocks data for a large number of countries. In this paper, we compared a wide range of methods, both parametric and non-parametric, based on statistical models and MLs. In the future, it will be interesting to use additional methods, for example more flexible regression models such as generalized additive models that are able to cope with changing trends over time, or such as multi-layer neural networks that offer great flexibility. However, these methods are data-intensive, and it is not certain that they will give good results with relatively short time series. It is worth noting that our methodological framework could be easily adapted to implement and compare other forecasting techniques.

Acknowledging data constraints and local market dynamics, our methodology could benefit from alternative data sources like satellite-derived vegetation indices (Anderson *et al* 2023). Further research should address data limitations and regional market dynamics to enhance reliability. Overall, our study provides an accessible toolbox for analyzing and forecasting agricultural commodity prices in the medium term, aimed at non-specialist users. Despite its limitations, this methodology could contribute to informed decision-making in global food markets.

## 5. Conclusion

Food security is a complex issue that refers to the state of having reliable access to a sufficient quantity of nutritious and affordable food to meet one's dietary needs. It encompasses the four dimensions of food availability, access, utilization, and stability. Access to food is a crucial aspect of food security, often determined by the relationship between the consumers' incomes and the price they pay for food. While food is considered a fundamental right, disparities in access to food still exist, particularly in low-income countries (FAO 2018). However, governments can mitigate the instability of food prices in their country by monitoring global markets, and farmers can plan their harvest and maximize profits by using existing models. According to Gilbert and Morgan (2010), many governments try to stabilize food prices in their countries. Yet, understanding and analyzing global food markets require financial knowledge and economic means that may not be available to all stakeholders (Olofsson 2020).

To contribute in the development of accurate and accessible price forecasting tools, we designed a methodological framework combining simple ML models and open-access databases. We showed that, when rigorously trained, these models could lead to accurate commodity price forecasting and allow

decision-makers to understand the origins of price fluctuations.

By implementing the proposed framework, stakeholders can gain insights into global food markets, enabling them to make informed decisions and take actions to promote food security, based on public data and software.

## Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

## Acknowledgments

The authors thank Nikolay Khabarov, Anna Shchiptsova, Sebastian Poledna and the Advancing Systems Analysis Program at IIASA for fruitful and valuable input. The authors thank Ansa Heyl for her assistance in proofing the manuscript.

## Ethical statement

The project participants respect and follow good research practices<sup>9</sup>. The project does not give rise to any potential ethical, safety-related, or regulatory issues<sup>10</sup>.

## Funding

The authors gratefully acknowledge funding from IIASA along with the Israeli Ministry of Science, Technology and Space, and the other National Member Organizations that support IIASA. This work was partly funded by the CLAND Institute of convergence in France (16-CONV-0003).

## Appendix A. Further information on models

### A.1. Commodities in focus

We focus our study on three globally traded ACs, namely maize, soybean, and cocoa. Maize is a crucial AC, used as bio-energy, feed, and food, both in developed and developing countries (FAO 2023). The USA is the largest maize producer, responsible for over 30% of the global supply.

Soybean is the most traded tropical grain worldwide and is produced in over 100 countries (De Maria *et al* 2020). While the USA was responsible for 70% of

<sup>9</sup> For example, see [The European Code of Conduct for Research Integrity](#).

<sup>10</sup> According to [ETHICS ISSUES table—CHECKLIST of H2020 programme](#) and [EU's Ethics for Researchers](#).

the global soybean market share in 1961, its share has since decreased to less than 30%.

Cocoa is mainly traded by major importers in New York and London, far from its countries of origin (ITC and UNCTAD/WTO 2001). Western Africa and South America are the primary regions for cocoa production, with smallholder farmers in family farms typically producing it (ITC and UNCTAD/WTO 2001). The prices received by these cocoa producers are closely tied to the international market price, a factor that significantly influences their decisions regarding land allocation and crop choices. The fluctuations in international cocoa prices can directly impact the livelihoods and economic prospects of these farmers, influencing their choices between cultivating cocoa and other crops such as coffee trees (Gilbert 2016). By including cocoa in this study, along maize and soybean, we aim to shed light on the multifaceted impact of production shocks on diverse ACs, ranging from those with direct nutritional implications to those with intricate socio-economic ramifications.

### A.2. Model comparison and the rationale behind it

The model comparison process is grounded in the RA metric and RMSE. RA, calculated as 1 minus RMSE divided by the data's standard deviation, serves as a normalized RMSE and equivalent to a standard model skill score. RMSE, a widely adopted measure of forecasting accuracy for quantitative predictions, combines bias and variance errors into a unified metric. While RMSE remains the prevalent evaluation criterion, alternative approaches could be valuable in specific scenarios.

This criterion measures the RA of the model compared to hypothetical RMSE of a constant forecast equal to the mean price change. The mean value can be seen as a naïve forecast as it assumes that the future value are equal to the mean of the past observations. In addition to this naïve forecast, we considered standard linear models owing to their prominence in statistical analysis, along with TBATS, as univariate time series models commonly applied in commodity and agricultural price prediction.

Through application to three major ACs-maize, soybean, and cocoa-this methodology's practical significance is demonstrated.

