Master of Data Science Capstone Project - Part 1 Report
Machine Learning Estimation of the Future Climate Risk Amplification of Food Security-Induced Conflict

Xiangyi He[1], Ritwik Giri[2], Jihang Yu[3], Jiaqi Hu[4], & Sunchuangyu Huang[5]

October 22, 2023

[1]Email: xiangyih1@student.unimelb.edu.au
[2]Email: rgiri@student.unimelb.edu.au
[3]Email: jihangy@student.unimelb.edu.au
[4]Email: jiaqih7@student.unimelb.edu.au
[5]Email: sunchuangyuh@student.unimelb.edu.au

Introduction

Nowadays, with the growing population and the increasing demand for sustainable development around the globe, food security and sustainability have become crucial topics. Meanwhile, climate change and variability have been aspects that are constantly monitored by human society. Thus, we are keen to visualize this chain of influence and discover the underlying correlations between food security and climate variability.

Climate variability is a crucial factor in the worldwide food security issue, for example, the food insecurity in the 2007 Lesotho-South Africa drought [16] and the annual crop failure risks for maize and winter wheat from 1960 to 2016 in the United States [14]. It impacts food security in many aspects, including population, agricultural production, market prices, and trade relationships between countries.
Therefore, understanding the correlations between climate and food security is essential for improving assessments and proactive management of climate-related impacts. It becomes crucial in a future scenario characterized by a growing population and limited resources.

Climate variability risks escalating food prices. Adverse weather conditions could disrupt crop growth cycles, making them unpredictable [15]. According to Bartos [2], changes in temperature and precipitation patterns may reduce the nutritional value of basic grains such as rice and wheat. Furthermore, pests and diseases may spread because of climate variability. All these factors above will consequently impact the yield of agricultural products. A decline in yield will result in expensive commodities and increase transportation costs [6].

As a result, this study aims to investigate and model the impact of climate variability on commodity prices using time series with ENSO as an exogenous factor for climate variability. One probable challenge is determining the appropriate lag between the ENSO and the commodity prices.

Related work

Climate variability's impact on global agriculture has become an important research topic. It affects food security, including population, agricultural production, market prices, and trade relations between countries. Food prices directly impact consumers, affecting their purchasing power and access to nutritious food.

An ARIMA (1,2,1) model was used by Priyanga [13] to forecast 2019 coconut oil prices in Kerala's market based on monthly price data from 2008 to 2018. They selected models with the lowest Akaike Information Criterion (AIC) and Schwarz Information Criterion (SIC) values. To ensure a good fit, the residuals of chosen models were confirmed as white noise using autocorrelation (ACF) and partial autocorrelation (PACF) functions [13]. This report will likewise utilize the ARIMA model for coconut oil price forecasting, with further details to follow.

Wheat is an essential source of food nutrition. It is the main source of vegetable protein and the source of about 20% of the world's human calorie consumption [7]. Monitoring wheat production and prices

is crucial for ensuring an adequate supply of wheat. Jadhav, CHINNAPPA, and Gaddi [10] forecasted rice, Ragi, and maize prices in Karnataka from 2002 to 2016 by the ARIMA model. They determined that ARIMA was effective in predicting variable magnitudes, but it required substantial sample sizes.

Gutierrez (2017) and Gutierrez, Piras, and Olmeo [8] both utilized Global Vector Autoregressive (GVAR) models to predict wheat prices in six key export regions, including Australia. Gutierrez [7] studied ENSO's impact on the global wheat market, particularly in Australia, affecting prices, production, and exports. They developed a model for wheat prices, considering supply and demand factors, exchange rates, and oil prices, and evaluating its accuracy using RMSFE and MAPE statistics.

In India, a major wheat producer, Hemant Sharma (2015), examines the effectiveness of the ARIMA model in predicting wheat prices in Rajasthan. The author tunes different seasonal parameters in the ARIMA model and then examines models based on criteria like AIC, SBC, root mean square error (RMSE), mean absolute deviation (MAD), and mean absolute percentage error (MAPE). The best model with the most appropriate parameter was ARIMA(1,1,1).

Darekar and Reddy [4] also employed the ARIMA model to forecast India's future wheat prices, using monthly data from January 2006 to June 2017. Their findings identified the ARIMA(0,1,1)(0,1,1) model as the most suitable for such predictions. The model's accuracy was confirmed by evaluating the RMSE and MAPE.

