Climate Extremes Impact on Soybean Commodities in the USA

Cristian Viorel Hrenciuc
13274740
University of Amsterdam
Master of Information Studies
Data Science

Name	Dr. Ing. C.M. Rodriguez Rivero	Assoc. Prof. Dim Coumou
Institution	University of Amsterdam	Vrije Universiteit Amsterdam
Email	c.m.rodriguezrivero@uva.nl	d.coumou@vu.nl

Table 1: Thesis Supervisors

ABSTRACT

This study analyses the impact of climate extremes and catastrophes on the price of soy futures in the USA. The climate impact is separated from other drivers of variation through the use of a benchmark model. The benchmark model contains factors from several fields, such that the additional effect of climate variables is isolated and analysed as intrinsic. The methods used are dynamic regression, ARIMA and XGBoost. Our analysis shows that drought has a significant impact on prices, but most climate-derived information does not indicate influential shifts on the price.

1 INTRODUCTION

Soybean is the most important oilseed and livestock feed crop in the world, accounting for 58% of total oilseed production and 69% of protein meal consumption by livestock. Its characteristics - the high protein content and oil content - are not rivaled by any other agronomic crop. Apart from its main uses, which are high-protein livestock, poultry feed and oilseed crop, soybean also has various other industrial uses, such as biodiesel, fatty acids, plastics, coatings, lubricants and hydraulic fluids. In many Asian countries, the whole seed is directly consumed as human food or it is incorporated into human food items.

Soybean production has an increasing importance on the world agricultural stage, justifying the need for further research to create a strong industry. Soybean is among the top ten of the most widely grown crops, with a total production of over 260 million tonnes in 2010 and the cultivated area of soybean occupies more than 100 million hectares worldwide, with about half in the U.S.A. and Brazil.

In 2016 croplands covered 1.53 billion hectares (12% of the earth's ice-free land), with agriculture occupying 38% of the earth's terrestrial surface [24]. Throughout the year 2020, the agricultural output of primary crops has been 9.8 billion tonnes, an increase of 50% from 2000 [11]. The world's population is expected to grow to almost 10 billion by 2050, boosting agricultural demand – in a scenario of modest economic growth – by some 50 % compared to 2013 [10]. Moreover, the size of the agricultural sector has reached 3.4 trillion USD in 2018, and it employs 884 million people, roughly 27% of the global workforce [11]. Agriculture retains its place as an important field of study, in particular when combined with the finance sector. As such, agricultural derivatives are important for both the sustainability sector, as well as from a general financial perspective.

The question of the exact effects climate extremes have on soy yield, while still being a filed of active research, are determined to some extent [17] [22]. The novelty of the current paper in terms of research is modelling the price of soybean through climate variables. Except for the factoring of ENSO variation [16], no research was done with regards to the impact climate variables would have on the price of agricultural commodities or financial instruments derived from them. This paper tries to converge two fields of study, finance and environmental, into an encompassing analysis over soybean futures. The analysis will focus on soybean price in the USA and how it behaves in relation to climate extremes, alongside exogenous variables. Findings conclude that the price of soybean futures is highly autoregressive and it is well impacted by financial factors, such as the behaviour of the S&P 500 Index and the trading volumes. The price does not show significant impacts of seasonality, nor do climate variables have a well determined effect. Droughts tend to decrease prices during autumn and have a slight tendency to increase them during summer.

2 RELATED LITERATURE

It was observed that the price of soy is dependent on the percentage content of protein, which itself is negatively affected by drought [22]. Moreover, synthesis of all the major seed components was inhibited during water stress or high temperature, but protein accumulation was not significantly affected from high temperature by itself [22]. Greenhouse and field studies showed that drought stress led to significant reduction in seed yield from distinct locations and time [2].

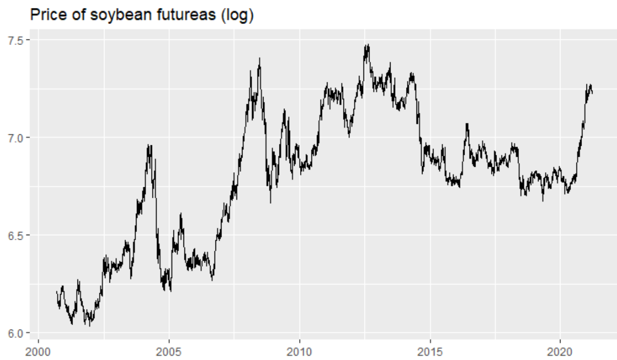

Figure 1: Price of soybean futureas (log)

The biological and chemical effects of drought are analysed in a 2013 meta-analysis [17]. Findings include that the protein and fat oil content are inversely correlated and the protein content increases under severe drought [8], mainly due to differential rainfall throughout the seed filling stage. Soybean seed weight (as measured per 100 seeds) decreases under drought, but the effect is not uniform [9]. The proportion of large seeds (>4.8mm diameter) decreases by 30% - 40%, while the proportion of small seeds (<3.2mm diameter) increases by 3%-15%. The results show the drought tends to affect the size and vigour of the seeds, rather than their number, but several studies show that differences in seed quality have no relation to field performance [3] [17].

As mentioned in [14], soybean culture increased strongly throughout the 20th century in the US, one reason for this expansion being its similarity to corn culture. Most seeds are planted during late spring, but they may be planted as late as early July, depending on the state and general conditions. Crops that are planted late runs the risk of being caught by early autumn frosts and may have issues flowering and setting pods in August [14]. The harvesting period alternates between years and states, but generally it occurs during the autumn, with most of it being done throughout the second half of the autumn [21]. As such, one could have expected that soy is cheapest during the winter, as the harvest is fresh and available nationally, and most expensive during the summer, as the stock of the previous year is dwindling.

A paradigm shift in the production of soy has happened throughout the 21st century. Whereas in 2007 USA produced 38% of the global output of 155 million tonnes [23], in 2019 it decreased to 29% [10]. The top producers were Brazil - 114, United States - 97, Argentina - 55, China -16 (quantities expressed in millions of tonnes per annum), with a global output of 334 millions of tonnes. The shift was done by the discovery of a cultivar that was particularly well adapted to the low latitudes, which boosted the growth and output in Southern America [5]. Moreover, the global output growth was lead by the indirect effect of the increase in meat consumption, with China accounting for 60% of the imports.

Soy is a highly nutritional plant, particularly in terms of the percentage of protein, with defatted soy flour reaching levels of 53 grams of protein per 100 grams of product, with only 0.6 grams of fat [23]. It was estimated that in 2006 humans ate 1.68 billion kg of edible protein [23].

As of November 2016, more than 80 percent of soybeans of USA origin are cultivated in the upper Midwest, with around half of the national output coming from the top five states, which are: Illinois - 14.5, Iowa - 13.7, Minnesota - 8.1, Nebraska - 7.7, Indiana - 7.4 (quantities expressed in million tonnes per annum). Total US exports amounted to 49.7 million tonnes, a decrease from 58.9, the quantity exported in 2016. In financial terms, 78% of the exports were whole soybeans, with China alone purchasing 8 billion USD worth of products [1]. With regards to the trade balance of soy, the top exporters of Soybeans in 2017 were Brazil ($25.9B), the United States ($22B), Argentina ($2.82B), Paraguay ($2.19B) and Canada ($1.91B), summing up to $58.1B. The top importers were China ($36.6B), Mexico ($1.72B), the Netherlands ($1.6B), Japan ($1.41B) and Spain ($1.31B).