### A.3. Predictive performance according to MAE criteria

We present three tables showing the MAE of the selected models for predicting the price changes of

**Table 2.** Maize.

Month	ML model (input)	MAE <sub>ml</sub>
1	GBM (regional production)	0.05
2	GBM (regional production)	0.06
3	GBM (regional production)	0.06
4	GBM (countries' production)	0.05
5	GBM (countries' production)	0.04
6	GBM (countries' production)	0.08
7	RF (regional production)	0.08
8	GBM (regional production)	0.06
9	GBM (regional yield)	0.07
10	GBM (regional production)	0.06
11	RF (regional production)	0.07
12	GBM (regional production)	0.06

**Table 3.** Soybean.

Month	ML model (input)	MAE <sub>ml</sub>
1	GBM (countries' yield)	0.02
2	GBM (regional yield)	0.01
3	GBM (regional yield)	0.00
4	RF (countries' production)	0.01
5	RF (countries' yield)	0.01
6	GBM (countries' yield)	0.02
7	RF (regional yield)	0.05
8	RF (countries' yield)	0.05
9	RF (countries' production)	0.06
10	GBM (regional yield)	0.03
11	GBM (countries' yield)	0.04
12	GBM (countries' yield)	0.04

**Table 4.** Cocoa.

Month	ML model (input)	MAE <sub>ml</sub>
1	RF (countries' production)	0.06
2	RF (countries' yield)	0.05
3	RF (countries' production)	0.07
4	RF (countries' production)	0.11
5	RF (countries' production)	0.09
6	GBM (countries' yield)	0.07
7	RF (regional yield)	0.07
8	RF (regional yield)	0.04
9	RF (countries' yield)	0.08
10	GBM (countries' production)	0.04
11	GBM (countries' production)	0.07
12	RF (countries' production)	0.08

maize, soybean, and cocoa relative for each month. MAE is a robust metric for assessing the accuracy of predictions. The MAE is calculated as follows:

$$\text{MAE}_m = \frac{1}{T} \sum_{y=45}^T |\hat{p}_{m,y} - p_{m,y}|. \quad (7)$$

## Appendix B. Data and variables

**Table 5.** List of variables and indices used in the paper.

Symbol	Values	Description
Raw data		
$p_{m,j}^d$		Time series of observed monthly prices, deflated
$q_{k,j}$		Time series of observed annual agricultural output
Variables, in relative annual change		
$p$	$P = (p_1, p_2, \dots, p_T)$	Model input, observations in training set.
$x$	$X_k = (x_{t,1}, x_{t,2}, \dots, x_{t,K})$ $X_t = (x_{1,k}, x_{2,k}, \dots, x_{T,k})$	Model input, observations in training set (no TBATS)
$\hat{p}_y^f$	—	Model output, forecasted by the model
$\hat{p}_{d^{tb}}$	—	Model output, price forecasted by TBATS
$p_y$	—	Price to forecast
Indices		
$p_d$	—	Price to forecast, using TBATS
$j$	$j = 1960, 1961, \dots, J$	Years observed ( $j_1 = 1960$ )
$y$	$y = 1, 2, \dots, Y$	Years observed, transformed to relative change ( $y \geq \frac{j_1 - j_0}{j_0}$ )
$k$	$k = 1, 2, \dots, K$	Number of features in model
$m$	$m = 1, 2, \dots, 12$	Month, fixed, except for TBATS
$d$	$d = [y_1, 1], [y_1, 2], \dots, D$	TBATS Date, composed of $[y, m]$ , $D = [Y, m]$
$h$	$h = 1, 2, \dots, H$	Lag/Forecasting horizon ( $1 \geq H \geq 12$ ), in monthly units
$t$	$t = 1, 2, \dots, T$	Observations in training set
$t^{tb}$	$t^{tb} = 1, 2, \dots, T^{tb}$	TBATS training set
$y^f$	$y_{45} \geq y^f \geq Y + 1$	An instance (year) in testing set, one-step ahead forecast ( $f$ )
$d^{tb}$	$d_{30} \geq d^{tb} \geq D + H$	An instance (date) of price to forecast with TBATS ( $tb$ )

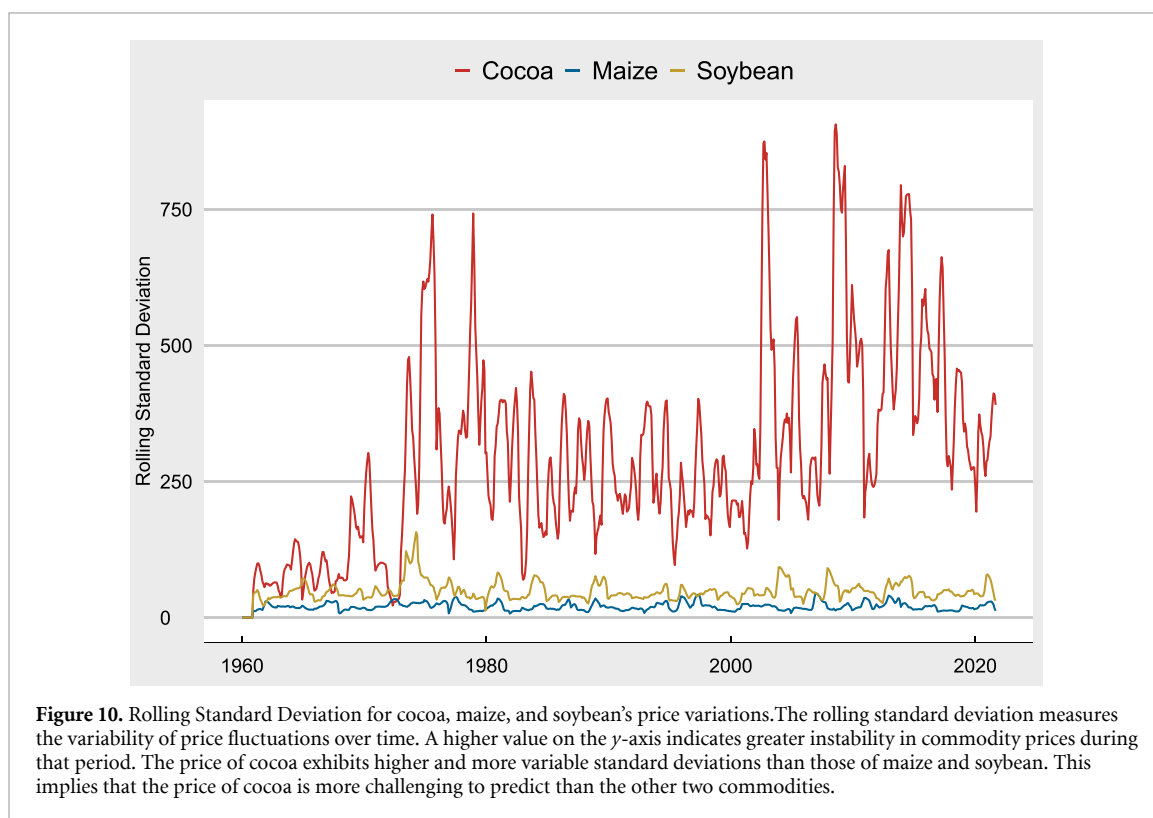
**Table 6.** Variable description and data sources.