Joana Dias and Humberto Rocha [5] highlight the limitations of time series modeling for forecasting wheat prices, indicating the variability of parameters in ARIMA models. They compared five different modeling approaches in the article using the out-of-sample data: ARIMA, Classification and Regression Tree, Random Forests, Support Vector Machines, and Multivariate Adaptive Regression Splines. Abilities of one to six months of forecasting are evaluated using four methods which also provide guidelines for this research as criteria: MAE, MaxAE, RMSE, and Accuracy in Trend. Although different models perform differently under each forecasting time scale, SVM and MARS present better results and consistent results.

In Kitsios et al. [11], it is clearly shown that the Multivariate ENSO factor has a significant effect on coconut oil prices. When comparing AR models with and without the ENSO factor, models with the ENSO factor outperform those without. Therefore, ENSO, as a specific climate variability, guides the further direction of the study.

Weston Anderson [1] and Kyungsik Nam [12] provide detailed insights into the relationship between the Trans-Pacific ENSO and maize, wheat, and soy. They express in a geographical perspective what El Nino and La Nina are and how they impact the production of maize, wheat, and soy in different regions around the globe. It will be helpful when it comes to understanding the fundamentals of this research project. The impact of El Niño events and natural variability on rice agriculture in 2050 under climate change conditions was noted by Naylor, Battisti, Vimont, Falcon, and Burke (2007). The study shows that reduced rainfall, increased temperature, and carbon dioxide concentration due to climate change

could impact rice yields.

Data analysis and preliminary model development

The data was provided by our client and was received in the form of a zipped folder. The received data was divided into two parts: Commodity Price information and Climate teleconnection files.

World Bank Monthly Commodities Price: Monthly information about the prices for different commodities. The data spans from January 1960 to February 2020. The data we received was clean and contained no-null values for the columns of interest. A heat map was generated to analyse the same.

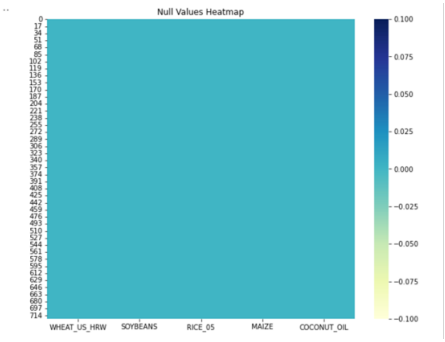


Figure 1: Heatmap of Null Values in Selected Columns

We convert the raw prices to log returns as they can help stabilise the time series variance.

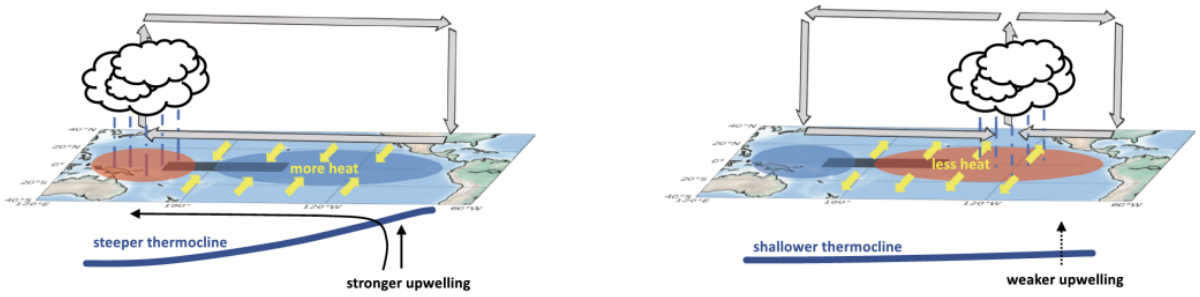
Climate Teleconnection Files: The information provides a brief explanation of climate teleconnections. The first column in all the files contains the date. The other column contains the teleconnection data itself.

All the data is recorded monthly from January 1958 to December 2018. For our initial research, we use Multivariate ENSO data and find its trends with the prices of the above-listed commodities. In this section, we use the dataset between Jan 1980 and Dec 2018 for our exploratory purposes.

El Niño-Southern Oscillation – The Phenomenon and Observations

ENSO (El Niño-Southern Oscillation) is a naturally occurring climate mode of variability that causes multiyear variations in sea-surface temperature and winds over the tropical Pacific Ocean. It is a natural climate variability phenomenon. It can lead to changes in atmospheric circulation patterns and wind speeds, which can in turn impact weather patterns around the world. It comprises the mature phases La Niña and El Niño, and the neutral phase.