The financial world has evolved greatly in the past decades, and agriculture was one of the main drivers of such shift. Among such changes, the concept of futures and derivatives were the most revolutionary, and they are now present in any developed financial system. Futures can be defined as derivative financial contracts that obligate the parties to transact an asset at a predetermined future date and price. The buyer must purchase or the seller must sell the underlying asset at the set price, regardless of the current market price at the expiration date. As such, agriculture commodities are contracts in their own right, which can be sold directly, at the spot price, but they can also be transacted through forwards and futures.

Soybean futures are just one example of commodities futures, along oil, gold, corn and many others. The standard soy future (and the one used in this analysis) contains 5000 bushels of soy (300000 pounds). The reasons why someone might use such financial derivatives is simple. A producer would prefer to lock-in a selling price for the next harvest and thus forecast the costs and income much better. A consumer of soy, conversely, will do the same thing, but on the buy side. It is important to mention that futures are tradable assets, which means that traders may use it for speculative purposes and not for direct commercial benefit.

There were several critical shifts in the price of soy, observable in Figure 1. These include the decrease at the end of 2004, the sharp drop in 2008 and the increase from the summer of 2020 onwards. September 2008 presents the top two days with the highest fluctuation throughout the entire history, with price increasing 274 dollars on the 12th and then dropping 311 dollars on the 15th. Moreover, the spike and drop of 2008 is highly correlated with the financial 'bubble' rising and then bursting. The current spike in prices is thought to be provoked by La Niña, which was forecasted to significantly decrease soy production. At the beginning of 2020, the coronavirus pandemic has forced many meat producers to close their factories, thus reducing the demand in the short-term. When they reopened, the demand increased abruptly, which has started a still ongoing sharp increase in price.

3 METHODOLOGY

name	n	mean	st_dev	median	min	max	range	skewness	kurtosis	st_error
soy	5112	942.401	309.896	937.5	418.5	1771	1352.5	0.241	-0.724	4.334
Adjusted.soy	5122	941.519	310.236	936.5	418.5	1771	1352.5	0.242	-0.727	4.335
diff	5121	0.17	17.07	0.5	-311	274	585	-1.14	39.779	0.239
Volume.soy	5122	47244.753	54286.339	17651	1	352462	352461	0.903	-0.049	758.527
Volume.snp	5122	$3291 * 10^6$	$1498 * 10^6$	$3358 * 10^6$	$356 * 10^6$	$11456 * 10^6$	$11100 * 10^6$	0.644	0.93	$21 * 10^6$
Adjusted.snp	5122	1685.504	720.265	1376.225	676.53	3974.54	3298.01	1.067	0.235	10.064
log_soy	5112	6.79	0.349	6.843	6.037	7.479	1.443	-0.352	-0.775	0.005
log_Adjusted.soy	5122	6.789	0.35	6.842	6.037	7.479	1.443	-0.349	-0.784	0.005
log_diff	5121	0	0.016	0.001	-0.234	0.203	0.437	-0.914	17.515	0
log_Volume.soy	5122	8.804	2.937	9.779	0	12.773	12.773	-0.758	-0.34	0.041
log_Volume.snp	5122	21.798	0.509	21.935	19.691	23.162	3.471	-0.572	-0.286	0.007
log_Adjusted.snp	5122	7.349	0.392	7.227	6.517	8.288	1.771	0.501	-0.781	0.005
Drought	5122	0.252	0.434	0	0	1	1	1.141	-0.699	0.006
Earthquake	5122	0.001	0.031	0	0	1	1	31.95	1019.002	0
Epidemic	5122	0.038	0.192	0	0	1	1	4.799	21.031	0.003
Flood	5122	0.098	0.298	0	0	1	1	2.7	5.289	0.004
Storm	5122	0.193	0.394	0	0	1	1	1.558	0.427	0.006
Wildfire	5122	0.157	0.364	0	0	1	1	1.885	1.555	0.005
Drought_lag	5032	0.257	0.437	0	0	1	1	1.113	-0.761	0.006
Max.Temperature	3460	13.447	11.983	14.342	-20.175	38.191	58.366	-0.178	-1.033	0.204
Min.Temperature	3460	4.311	10.012	5.385	-25.883	25.085	50.968	-0.398	-0.686	0.17
Precipitation	3460	2.717	3.475	1.357	0	27.887	27.887	2.285	6.303	0.059
Wind	3460	4.34	1.374	4.148	1.821	10.503	8.682	0.807	0.602	0.023
Relative.Humidity	3460	0.743	0.125	0.753	0.293	0.978	0.685	-0.324	-0.635	0.002
Solar	3460	14.393	7.625	13.422	1.048	31.583	30.536	0.251	-1.183	0.13

Table 2: Database Statistics

The data is obtained from numerous sources, each providing specific input. The main information is obtained from EM-DAT, which contains the climate catastrophes and extremes that have happened worldwide since 1900 [4]. The database was selected to span 01.01.2000 - 31.03.2021 and subsetted to events in the USA alone.

Each event has the information of a starting time and an ending time, segmented into year, month and day. In case of a missing day, it was assumed that the event lasted the entire month. In the particular case of drought it is usually difficult to set a particular week or even a month from which the droughts emerge, as they are events that require a prolonged absence of water and last for a long period of time. All types of drought (meteorological, hydrological and agricultural) are grouped in a single bucket. In terms of length, it was assumed that the drought lasted throughout the spring and summer of the years where drought was present.

The data was further subsetted to contain only natural disasters, thus excluding the technological ones. As such, the disasters to be analysed are the following: droughts, earthquakes, epidemics, floods, storms and wildfires.

Information regarding both soy futures and the S&P 500 index was obtained through the quantmod package in R, using symbols GSPC for S&P and ZS=F for soy futures. The database contains information regarding both the price and the volume traded at each day during the time interval 2000-09-15 to 2021-03-31. The adjusted closing price was one of the variables contained and it represents the closing price after adjusting to any corporate action, such as dividend payment (which may abruptly decrease the price); as such, the adjusted price will be the variable of concern, for ease of interpretability. Three observations were excluded, due to the fact that in those days there was no soy being traded, thus having a 0 value for the volume of traded soy. Given that it was impossible to take the logarithm, it was preferred to remove these observations, rather than change their values.

For each day in the interval, all the climatic events were aggregated into a column and then separated into columns with binary variables which specify if the event took place in that day or not. The season is extracted from the date and used as a variable. The more formal terminology is used, which abbreviates the names of the month, but the interpretation will refer to the seasons of the Northern Hemisphere (DJF - winter, MAM - spring, JJA - summer, SON - autumn).

A further set of information was obtained from CFSR [12] with information regarding observed climate variables for the US, including Temperature (°C), Precipitation (mm), Wind (m/s), Relative Humidity (fraction) and Solar intensity(MJ/m2). The data was selected for the entire north-Midwestern area, in order to better aggregate the climatic conditions of the highest soy-producing regions. Preliminary analysis also contained such observations for each state separately, but they were proven to be much weaker predictors. The CFSR data ends on 2014-07-31, much sooner than the other data available, so its use alters the training and testing sets. As such, the results are segmented into two separate sections, depending on the presence of CFSR variables.

The reason for the segmentation according to the presence of the CFSR variables is data availability. The initial dataset contains information spanning 01.01.2000 - 31.03.2021, with the train/test cutoff on 15.02.2017. The CFSR variables are only available until 31.07.2014, which completely miss the entire test data. Moreover, it results in a database imbalance, as there are 5122 observations in the initial database and CFSR variables only populate 3460 rows. As such, the benchmark was adjusted to both sets of data. This also enables us to observe the change in parameters and significance regarding all variables by comparing the coefficients of the same regression between databases. This way, comparisons are not skewed and we allow for interpretability of the models throughout time.