Data	Unites	Time-range	Source
Final data			
Production	% change/year	1962 – 2019	
Yield	% change/year	1962 – 2019	
Price	% change/year	01/1961 – 11/2020	
Initial information			
Price	Nominal USD/mt <sup>a</sup>	01/1960 – 11/2020	World Bank,
Price index, Agriculture	USD (2010 = 100)	01/1960 – 11/2020	Pink Sheet (2023)
Production	tonnes/year	1961 – 2019	FAO STAT
Yield	hg/ha	1961 – 2019	(2023)

<sup>a</sup> Cocoa prices are given by kg. They were manually converted to units of metric tonnes



## Rolling standard deviation



## Maize

**Table 7.** Annual production and yield of maize, relative to region.

Area	Production (1000 tonne)			Yield (hg/ha)		
	Average	Min.	Max	Average	Min.	Max.
Caribbean	460	280	816	11 153	8536	16 556
Central America	18 051	7469	32 300	20 404	9738	36 458
Central Asia	1287	454	2275	48 623	25 634	69 552
Eastern Africa	14 897	5690	34 121	13 682	9523	20 753
Eastern Asia	106 304	17 712	267 523	39 413	12 276	62 944
Eastern Europe	34 855	18 752	88 839	37 404	18 402	69 972
Middle Africa	2674	1144	7645	8542	6741	10 903
Northern Africa	4982	1762	8688	41 231	15 914	69 852
Northern America	222 359	89 846	426 151	73 601	39 229	116 691
Northern Europe	41	1	203	35 396	10 000	75 841
Oceania	426	146	742	50 793	17 335	87 976
South-eastern Asia	19 551	4712	52 321	22 239	9017	46 328
South America	55 077	16 314	171 819	28 285	12 947	61 406
Southern Africa	9260	3445	17 891	24 027	7878	54 049
Southern Asia	16 435	6421	43 697	17 540	10 024	35 815
Southern Europe	20 903	10 388	29 670	54 903	21 126	90 657
Western Africa	8297	1803	26 043	12 141	6964	19 225
Western Asia	3140	847	7760	35 621	11 398	81 268
Western Europe	15 334	2121	26 826	69 800	22 556	103 049



**Table 8.** Annual production of maize, relative to countries.

Area	Production (1000 tonne)			Yield (hg/ha)		
	Average	Min.	Max	Average	Min.	Max.
Argentina	16 277	4360	60 526	44 173	16 481	78 615
Brazil	35 824	9036	103 964	26 840	11 606	57 734
Canada	7147	742	14 191	68 974	41 108	102 055
China	109 216	16 250	272 552	40 158	11 842	63 177
Egypt	4787	1617	8543	56 287	24 014	83 705
Ethiopia	3110	695	10 722	18 082	9000	42 404
European Union (27)	48 096	16 876	77 575	53 224	20 522	83 617
India	12 060	4312	31 650	16 654	8999	32 099
Indonesia	9316	2254	30 254	25 774	9214	57 237
Kenya	2417	940	4014	15 485	10 713	20 712
Mexico	15 804	6246	28 250	22 087	9867	40 697
Nigeria	4676	488	12 745	13 432	5731	22 545
Pakistan	2255	483	10 635	21 978	9956	64 356
Philippines	4376	1266	8300	16 099	6280	32 368
South Africa	9272	3277	17 551	26 274	7853	58 596
Thailand	3470	598	5300	30 389	14 140	45 981
Türkiye	2633	800	6750	44 074	11 994	96 358
Tanzania	2750	488	7039	13 023	4808	31 359
USA	220 543	88 504	412 262	74 988	39 184	117 433

## Soybean

**Table 9.** Annual production and yield of soybean, relative to region.

Area	Production (1000 tonne)			Yield (hg/ha)		
	Average	Min.	Max	Average	Min.	Max.
Central America	372	20	1041	18 198	12 851	21 343
Central Asia	96	3	296	15 181	6471	21 680
Eastern Africa	236	10	909	13 232	5535	22 381
Eastern Asia	12 127	6810	18 692	13 833	6421	19 349
Eastern Europe	1852	289	9367	10 194	3220	19 528
Middle Africa	21	1	89	7069	4611	8868
Northern Africa	60	1	178	25 726	9494	32 987
Northern America	62 764	18 393	127 931	23 266	15 346	34 600
Northern Europe	1	0	3	15 330	12 692	17 969
Oceania	51	0	130	15 412	3858	23 289
South America	54 264	297	186 189	20 080	9388	32 614
South-eastern Asia	1270	423	2508	10 668	6369	14 847
Southern Africa	235	2	1540	12 457	2500	22 929
Southern Asia	4504	8	14 928	8836	4656	13 589
Southern Europe	783	3	2114	24 190	6757	36 850
Western Africa	348	48	1359	5632	2293	12 655
Western Asia	69	4	252	22 692	7833	43 578
Western Europe	220	1	736	23 577	12 512	29 969