During La Niña, there are cooler than average waters in the eastern Pacific and warmer than average waters in the west, which is accompanied by enhanced evaporation. During El Niño, the opposite occurs, with warm waters in the east and cool waters in the west. ENSO can affect rainfall and temperature in distant regions of the Earth, with the strongest impacts in the tropics. The phases of ENSO have heterogeneous impacts on rainfall, temperature, and potential economic activity. Detailed information about the phases is described in the figure shown.



(a) La Niña Phase

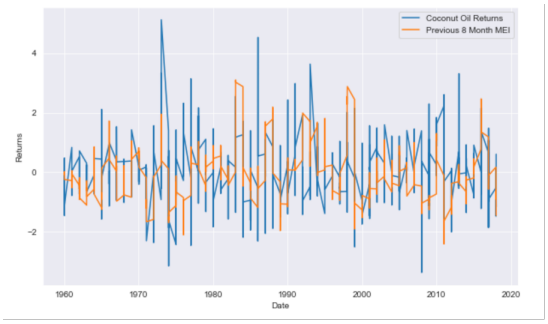
(b) El Niño Phase

Figure 2: Atmospheric and oceanic structure during phases of El Niño Southern Oscillation (ENSO) (Kitsios et al.[11])

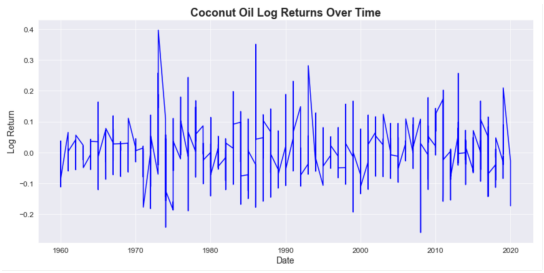
We can calculate the phase and magnitude of ENSO using indices that are derived from sea surface temperature measurements in the equatorial Pacific Ocean. In this project, using the Niño4 index as it is useful in assessing the potential impacts of climate change, as well as for predicting ENSO events on longer timescales than the historical record provides. The positive values of the Niño4 index indicate warmer than average conditions (El Niño) and negative values indicate cooler than average conditions (La Niña) in the eastern tropical Pacific.

El Niño-Southern Oscillation and Coconut Oil

We find the relationship between the previous month's ENSO values and the log return of Coconut Oil. It is clear from the chart that there is a negative correlation between the two factors.



(a) Correlation between the 8-month lagged value of MEI and Coconut oil log returns.



(b) Log Returns of Coconut Oil Price

Figure 3: Relations between ENSO factor and Coconut Oil Log Returns

We have used the log returns of the commodities and the ENSO index as the input variables. It can make the model more interpretable, as they represent the percentage change in the variable rather than the absolute change. They are used in AR models because they stabilize time series variance. As the time series level increases, raw return variance tends to grow, posing challenges for stable AR model fitting.

El Niño-Southern Oscillation with Other Agricultural Commodities

Based on a previous study that found a negative correlation between ENSO factors and the log returns of commodities like coconut oil, we investigated whether this approach could be applied to other agricultural commodities. Our study focused on WHEAT_US_SRW, RICE_05, MAIZE, and SOYBEANS, and we constructed two comparison AR models to address:

- The significance of ENSO's impact on the log return of these crops.
- The potential improvement of model performance by incorporating more climate features.

Our analysis began by examining the correlation between the four commodities and all climate teleconnection features from the JRA55 dataset, resulting in a correlation plot (Figure 4). This plot revealed a weak, negative correlation between the commodities' log returns and the ENSO factor, consistent with our previous findings. The log returns also showed weak or extremely weak correlations with other teleconnection features, such as the Arctic Oscillation (AO) index and the Southern Annular Mode (SAM) index.