The variables used are lagged by 10 working days, such that the forecasting accuracy is maintained over longer horizons. Technically, this has meant that a lag of 10 is applied to all variables, with only the log-price being on the day of analysis. Drought is a particular exception, as two versions are analysed, one lagged by 10 days (Drought) and one lagged by 100 days (Drought_lag), in order to capture the long-term effect droughts have. The variables of interest are analysed in the logarithmised version, as it decreases the variation in the order of magnitude and presents a distribution that is better mapped to the normal one. This is particularly relevant in the case of volumes, where $\frac{mean(Volume.snp)}{mean(Adjusted.soy)} = 3496311$ and $\frac{mean(log\_Volume.snp)}{mean(log\_Adjusted.soy)} = 3.21$. The decrease in direct interpretability is small, as a log-log regression has the benefit of enabling the percentage interpretation, (e.g 'a 1% increase in the volume of soy traded leads to a 0.44% increase in price'). In addition, it was decided that the analysis will not take differences of any order, regardless of the non-stationarity of the data and the high autoregressivity. In a dummy example, the first order difference of log-price, log_diff, was regressed against all the variables of the database, resulting in an R^2 value of 0.023. Nevertheless, the variables found in the benchmark model had a tendency to remain relevant at 10%.

The benchmark model contains several variables, in order to more accurately forecast and contain drivers of explainability. A lag of 10 for the explainable variables was chosen to balance accuracy in forecasts and a forward looking prediction. The price is lagged by 10 days, in order to address autoregression, while still retaining forward looking forecasts. Moreover, the benchmark model will include the price of the S&P index, in order to address the fact that soy futures are financial products and, as such, are deeply dependent on the financial world. The trading volume of soy futures is included, as a large trading quantity of soy tends to increase the price and it is highly significant. The seasonal aspect is handled by implementing a dummy variable for the season of the observation. The volume of S&P, while highly significant when regressions do not contain autoregressive components, is discarded due to low fit, low predictability and high volatility.

The models implemented are mostly centered around dynamic regressions, which are extensions of generalised regression that also contain lagged regressors. The results section is split into two parts: one without the CFSR variables (initial dataset) and one that contains them (extended dataset). Each of these parts will contain the results to the following models: benchmark, final model, final model without autogregressive terms, AR1 and XGBoost, as well as what would be the effect of each variable in the database if added to the benchmark.

Given that the variables have complex inter-relations, a combination of measures of model suitability is used in order to succinctly define the model. RMSE (root mean square error) will be the first measure of forecasting accuracy. It is defined as $RMSE = \sqrt{\frac{1}{n}(\sum(\hat{y}_i - y_i)^2}$, where $\hat{y}$ represent predicted values and y represent observed values. Conceptually, it is similar to a normalized Euclidean norm. Furthermore, MAPE (mean absolute percentage error) is the second measure of forecasting accuracy . It is defined as $MAPE = \frac{100}{n} \sum \frac{\hat{y}_i - y_i}{y_i}$. It is an intuitive and non-dimensional measure of comparing models, conceptually quantifying the error in percentages. While it may be used to compare models between them, regardless of the structure, dataset size and variables, it is biased by over-representing errors that estimate above the observed value [18]. R^2 is used as the usual measure of fit for a regression. BIC (Bayesian Information Criterion) is used to balance explainability power and the complexity of the model, where a lower BIC is preferable.

The assumption that the distribution of the log-price is normal does not withstand the statistical Shapiro test, nor the graphical QQ analysis. Even after implementing the logarithm, the log-price fails the Shapiro test at high significance (< 2.2e-16) due to several outliers that render a number of 16 values under -0.2, which is 4 standard deviations away from the mean of 0.

Descriptive statistics of the database can be seen in Table 2, where diff refers to the first order difference of the linear variables and log_diff refers to the first order difference of the log variables.

XGBoost

Given that the other methods presented in this paper (dynamic regression and autoregression) are well known and of standard use in statistical analysis, it is assumed that the reader is familiar with them and only XGBoost will be properly introduced. XGBoost is an acronym and it stands for eXtreme Gradient Boosting, a boosting technique designed and optimized to be efficient, flexible and portable [13]. The R implementation [6] includes both linear model solver and tree learning algorithm. It supports various objective functions, including ranking, classification and regression [7]. It is built upon the gradient boosting framework, but with an efficient and scalable implementation and optimization structure [20].

In order to render the model simple and to avoid overfitting, a regularized model is used to control the complexity of the model. An added benefit of XGBoost is the fact that is supports parallelization as it implements parallelization when selecting the best splitting points for enumeration, which greatly enhances the training speed [25]. The training speed is accelerated through premature ending of the algorithm if the errors are convergent. Furthermore, it supports setting sample weight embodied in the firstly derivative and the second derivative. We can pay more attention to some samples by adjusting the weight.

Boosting algorithms may be summarised as

$$F_m(x) = \sum_{i=1}^{n} \beta_i f_i(x),$$

where m represents the number of weak learners, f is a weak learner, β is the coefficient and F stands for the general model. The objective of the model is to minimize the loss function (L), which means that

$$F_m argmin \sum_{i=1}^{n} L(y, F_m(x_i)) =$$

$$\beta_m argmin \sum_{i=1}^{n} L(y, F_{m-1}(x_i) + \beta_m f_m(x_i))$$

Given that F is weighted by several weak learners, it is not possible to solve the system analytically simultaneously, problem which is solved by the implementation of a greedy algorithm. Gradient boosting minimizes the loss function by 'cancelling' the errors through making each new term equal to the negative gradient of the loss function in the previous step, which results in a β of

$$\beta_m f_m(x_i) = \gamma \frac{\delta L(y_i, F_{m-1}(x_i))}{\delta F_{m-1}(x_i)},$$

where γ is the step size. Thus, the equation for the model is

$$F_m(x) = F_{m-1}(x) + \beta_m f_i(x)$$

XGB improves on this general structure through optimizing the regularization learning target and it is particularly well suited for regressions with a low number of variables, such as is our case. The data split follows the same treatment as before, 80% allocated for training and 20% for testing.

4 RESULTS - INITIAL DATASET

Benchmark

Variable	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.2881	0.0282	10.196	< 2e-16 ***
log_Adjusted.soy	0.9628	0.0031	309.896	< 2e-16 ***
log_Volume.soy	0.0050	0.0004	12.345	< 2e-16 ***
log_Adjusted.snp	-0.0103	0.0034	-2.959	0.0030 **
SeasonMAM	0.0071	0.0023	3.038	0.0023 **
SeasonJJA	-0.0104	0.0024	-4.354	1.37e-05 ***
SeasonSON	-0.0077	0.0023	-3.366	0.0007 ***

Table 3: Benchmark Model

Table 3 presents the results of the benchmark regression, where we can clearly see the critical importance of the autoregressive term from the value of the coefficient (0.96) and

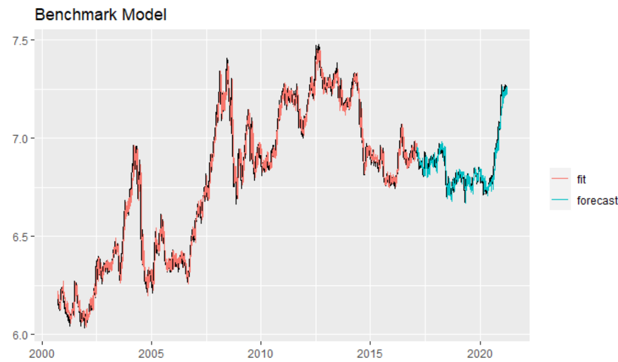

Figure 2: Benchmark Model

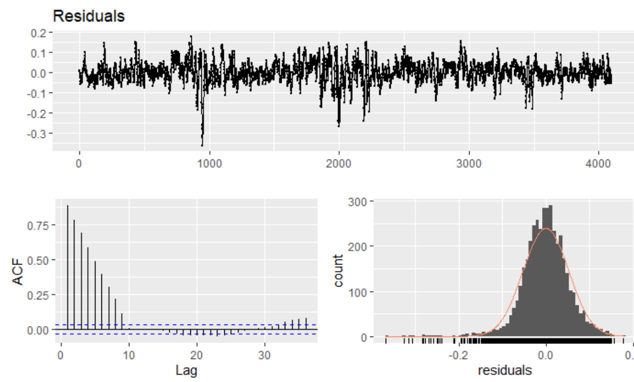

Figure 3: Benchmark Model - Residuals

the low p-value. Figure 2 reveals visual confirmation that the regression is a good fit for the data, as it closely follows the price during both the training and the testing period.