**Table 10.** Annual production of soybean, relative to countries.

Area	Production (1000 tonne)			Yield (hg/ha)		
	Average	Min.	Max.	Average	Min.	Max.
Argentina	18 911	1	61 447	20 943	9765	33 340
Bolivia	999	0	3318	17 877	9302	23 981
Brazil	35 233	271	134 935	20 536	8478	34 450
Canada	2244	136	7717	24 034	14 754	31 151
China	11 609	6140	19 600	14 095	6243	19 833
Korea, North	321	135	455	11 174	5625	17 041
European Union (27)	1102	4	2912	20 273	4682	32 297
India	4621	5	14 666	8556	4348	13 530
Indonesia	823	273	1870	10 702	6203	15 690
Japan	207	99	392	15 363	11 059	20 195
Mexico	338	20	992	17 468	10 925	21 218
Nigeria	303	42	994	5589	2100	12 951
Paraguay	3012	2	11 046	21 422	11 304	33 938
Philippines	3	0	11	10 938	7225	14 286
Korea, South	180	75	319	13 336	5421	20 348
South Africa	279	2	1897	12 714	2500	22 936
Thailand	188	19	672	12 940	7365	17 712
USA	62 529	18 213	120 707	23 660	15 309	34 936
Uruguay	540	1	3212	16 129	7000	29 495

**Cocoa****Table 11.** Annual production and yield of cocoa, relative to region.

Area	Production (1000 tonne)			Yield (hg/Ha)		
	Average	Min.	Max.	Average	Min.	Max.
Caribbean	59	36	102	3591	2214	5203
Central America	49	31	75	4893	2834	6649
Eastern Africa	14	1	62	3956	1716	5839
Middle Africa	172	107	323	2974	1915	3912
Oceania	39	14	66	4122	3235	5112
South-eastern Asia	354	5	866	6128	2649	10 323
South America	437	223	822	3995	2655	5680
Southern Asia	9	1	25	3154	1013	5741
Western Africa	1616	693	3377	4001	2553	5065

**Table 12.** Annual production and yield of cocoa, relative to countries.

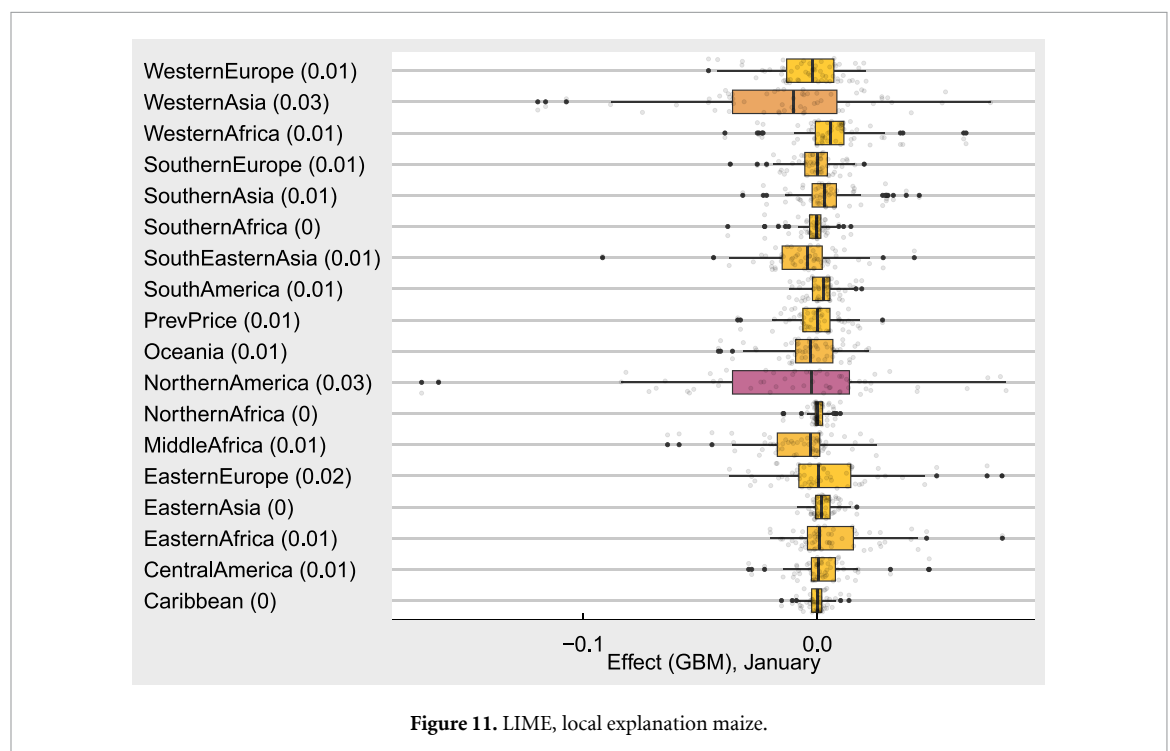
Area	Production (1000 tonne)			Yield (hg/Ha)		
	Average	Min.	Max.	Average	Min.	Max.
Brazil	254	140	459	4411	2788	7415
Cameroon	146	75	310	3184	1976	4104
Colombia	40	14	102	4788	2757	9266
Côte d'Ivoire	840	85	2235	5194	3269	7006
Dominican Republic	46	25	87	3777	2047	5737
Ecuador	94	35	284	2934	1166	5612
Equatorial Guinea	10	1	38	1500	643	4244
Ghana	460	167	969	3369	2054	5495
Guinea	6	2	40	3688	1000	9429
India	9	0	24	3612	1970	8000
Indonesia	291	1	845	5363	1216	11 323
Malaysia	57	0	247	7214	528	13 056
Mexico	37	20	60	5133	2934	7617
Nigeria	269	140	485	3159	2000	4980
Papua New Guinea	34	9	59	4156	3137	5266
Peru	26	2	142	5656	3849	8594
Sierra Leone	13	3	50	4039	2750	5996
Togo	23	4	142	5605	1464	10 274
Uganda	6	0	35	1873	222	4834
Venezuela	18	11	32	2894	1746	5115