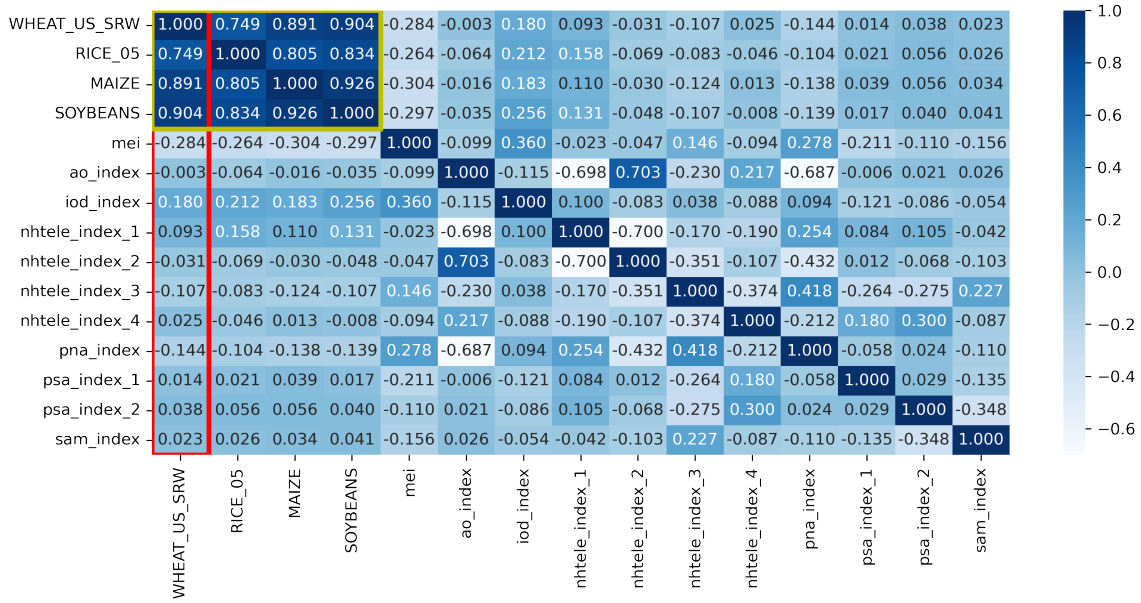


Figure 4: Correlation between WHEAT_US_SRW, RICE, MAIZE log returns, and other JRA55 features

Additionally, we observed that teleconnection features might exhibit interactions due to plausible seasonal factors and the complexity of the climate system, implying that JRA55 features may not be independent and identically distributed. Moreover, economic factors impact log returns, as seen by the strong correlation among the prices of these agricultural commodities within the yellow box.

Based on these observations, we built two models using log returns as the response variable: a single ENSO factor model and a full model including all JRA55 features. Given the potential unknown relationships between teleconnection features observed in Figure 4, both models may yield sub-optimal inference results. We will use the Root Mean Absolute Error (RMAE) and Bayesian Information Criterion (BIC) as evaluation metrics to determine the best model, serving as the baseline for future research.

Auto Regressive Commodity Log Return Forecasting Model Selection

To deploy the AR model, we utilize the `SARIMA` method from the `statsmodels` library. The `SARIMA` method allows us to adjust lags based on non-seasonal or seasonal order. Hyperparameter setting is crucial for model deployment. According to Kitsios et al.[11], ENSO may have an impact after 8 months. Using a 1-month lag in the `SARIMA` model would yield poor inference results, as ENSO requires time to create impacts. Therefore, in this project, we employ an 8-month lag for both seasonal and non-seasonal order in model deployment. The model will be trained on data from January 1979 to March 2012. After building the models, we generate inferences for the next 82 months from April 2012 and evaluate the performance using RMSE and BIC. Single ENSO factor model will use the fixed hyperparameter setting and full model will adjust parameters based on commodities accordingly. The results are presented below.

Model	Commodity	Order	Seasonal Order	BIC	RMAE
Single ENSO (MEI)	WHEAT_US_SRW	(6,1,0)	(6,1,2,12)	-1124.849	0.220
	RICE_05	(6,1,0)	(6,1,2,12)	-1307.106	0.235
	MAIZE	(6,1,0)	(6,1,2,12)	-1248.832	0.232
	SOYBEANS	(6,1,0)	(6,1,2,12)	-1330.029	0.227
Full Model	WHEAT_US_SRW	(8,2,2)	(8,1,1,12)	-731.767	0.237
	RICE_05	(8,1,3)	(8,1,1,12)	-917.000	0.240
	MAIZE	(8,2,5)	(8,1,1,12)	-841.786	0.255
	SOYBEANS	(8,1,4)	(8,2,1,12)	-866.585	0.272

Table 1: SARIMAX evaluation metrics between the single MEI model and the full model.

Both approaches result in an RMSE of approximately 0.2; however, the method incorporating all JRA55 features increases the BIC scores by around 300. In theory, a model with a lower BIC suggests superior inference capabilities. Nevertheless, as shown in figure 5, utilizing a single ENSO factor leads to flat lines across the log returns of the four commodities. Interestingly, rice and maize exhibit some trends that align with the actual log returns, but they appear to be missing fluctuations or seasonal patterns.

On the other hand, there is some delay observed in figure 6, the full model captures trends more effectively and corresponds more closely to the true values, albeit with a higher RMAE score. We previously noted that hyperparameter settings significantly impact model performance. Here, we conclude that a lag of 6 to 8 months, which matches ENSO patterns, does influence crop production and log returns. It appears that even with the optimal hyperparameter setting for the single ENSO model, there is no guarantee that the AR model will produce reasonable inferences. For instance, regardless of the parameter tuning for the wheat model, it continues to yield a relatively flat inference plot. Therefore, despite the full model having a higher BIC value and a slightly increased RMAE, we still favour it over the single ENSO model.