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
Drought	0.0078	0.0023	3.357	0.0007 ***
Earthquake	0.0187	0.0262	0.714	0.4750
Epidemic	-0.0030	0.0040	-0.736	0.4618
Flood	0.0001	0.0026	0.068	0.9454
Storm	-0.0003	0.0021	-0.163	0.8704
Wildfire	-0.0068	0.0026	-2.597	0.0094 **
Drought_lag	-0.0180	0.0021	-8.414	< 2e-16 ***

Table 4: Additional Variables

Table 4 shows the effect that each variable would have if added to the benchmark model. It must be noticed that Drought and Drought_lag have opposing effects and that the presence of wildfires tends to decrease the price by 0.68%. The presence of the epidemics, floods and storms seems not to have a profound impact on the price, nor do they have any statistical significance. Earthquakes have a high impact on the price, reducing it by almost 2%, but this is not statistically relevant, with a p-value of 0.47. The implementation of each variable has a very small impact on the coefficients of the variables in the benchmark model (<2%, except for the intercept), so one can assume that they remain identical to those already existing in the benchmark. As such, one can interpret the changes brought by the environmental variables at face value, as shifts brought by the presence of the environmental event.

Final Model

Variable	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.3264	0.0281	11.597	< 2e-16 ***
log_Adjusted.soy	0.9606	0.0031	307.949	< 2e-16 ***
log_Volume.soy	0.0051	0.0004	12.713	< 2e-16 ***
log_Adjusted.snp	-0.0135	0.0034	-3.883	0.0001 ***
SeasonMAM	0.0053	0.0028	1.860	0.0629 .
SeasonJJA	-0.0162	0.0029	-5.518	3.64e-08 ***
SeasonSON	0.0054	0.0029	1.890	0.0587 .
SeasonDJF:Drought_lag	-0.0009	0.0041	-0.229	0.8186
SeasonJJA:Drought_lag	-0.0271	0.0042	-6.459	1.18e-10 ***
SeasonSON:Drought_lag	-0.0326	0.0033	-9.851	< 2e-16 ***
SeasonMAM:Drought	0.0015	0.0032	0.467	0.6405
SeasonJJA:Drought	0.0255	0.0036	6.922	5.16e-12 ***

Table 5: Final Model

Table 5 shows the final model, which included all the variables found in the benchmark model, alongside Drought and Drought_lag, which are segmented by each season.

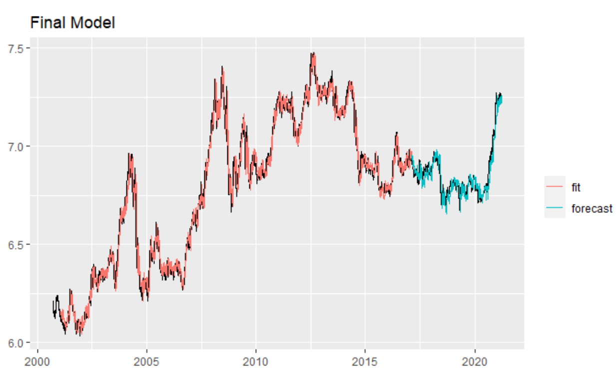

Figure 4: Final Model

Graphically, figures 4 and 5 present very little difference to the ones for the benchmark model. The histogram distribution of the residuals better resembles a normal one in the case of the final model, but the variation is not statistically significant and the extreme values found in the left tail are present in both cases.

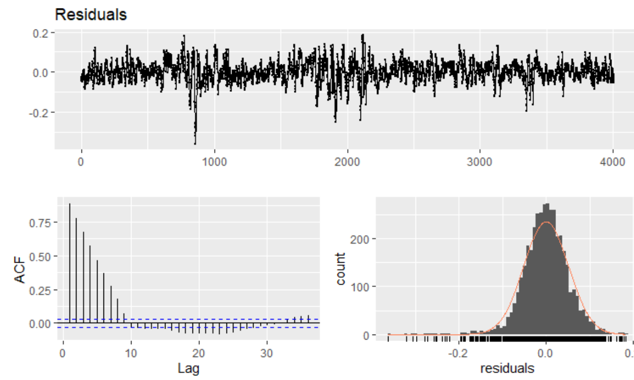

Figure 5: Final Model - Residuals

No Autoregressive Component

We shall further analyse the final model without the variable Adjusted.soy, thus excluding any autoregressive component.

Variable	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	4.92	0.1187	41.443	< 2e-16 ***
log_Volume.soy	0.0860	0.0015	55.458	< 2e-16 ***
log_Adjusted.snp	0.1480	0.0171	8.614	< 2e-16 ***
SeasonMAM	0.1051	0.0142	7.384	1.86e-13 ***
SeasonJJA	0.0985	0.0145	6.779	1.39e-11 ***
SeasonSON	0.0300	0.0144	2.082	0.0374 *
SeasonDJF:Drought_lag	0.0117	0.0204	0.575	0.5655
SeasonJJA:Drought_lag	0.0048	0.0209	0.231	0.8173
SeasonSON:Drought_lag	-0.0451	0.0164	-2.738	0.00621 **
SeasonMAM:Drought	0.0043	0.0162	0.266	0.7902
SeasonJJA:Drought	0.1077	0.0183	5.882	4.39e-09 ***

Table 6: Final Model without AR component

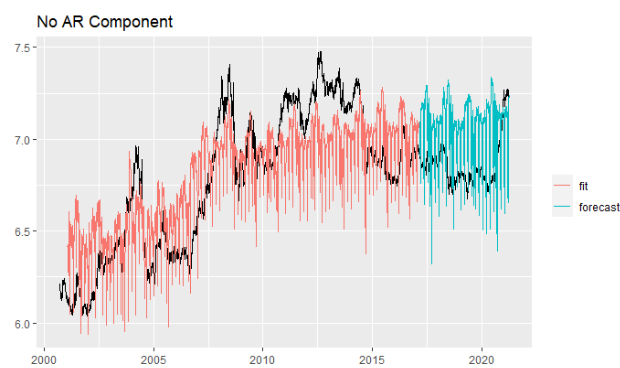

Figure 6: Model Without AR Component

It is clear that the model is extremely volatile and it does but a mere resemblance of the general trend, as seen in Figure 6. Without any element that would smooth the results, the

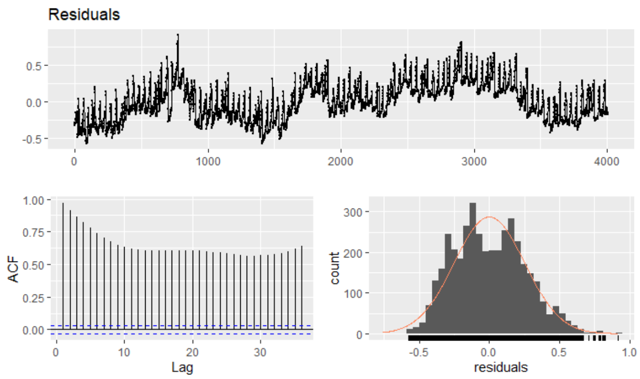

Figure 7: Model Without AR Component - Residuals

forecasts are unreliable. Figure 7 shows that, while the residuals do not have the relative extreme observations found in the previous models, the distribution is even further from the normal and the ACF function plateaus at 0.6 and does not decrease after 10 lags.