### Appendix C. Detailed model agnostic

Three graphs illustrate the distribution of Shapley values obtained for the most accurate model for each commodity. The interpretation of the results appear in the left-hand plot of figures 4, 6, and 8 for maize, soybean, and cocoa, respectively. In each plot, the individual Shapley values are combined into one graph displayed as a box plot, which shows the overall distribution of the values. The features' names are written right to the Y-axis with the corresponding mean Shapley (absolute) value in brackets. The X-axis represents the impact of each feature on the price change, with features having a negative impact on the left and with a positive impact on the right. The bars are colored in accordance with the right-hand plot (global agnostic analysis), where dark purple represents highly influential features, and orange indicates features with a low impact on the model's accuracy. Each bar displays the distribution of individual Shapley values for each feature, as calculated from the second and third quarterlies. The bars are positioned on the X-axis based on the magnitude of each feature's Shapley value. The median of the Shapley values is represented by the black vertical line within each box, and the black horizontal lines represent the lowest and highest inter-quartile ranges (IQR) multiplied by 1.5. Data points located outside the box are considered outliers.

The figures below utilize the model-agnostic method LIME to interpret the results obtained from the most accurate models developed for maize, soybean, and cocoa. Specifically, figures 11, 12, and 16 display the distribution of all the local effects obtained for each crop. These figures utilize a box plot format to combine individual effects into a single graph, presenting the overall distribution of the values. The names of the features are written next to the Y-axis, with the corresponding mean effect (absolute values) displayed in brackets. The X-axis shows the impact of each feature on price change, with features that negatively impact the price positioned on the left and those that positively impact the price on the right. The bars are colored in accordance with the global agnostic analysis, (i.e. the relative importance of the feature). Dark purple bars represent highly influential features, while orange bars indicate features with a low impact on the model's accuracy. Each bar in the plot displays the distribution of individual effect values for each feature, calculated from the second and third quarterlies. The bars' position on the X-axis corresponds to the magnitude of each feature's effect on the price, with the median effect values represented by a black vertical line within each box. The black horizontal lines represent the lowest and highest IQR multiplied by 1.5. Any data points located outside the box are considered outliers.