AR1

In order to have a full image of the trajectory of the price, an AR(1) model is implemented. The resulting equation is

$$log\_soy_t = 0.0088 + 0.9987 * log\_soy_{t-1}$$

with a p-value of 0.0676 for the intercept and extremely small (<2e-16) for the AR1 component.

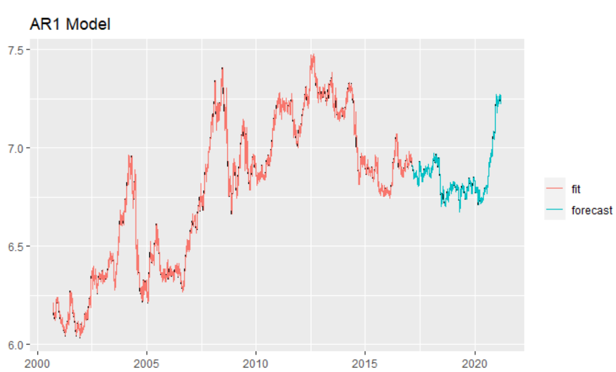

Figure 8: AR1 Model

Figure 8 shows an almost perfect overlap, while the residuals plotted in Figure 9 show no significant sign of autocorrelation. Even if the distribution of residuals is highly leptokurdic (kurtosis = 16.4), it is the one that best matches a normal distribution so far.

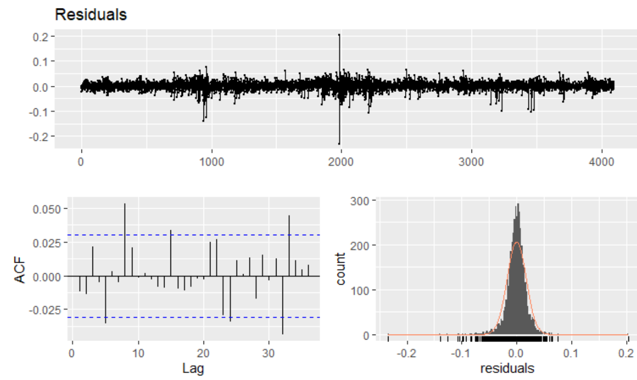

Figure 9: Residuals AR1

XGBoost

The target indicators quickly converge to similar values after a number of 30000 runs, regardless of other parameters. The reguralizer parameters are empirically optimised on both the training and testing set, fact which led to $\lambda = 3$ as the L2 regularizer and $\alpha = 2$ for the L1 regularizer, as the implementation of both resulted in consistently optimal values for both RMSE (0.047) and MAPE (0.45).

Feature	Weight
soy	9.80828e-04
Adjusted.snp	3.22585e-05
Adjusted.soy	2.09146e-05
Volume.soy	1.56865e-07
Volume.snp	2.65632e-11

Table 7: Feature Importance XGB

The weight of almost all the other features is fixed at 1.40130e-45 with an alternating sign, with the probable goal of cancelling each other. The only variables set to exactly zero were Earthquake, Epidemic, Flood, Storm and log_diff. The reason for the fact that the climate variables are regarded as irrelevant are the binary distribution they have, as well as the low number of positive observations. XGBoost results suggest that the most important elements in predicting the price of soybean futures are the autoregressive element and the financial data.

The model is quite accurate in the prediction, as it closely mirrors the trend, spikes and troughs. The significant increase in price from mid-2020 prompts the model to aim higher than the actual value, while still retaining the general trend.

5 RESULTS - EXTENDED DATASET

The same analysis will be done, but for the extended dataset, which includes the CFSR variables.

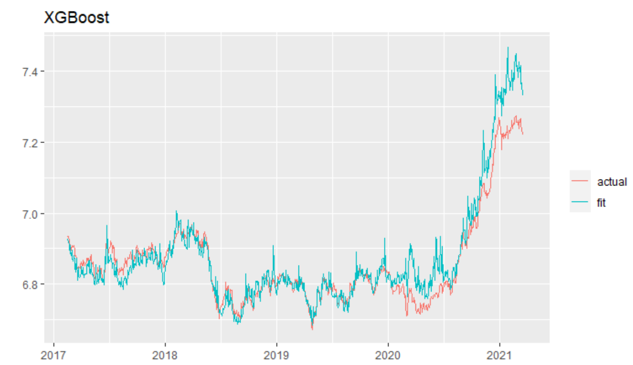

Figure 10: XGBoost

Benchmark

Variable	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.1831	0.0547	3.344	0.0008 ***
log_Adjusted.soy	0.9626	0.0043	222.805	< 2e-16 ***
log_Volume.soy	0.0044	0.0005	8.319	< 2e-16 ***
log_Adjusted.snp	0.0054	0.0069	0.782	0.4344
SeasonMAM	0.0048	0.0031	1.555	0.12
SeasonJJA	-0.0124	0.0031	-3.909	9.49e-05 ***
SeasonSON	-0.0059	0.003	-1.947	0.0516 .

Table 8: Benchmark Model Extended

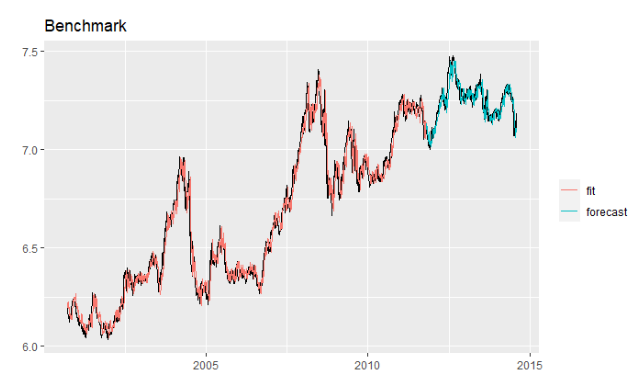

Figure 11: Benchmark Extended

The current benchmark model shows that past price of soy (log_Adjusted.soy), the past soy trading volume (log_Volume.soy) and the seasonality have the same effect, both in direction, as well as in magnitude. The correlation with log_Adjusted.snp is positive for this time interval, while in Table 3 it doubles and changes sign.

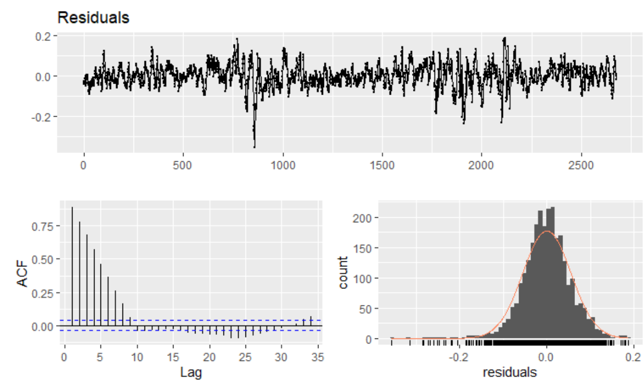

Figure 12: Benchmark Residuals Extended

Variable	Estimate	Std. Error	t value	Pr(>|t|)
Drought	0.0138	0.003	4.487	7.51e-06 ***
Earthquake	0.0085	0.0328	0.260	0.7952
Epidemic	0.0008	0.0048	0.177	0.8597
Flood	0.0008	0.0032	0.264	0.7920
Storm	-0.0038	0.0031	-1.230	0.2188
Wildfire	-0.0141	0.0034	-4.078	4.66e-05 ***
Drought_lag	-0.0205	0.0031	-6.543	7.21e-11 ***
Max.Temperature	-0.0004	0.0001	-2.462	0.0138 *
Min.Temperature	-0.0004	0.0001	-2.502	0.0123 *
Precipitation	0.0001	0.0003	0.339	0.7346
Wind	0.0012	0.0008	1.439	0.1501
Relative.Humidity	-0.0062	0.0111	-0.563	0.5737
Solar	-0.0001	0.0002	-0.590	0.5553

Table 9: Additional Variables Extended

Variable	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.1990	0.0816	2.437	0.0148 *
log_Adjusted.soy	0.9552	0.005	190.496	< 2e-16 ***
log_Volume.soy	0.0039	0.0006	6.493	9.99e-11 ***
log_Volume.snp	0.0074	0.0032	2.322	0.0203 *
log_Adjusted.snp	-0.0079	0.0076	-1.044	0.2966
SeasonMAM	0.0002	0.0045	0.063	0.9499
SeasonJJA	-0.0115	0.0062	-1.848	0.0647 .
SeasonSON	0.0125	0.0049	2.544	0.011 *
Storm	-0.0053	0.0031	-1.720	0.0855 .
Wildfire	-0.0129	0.0036	-3.528	0.0004 ***
Max.Temperature	-0.0004	0.0001	-2.485	0.0130 *
Precipitation	0.0005	0.0003	1.483	0.1381
Relative.Humidity	-0.0333	0.014	-2.382	0.0172 *
SeasonDJF:Drought_lag	0.0031	0.0066	0.470	0.6386
SeasonJJA:Drought_lag	-0.0158	0.0057	-2.770	0.0056 **
SeasonSON:Drought_lag	-0.0455	0.0047	-9.659	< 2e-16 ***
SeasonMAM:Drought	0.0155	0.0044	3.534	0.0004 ***
SeasonJJA:Drought	0.0256	0.005	5.103	3.57e-07 ***

Table 10: Final Model Extended

Final Model

Table 10 shows the final model with the extended database. As in the previous section, both variables representing drought are present in the model, highly relevant and segmented by season. The added variables are Storm, Wildfire, Max.Temperature, Precipitation and Relative.Humidity. Precipitation was included, even if it fails the significance test, due to the improvement it bring in forecasts and the intuition behind it's impact.

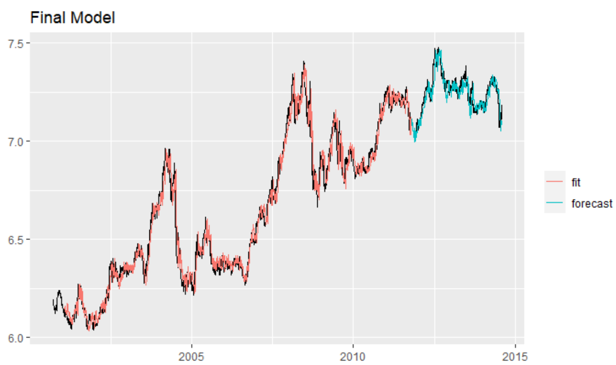

Figure 13: Final Model Extended

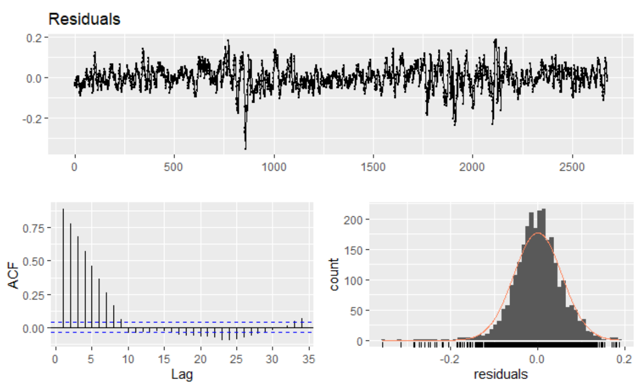

Figure 14: Final Model - residuals Extended

No Autoregressive Component

Similarly with the previous section, the lack of the AR component results in a highly volatile model due to the variations in the trading volume. The regressors in Figure 16 are highly correlated and present a significant left skewness when compared to the normal distribution. We notice that all variables (except for Max.Temperature) have an important increase in significance, with Precipitation dropping its p-value lower than 1%. On the other hand, when comparing to the final model in Table 10, we notice that the signs of many variables are switched, fact which decreases the trust one would have as to the direction these climate impacts would affect the price.

Variable	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-1.0796	0.3114	-3.467	0.0005 ***
log_Volume.soy	0.0412	0.0021	18.910	< 2e-16 ***
log_Volume.snp	0.2977	0.0108	27.537	< 2e-16 ***
log_Adjusted.snp	0.1118	0.029	3.849	0.0001 ***
SeasonMAM	0.0306	0.0175	1.751	0.0801 .
SeasonJJA	0.0467	0.0239	1.956	0.0505 .
SeasonSON	0.0193	0.0188	1.027	0.3047
Storm	-0.0298	0.0119	-2.497	0.0125 *
Wildfire	-0.0492	0.014	-3.501	0.0004 ***
Max.Temperature	0.0003	0.0007	0.415	0.6780
Precipitation	-0.0037	0.0014	-2.693	0.0071 **
Relative.Humidity	0.2249	0.0533	4.214	2.60e-05 ***
SeasonDJF:Drought_lag	-0.0844	0.0255	-3.306	0.0009 ***
SeasonJJA:Drought_lag	0.0652	0.0218	2.987	0.0028 **
SeasonSON:Drought_lag	-0.0850	0.0180	-4.725	2.42e-06 ***
SeasonMAM:Drought	0.0504	0.0168	2.995	0.002766 **
SeasonJJA:Drought	0.0949	0.0191	4.955	7.70e-07 ***

Table 11: Final Model without AR Component Extended

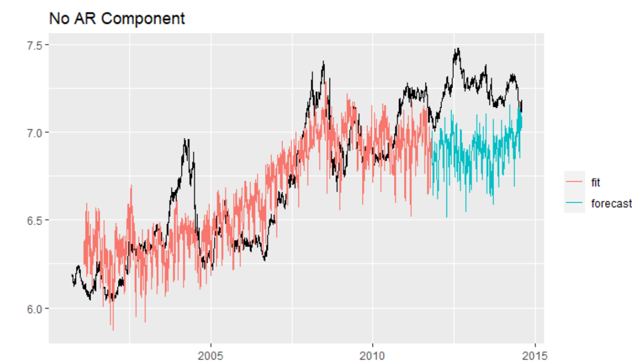

Figure 15: Final Model without AR Component Extended

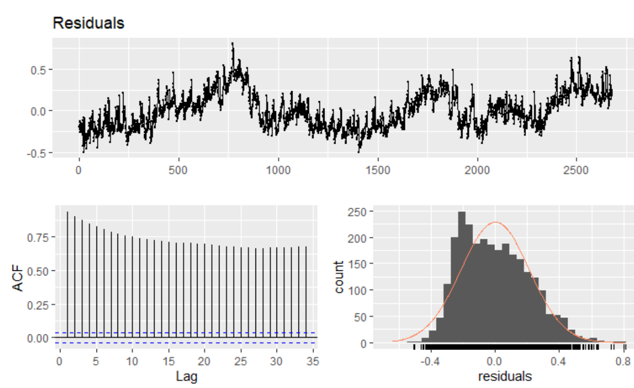

Figure 16: Final Model without AR Component Extended-residuals

AR1

The resulting equation is

$$log\_soy_t = 0.0085384 + 0.9987 * log\_soy_{t-1}$$

with a p-value of 0.18 for the intercept and infinitesimal (<2e-16) for the AR1 component. The residuals (Figure 18) and the predictions (Figure 17) follow exactly the same structure as in the previous section, but they are also present here for completeness.