#### Maize



### Soybean

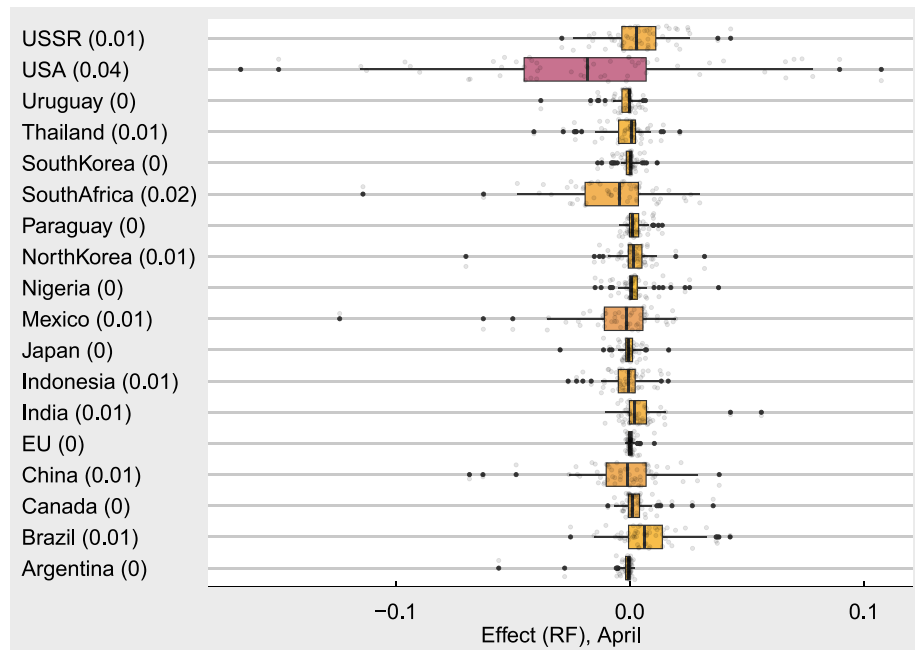


Figure 12. LIME, local explanation soybean

### Cocoa

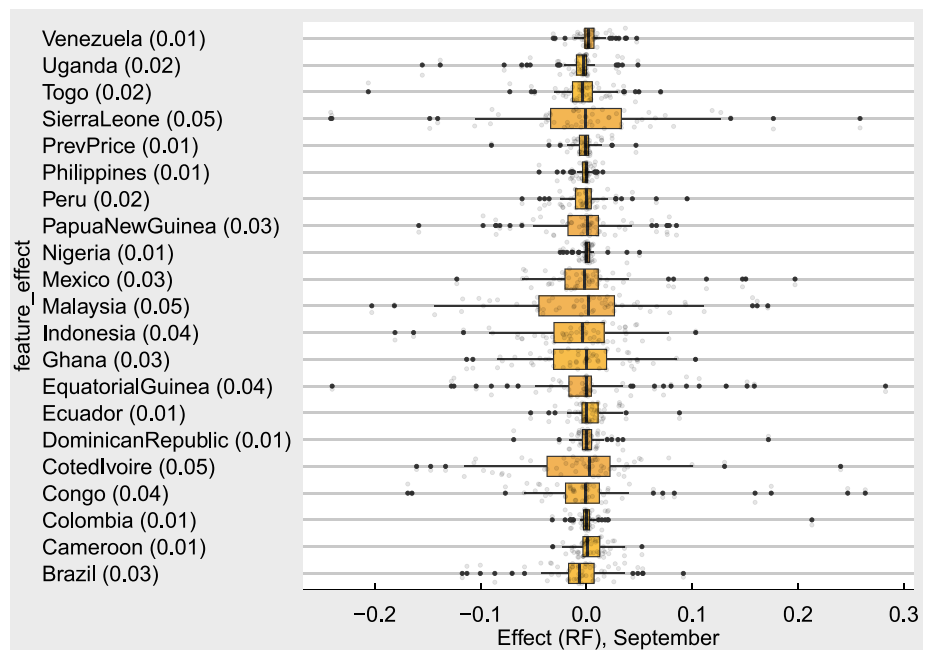


Figure 13. LIME, local explanation cocoa

### Appendix D. Price change forecasts of the models

#### Maize

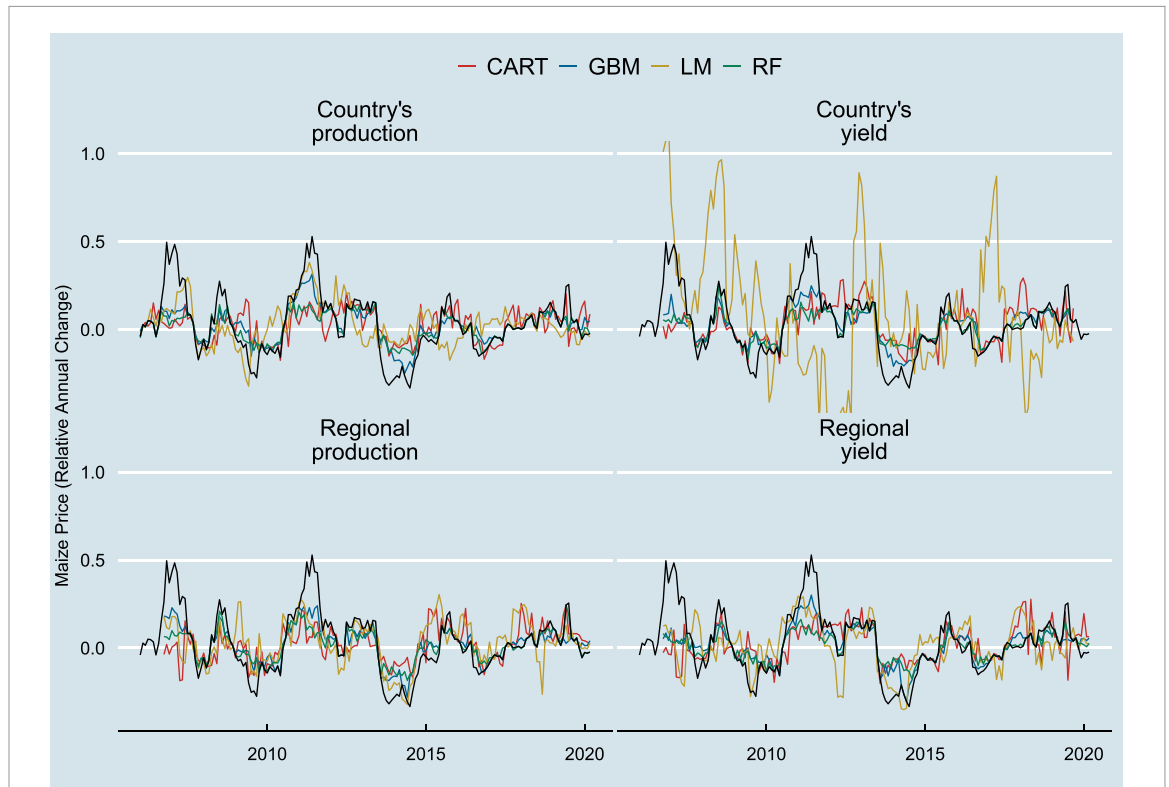


Figure 14. Forecasted maize monthly price changes obtained with all models (colored lines). Observed price dynamics are shown as a black line.