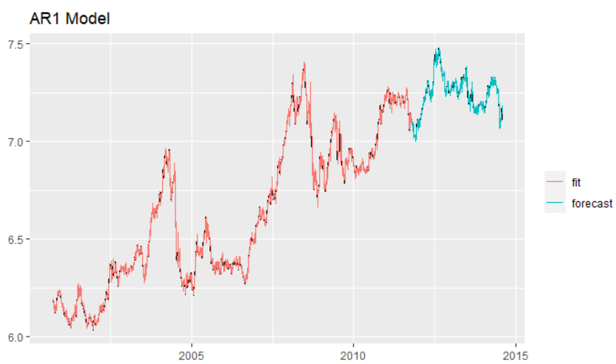

Figure 17: AR1 Extended

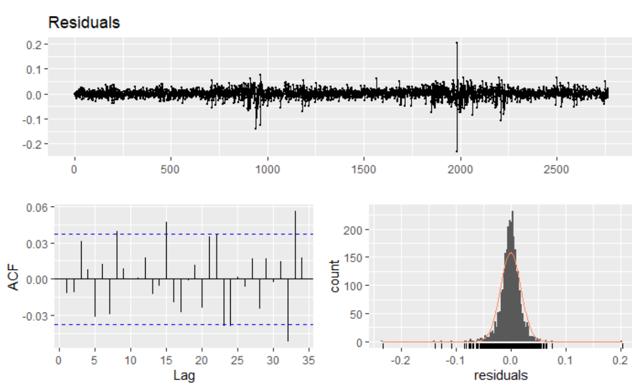

Figure 18: AR1 - residuals Extended

XGBoost extended

Feature	Weight
soy	1.08947e-03
Volume.soy	3.12569e-08
Volume.snp	1.85535e-11

Table 12: Feature Importance XGB Extended

For this instance of the database, the XGBoost algorithm focused on only 3 variables, as seen in Table 12. The same parameters were used, $\lambda = 3$ and $\alpha = 2$, while keeping the number of runs equal to 30000.

Figure 19 shows that the model accurately follows the data, even if it slightly overreacts to the trough - spike variation.

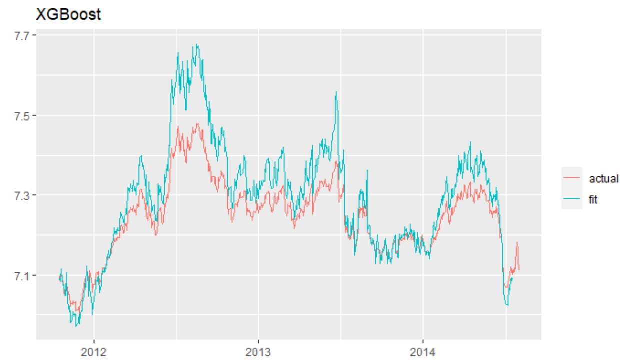

Figure 19: XGBoost Extended

The lack of a significant portion with trend renders this model a better estimation for the extended database than that found in the previous section.

6 DISCUSSION

The aggregated values for the models are presented in Table 13 and 14. By some standards, it would be obvious that the AR1 model is better than the rest. It is significantly and consistently better than all the other models, while having the least amount of parameters. However, it must be noted that it does a poor job of predicting over longer periods of time, as it would virtually be a naive forecast for periods longer than one day. Moreover, it has no resistance to shock, as there are no other regressors that may forecast any impact brought by exogenous variables. Lastly, the model was introduced as a means of analysing the strength of the autocorrelation regarding the price of soy and it does not answer any question related to the impact climate variables have on soybean price.

Model	RMSE	MAPE	R^2	BIC
benchmark	0.0523	0.5743	0.9813	-12480
final	0.0515	0.5669	0.9812	-12269
no AR	0.2566	3.1641	0.5358	579
AR1	0.0174	0.1780	0.9979	-21529
XGBoost	0.0477	0.4561	-	-

Table 13: Forecasting Accuracy

With regards to the measures of forecasting accuracy, it seems that the addition of climate variables to the benchmark model decrease both the RMSE and the MAPE. The improvements are minimal, if one is to compare them, as both decrease by approximately 1.5% (initial dataset) and 2% (extended dataset) when compared to the benchmark.

Model	RMSE	MAPE	R^2	BIC
benchmark	0.0566	0.6310	0.9762	-7758
final	0.0554	0.6189	0.9767	-7734
no AR	0.2123	2.6495	0.6587	-556
AR1	0.0184	0.1911	0.9975	-14200
XGBoost	0.07031	0.7419	-	-

Table 14: Forecasting Accuracy Extended Dataset

Nevertheless, this is a statistical indicator that the addition of climate variables helps with forecasting. Conversely, BIC increases, but the variation lacks interpretability due to the construction of the criterion. This suggests that the added benefit of the extra variables is smaller than the burden of having a more complex model with more variables.

Looking at the findings from the initial dataset, it is clear that the presence of drought is significant for soybean price, even if the impact is not significant in terms of size of change. Generally, Drought tends to increase the price, while Drought_lag tends to decrease it. Possible explanations for this could be the fact that by the time summer arrives, the information that there is going to be a drought is already seen in the price. This hypothesis is also backed by the fact that the net impact of Drought and Drought_lag is small in regressions where the variable log_Adjusted.soy exists, as it is an exponent of autoregression (tables 5 and 10).

season	initial + AR	extended + AR	initial no AR	extended no AR
Spring	0.1	1.5	0.4	5.2
Summer	-0.2	1.1	11.8	17.6
Autumn	-3.2	-4.6	-4.6	-8.8

Table 15: Drought Net Impact in Percentages

The values for the drought effect in winter (SeasonDJF:Drought_lag) are excluded from analysis. They are present for completeness and are a result of the way Drought_lag was constructed. One can only deduce from the fact that they do not pass the significance threshold that the presence of droughts in spring and summer does not impact the price throughout the winter. Table 15 leads to the fact that the presence of drought tends to increase the price in spring, but decrease it in autumn. The impact on the summer prices is quite large if there is no AR component, and this should hint towards the fact that soy is more expensive during the summer if droughts are present.

In the extended database, the presence of the Wildfire variable tends to decrease the price by 1.3% (Table 10) and 5% (Table 11). This information should be treated with caution, as there are 804 days of wildfire, but they are derived from only 87 wildfire event, out of which only a fifth happen in areas with intense soy cultivation. Similarly, storms have a marginal effect on the price in the case of the extended dataset, where the price decrease by 0.5%, alongside a marginal statistical relevance of 8.5%. Cyclones were included, even if they are more intense in the south of the country, because of their reach and intensity, whereas tornadoes and blizzards are not infrequent events in the North Midlands. There were a total of 670 storms during the time interval of analysis, amounting to 987 days with storm events.