#### Soybean

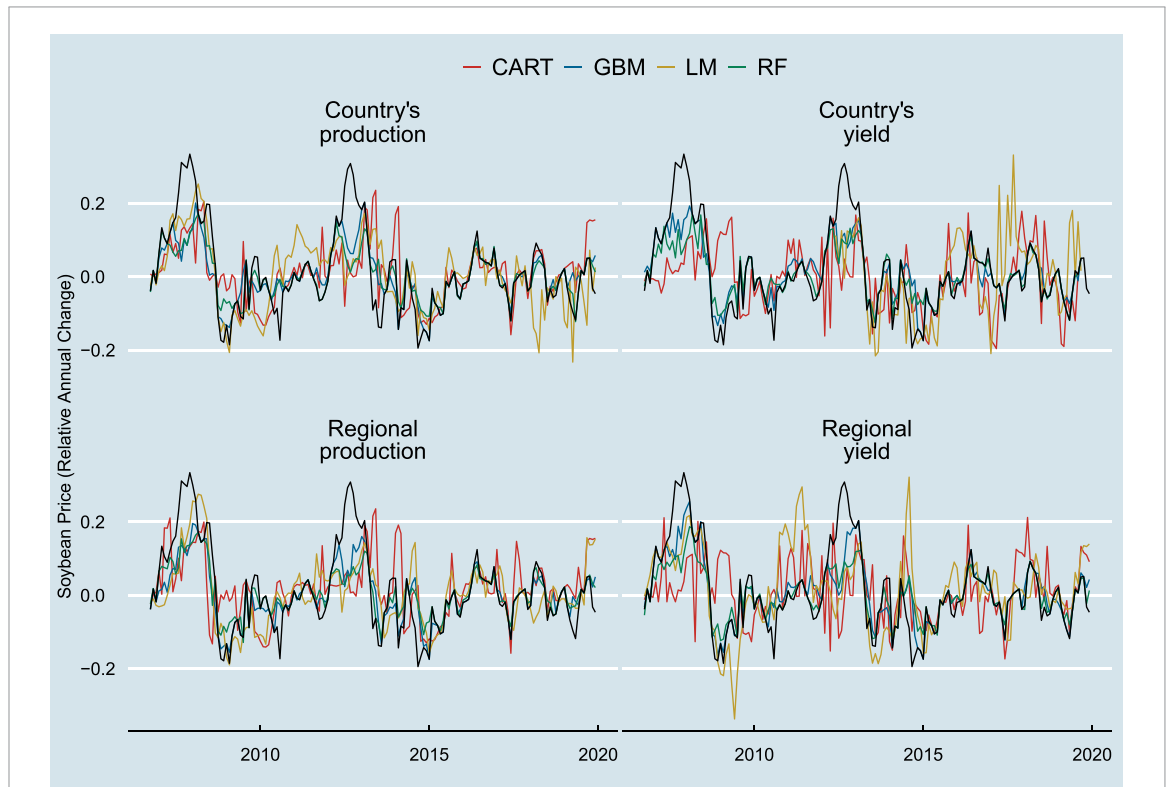
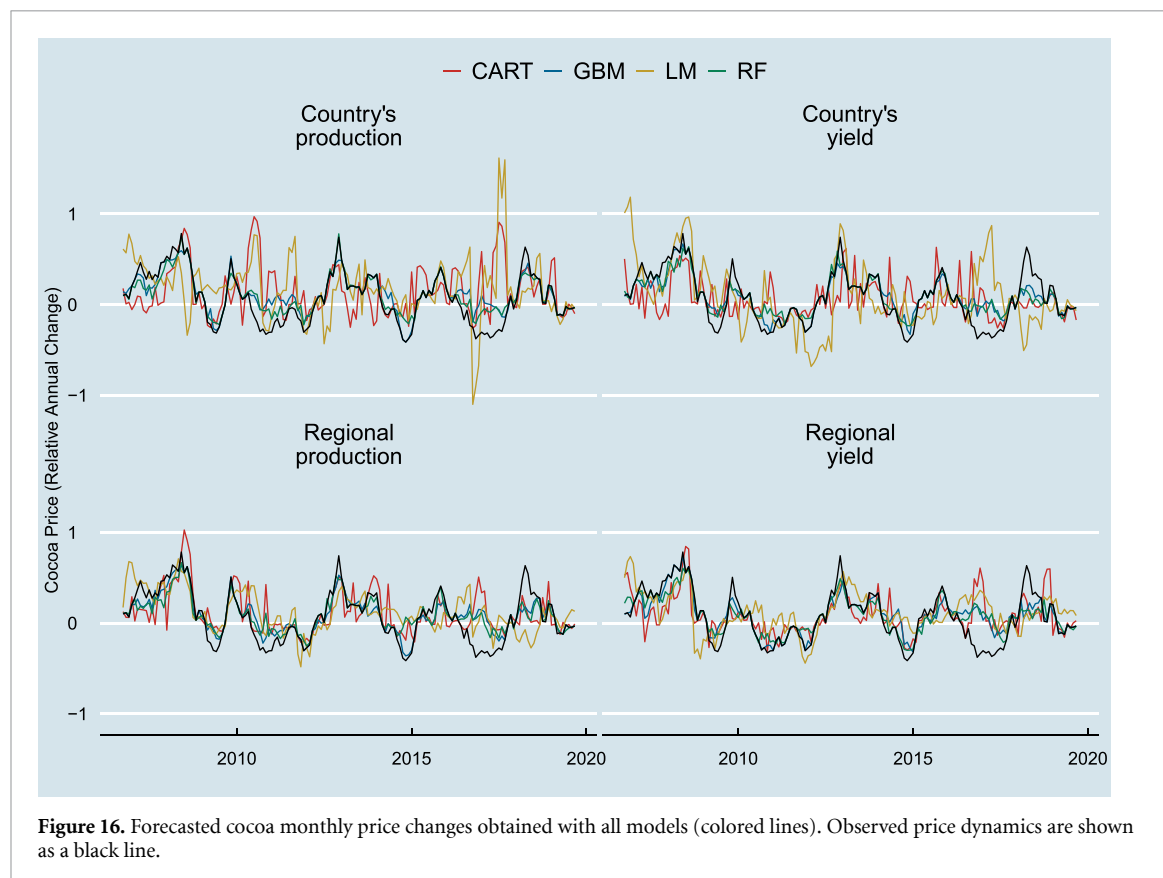


Figure 15. Forecasted soybean monthly price changes obtained with all models (colored lines). Observed price dynamics are shown as a black line.

## Cocoa



**Figure 16.** Forecasted cocoa monthly price changes obtained with all models (colored lines). Observed price dynamics are shown as a black line.

## ORCID iDs

Rotem Zelingher  <https://orcid.org/0000-0002-6383-2838>

David Makowski  <https://orcid.org/0000-0001-6385-3703>

## References

- Akaike H 1974 Stochastic theory of minimal realization *IEEE Trans. Autom. Control* **19** 667–74
- Anderson P, Davenport F, Baylis K and Shukla S 2023 Using earth observation products to predict maize prices in Southern Africa *SSRN Electron. J.* (<https://doi.org/10.2139/ssrn.4369127>)
- Bobenrieth E, Wright B and Zeng D 2013 Stocks-to-use ratios and prices as indicators of vulnerability to spikes in global cereal markets *Agric. Econ.* **44** 43–52
- Breiman L, Friedman J, Stone C J and Olshen R A 1984 *Classification and Regression Trees* (CRC Press)
- Chruslki T 2021 Using the R language computing environment in forecasting natural gas consumption *Polityka Energetyczna - Energy Policy J.* **24** 33–56
- Crespo Cuaresma J, Hlouskova J and Obersteiner M 2021 Agricultural commodity price dynamics and their determinants: a comprehensive econometric approach *J. Forecast.* **40** 1245–73
- De Livera A M, Hyndman R J and Snyder R D 2011 Forecasting time series with complex seasonal patterns using exponential smoothing *J. Am. Stat. Assoc.* **106** 1513–27
- De Maria M, Robinson E, Kangile J, Kadigi R, Dreoni I, Couto M, Howai N, Peci J and Fiennes S 2020 Global soybean trade - the geopolitics of a bean *Technical Report* UK Research and Innovation Global Challenges Research Fund (UKRI GCRF) Trade, Development and the Environment Hub
- FAO 2018 The future of food and agriculture. Alternative pathways to 2050 *Technical Report*, Markets and Trade Division, FAO, Rome
- FAO 2023 Faostat statistical database
- FAS-USDA 2023 The foreign agricultural service (FAS) data & analysis
- Fisher A, Rudin C and Dominici F 2018 All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously
- Gilbert C L 2016 The dynamics of the world cocoa price *The Economics of Chocolate* (Oxford University Press)
- Gilbert C L and Morgan C W 2010 Food price volatility *Phil. Trans. R. Soc. B* **365** 3023–34
- Glauber J W and Laborde Debucquet D (ed) 2023 *The Russia-Ukraine Conflict and Global Food Security* (International Food Policy Research Institute (IFPRI))
- Greenwell B, Boehmke B and Cunningham J 2020 Package 'gbm' Greg Ridgeway
- Hastie T, Tibshirani R and Friedman J 2009 *Random Forests* (Springer) pp 587–604
- Headley D 2011 Rethinking the global food crisis: The role of trade shocks *Food Policy* **36** 136–46
- Henrique B M, Amorim Sobreiro V and Kimura H 2019 Literature review: machine learning techniques applied to financial market prediction *Expert Syst. Appl.* **124** 226–51
- Hyndman R et al 2020 Forecasting functions for time series and linear models NA
- Hyndman R and Athanasopoulos G 2018 *Forecasting: Principles and Practice* 2nd edn (OTexts)
- ITC and UNCTAD/WTO 2001 Cocoa: a guide to trade practices International Trade Centre UNCTAD/WTO

- Kalkuhl M 2016 *How Strong Do Global Commodity Prices Influence Domestic Food Prices in Developing Countries? A Global Price Transmission and Vulnerability Mapping Analysis* (Springer) pp 269–301
- Kalkuhl M, von Braun J and Torero M 2016 *Food Price Volatility and Its Implications for Food Security and Policy* (Springer)
- Kourentzes N, Barrow D and Petropoulos F 2019 Another look at forecast selection and combination: evidence from forecast pooling *Int. J. Prod. Econ.* **209** 226–35
- Kyriazi F, Thomakos D D and Guerard J B 2019 Adaptive learning forecasting, with applications in forecasting agricultural prices *Int. J. Forecast.* **35** 1356–69
- Liaw A and Wiener M 2002 Classification and regression by randomforest *R News* **2** 18–22
- Molnar C 2022 *Interpretable Machine Learning* 2nd edn (available at: <https://christophm.github.io/interpretable-ml-book/>)
- Molnar C, Casalicchio G and Bischl B 2018 iml: an R package for interpretable machine learning *J. Open Source Softw.* **3** 786
- Olofsson M 2020 Socio-economic differentiation from a class-analytic perspective: The case of smallholder tree-crop farmers in limpopo, South Africa *J. Agrar. Change* **20** 37–59
- Organization I. C 2021 Quarterly bulletin of cocoa statistics = bulletin trimestriel de statistiques du cacao *Technical Report*
- Perone G 2022 Comparison of arima, ets, nna, tbats and hybrid models to forecast the second wave of covid-19 hospitalizations in Italy *Eur. J. Health Econ.* **23** 917–40
- Schmidhuber J, Pound J and Qiao B 2020 Covid-19: channels of transmission to food and agriculture *Technical Report* FAO
- Shapley L S 1952 *A Value for N-Person Games* (RAND Corporation)
- Spavound S and Kourentzes N 2022 Making forecasts more trustworthy *Foresight: The International Journal of Applied Forecasting* **66** 21–25
- Therneau T, Atkinson B, Ripley B and Ripley M B 2019 Package ‘rpart’ CRAN
- von Braun J and Torero M 2009 Exploring the price spike *Choices* **24** 16–21
- Wang X, Hyndman R J, Li F and Kang Y 2022 Forecast combinations: an over 50-year review *Int. J. Forecast.* **39** 1518–47
- Wollni M and Zeller M 2007 Do farmers benefit from participating in specialty markets and cooperatives? the case of coffee marketing in Costa Rica<sup>1</sup> *Agric. Econ.* **37** 243–8
- World-Bank 2023 Commodity market, “pink sheet” data
- Wright B 2009 *International Grain Reserves and Other Instruments to Address Volatility in Grain Markets* (The World Bank)
- Zelingher R and Makowski D 2022 Forecasting global maize prices from regional productions *Front. Sustain. Food Syst.* **6** 28