The variables added in the extended database show moderate impact and significance. The intrinsic change brought by a raise of 10^{o}C in daily temperature leads to a decrease of 0.4% (Table 10) or an increase of 0.3% (Table 11), with the latter one being statistically insignificant. The coefficient for Precipitation is opposed to the one for Max.Temperature, which compounds the effect (hotter days tend to have less rain). The effect from Precipitation tends to be small, as there are few observations that go beyond two standard deviations from the mean (Precipitation > 10 mm/day), and they have an uneven distribution in days per season (DJF-31, MAM-75, JJA-32, SON-47). Conclusions cannot be drawn from either of these variables, as the presence of the autoregressive component through log_Adjusted.soy reverses the direction of the sign, which hints that the effect is not significant enough.

The days of the working weeks were also analysed in the first steps of modelling. There was no model that had them statistically relevant at 5% p-value, including interaction between the day and other variables, and they were excluded from output due to the volume they would occupy in the tables. Volatility is considered to be higher on Mondays [19], with one reason for this being the impact of weekend news, but this had no effect on the segmentation of trading volume for either soy or S&P, nor on their prices. Seasons show minimal effect on the price and the alternating signs render them in a similar situation as the climate variables. While statistically relevant in benchmark models, the signs alternate and the size of the coefficients do not suggest any definitive conclusion regarding how prices should vary between seasons.

The trading volume of soy is highly significant and relevant, where a 1% increase in volume translates into a 0.4-0.5% increase in price in models with AR elements. Without them, the effect increases by an order of magnitude to 8.6% (Table 6) and 4.1% (Table 11), respectively. Conversely, the S&P Index has opposite patterns depending on AR components, with the models in the extended database even rendering it statistically insignificant (Tables 8 and 10). Thus, we reach no conclusion as to the effect it has on price. The fact that these variables hold the highest forecasting power is further underlined by the results of the XGBoost model, which identifies the volumes as strong predictors in both instances of the dataset.

7 CONCLUSION

The paper has analysed the way climate variables affect the price of soybean futures. Results from several models have shown mixed results regarding such variables. By far, the highest predictor is the autoregressive term, followed by information from the field of finance (S&P price index and trading volumes). While statistically significant in most models, climate variables have coefficients of highly alternating size and sign, which does not allow for a definitive interpretability of their effect. Drought is the climate extreme that consistently has the highest importance, but the effect it has on price is, in general, either minuscule, or fluctuating. There is some evidence that droughts decrease prices during the autumn, with slight evidence that they are increased during summer.

Future Work

The paper at hand has natural extensions in several dimensions. The only country analysed was USA, but several other countries have industries with soybean futures, Brazil and Argentina being some of them. A global analysis of climate impact on soybean could follow the same steps, such that a more unified view is obtained. Moreover, future research could be done for other crop futures, without any significant change in database building or other specifics. For example, corn, maize, rice and wheat could be great examples of that, and the analysis could start with USA, to follow the same pattern, and then extend to several other countries who have significant crop futures industries. Such extensions might take into account information about the commercial status between countries, alongside general economic data and propensity to produce and consume soy.

With regards to statistical modelling, there was an attempt to use the GARCH model [15], given that the log-price is clearly heteroskedastic. GARCH would allow for a varying volatility of the error term and would be a better theoretical fit for the data at hand. Significant issues were found with regards to inverting the Hessian matrix, but future research could find new computational ways of obtaining GARCH results. Additional variables of interest could be the ENSO variation, as their integration within the current paper could start from an existing article [16], with the addition of contemporary research on the topic.

REFERENCES

[1] American Soybean Association. 2020. *SOYSTATS*. http://soystats.com/

[2] James Board. 2013. A comprehensive survey of international soybean research: genetics, physiology, agronomy and nitrogen relationships. (2013).

[3] M Bouslama and WT Schapaugh Jr. 1984. Stress tolerance in soybeans. I. Evaluation of three screening techniques for heat and drought tolerance 1. *Crop science* 24, 5 (1984), 933–937.

[4] Université catholique de Louvain (UCL). 2021. *EM-DAT: The Emergency Events Database*. https://www.emdat.be/

[5] Alexandre José Cattelan and Amelio Dall'Agnol. 2018. The rapid soybean growth in Brazil. *Embrapa Soja-Artigo em periódico indexado (ALICE)* (2018).

[6] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, et al. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2* 1, 4 (2015), 1–4.

[7] Wenbin Chen, Kun Fu, Jiawei Zuo, Xinwei Zheng, Tinglei Huang, and Wenjuan Ren. 2017. Radar emitter classification for large data set based on weighted-xgboost. *IET radar, sonar & navigation* 11, 8 (2017), 1203–1207.

[8] DL Dornbos and RE Mullen. 1992. Soybean seed protein and oil contents and fatty acid composition adjustments by drought and temperature. *Journal of the American Oil Chemists Society* 69, 3 (1992), 228–231.

[9] DL Dornbos Jr and RE Mullen. 1991. Influence of stress during soybean seed fill on seed weight, germination, and seedling growth rate. *Canadian Journal of Plant Science* 71, 2 (1991), 373–383.

[10] The Food and Agriculture Organization of the United Nations. 2020. *The future of food and agriculture*. http://www.fao.org/3/i6583e/i6583e.pdf

[11] The Food and Agriculture Organization of the United Nations. 2020. *World Food and Agriculture 2020*. http://www.fao.org/3/cb1329en/online/cb1329en.html#

[12] National Centers for Environmental Prediction (NCEP). 2021. *Climate Forecast System Reanalysis (CFSR)*. https://globalweather.tamu.edu/

[13] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics* 28, 2 (2000), 337–407.

[14] Hélyette Geman. 2009. *Commodities and commodity derivatives: modeling and pricing for agriculturals, metals and energy*. John Wiley & Sons.

[15] Lajos Horv, Piotr Kokoszka, et al. 2003. GARCH processes: structure and estimation. *Bernoulli* 9, 2 (2003), 201–227.

[16] Christian L Keppenne. 1995. An ENSO signal in soybean futures prices. *Journal of Climate* 8, 6 (1995), 1685–1689.

[17] Yee-Shan Ku, Wan-Kin Au-Yeung, Yuk-Lin Yung, Man-Wah Li, Chao-Qing Wen, Xueyi Liu, and Hon-Ming Lam. 2013. Drought stress and tolerance in soybean. *A comprehensive survey of international soybean research—Genetics, physiology, agronomy and nitrogen relationships* (2013), 209–237.

[18] Spyros Makridakis. 1993. Accuracy measures: theoretical and practical concerns. *International journal of forecasting* 9, 4 (1993), 527–529.

[19] Xiaoyi Mu. 2007. Weather, storage, and natural gas price dynamics: Fundamentals and volatility. *Energy Economics* 29, 1 (2007), 46–63.

[20] Alexey Natekin and Alois Knoll. 2013. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics* 7 (2013), 21.

[21] United States. Department of Agriculture. Statistical Reporting Service. 1984. *Usual planting and harvesting dates for US field crops*. Number 628. US Department of Agriculture, Statistical Reporting Service.

[22] José L Rotundo and Mark E Westgate. 2009. Meta-analysis of environmental effects on soybean seed composition. *Field Crops Research* 110,

2 (2009), 147–156.
[23] Preeti Singh, R Kumar, SN Sabapathy, and AS Bawa. 2008. Functional and edible uses of soy protein products. *Comprehensive reviews in food science and food safety* 7, 1 (2008), 14–28.
[24] M Thrane, PV Paulsen, MW Orcutt, and TM Krieger. 2017. Soy protein: Impacts, production, and applications. In *Sustainable protein sources*. Elsevier, 23–45.
[25] ZHAO Wenpeng, LI Jincai, ZHAO Jun, ZHAO Dandan, LU Jingze, and WANG Xiang. 2020. XGB Model: Research on Evaporation Duct Height Prediction Based on XGBoost Algorithm. *Radioengineering* 29, 1 (2020).