Incorporating ENSO alone does not significantly improve log return predictive performance. In our experiment, a comprehensive model using all factors leads to enhanced model capabilities. Although both models display similar RMSE values, the comprehensive model reveals a more precise seasonal pattern that aligns better with the true value. This may be affected by the model hyperparameter settings due to the unknown connection of lagged values in the past. Additionally, underlying dependent relationships also impact performance, such as interactions or features irrelevant to log return inference.

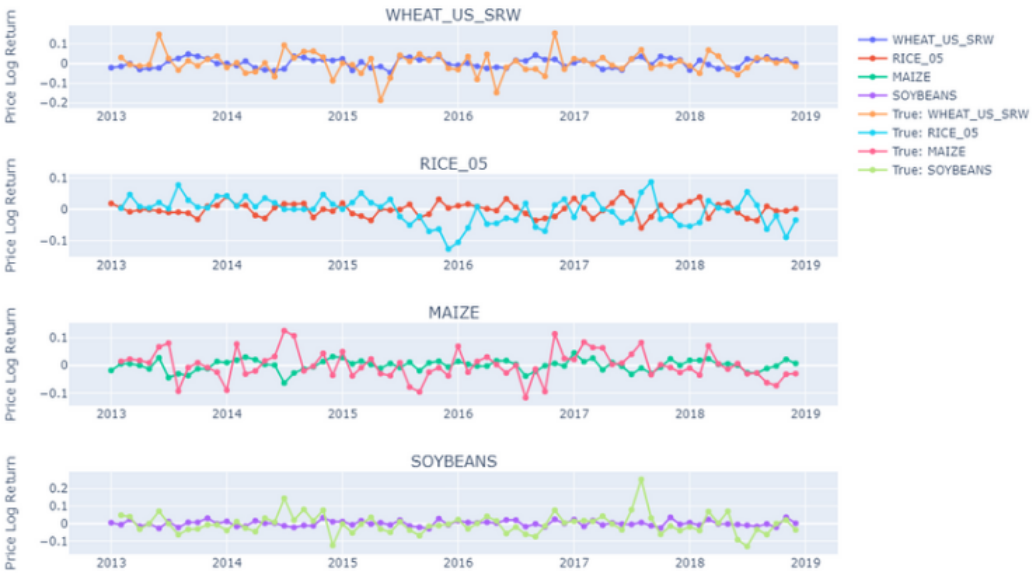


Figure 5: SARIMAX model with single ENSO (MEI) feature from JRA55

Furthermore, climate exhibits some Bayesian dependencies, as today's events will affect tomorrow. At the current stage, we will consider the comprehensive model as a baseline model; any future deployed models, such as dynamic Bayesian networks, must outperform this baseline.

Enhanced Commodity Forecasting: Incorporating Inflation and Validating Model Assumptions

In the previous section, we established a baseline model based on our initial understanding of time series analysis and the insights gained from our exploratory data analysis (EDA). However, to enhance the accuracy of our model, it is imperative to consider additional factors that may impact commodity prices beyond the influence of the ENSO indexes. One such influential factor is inflation, which has been recognized for its effect on commodity prices.

To incorporate the influence of inflation, we will extend our model using the AutoRegressive (AR) framework proposed by Kitsios et al[11]. In this extended model, denoted as formula 1.

$$p(t) = a^{pp} + \sum_{l=1}^{11} a_l^{pp} D_l(t) + \sum_{K \in \mathbb{I}^{pp}} b_k^{pp} p(t-k) + \sum_{K \in \mathbb{I}^{pE}} b_k^{pE} E(t-k) \tag{1}$$

$$p(t) = (1 + i(t)) \times (1 + \tilde{p}(t)) - 1 \tag{2}$$

$$\tilde{p}(t) = \log(P(t)/P(t-1)) \tag{3}$$

$$i(t) = \log(C(t)/C(t-1)) \tag{4}$$

We will include real commodity log returns (formula 2), nominal rates (formula 3), and the inflation rate (formula 4) as exogenous variables. By incorporating inflation to normalise the endogenous variable (commodity log return) to remove the influcen of economic growth, our aim is to capture its potential impact on commodity prices alongside the ENSO indexes. Additionally, we will judiciously select the appropriate lags of the log returns as endogenous variables and incorporate relevant lags of the ENSO

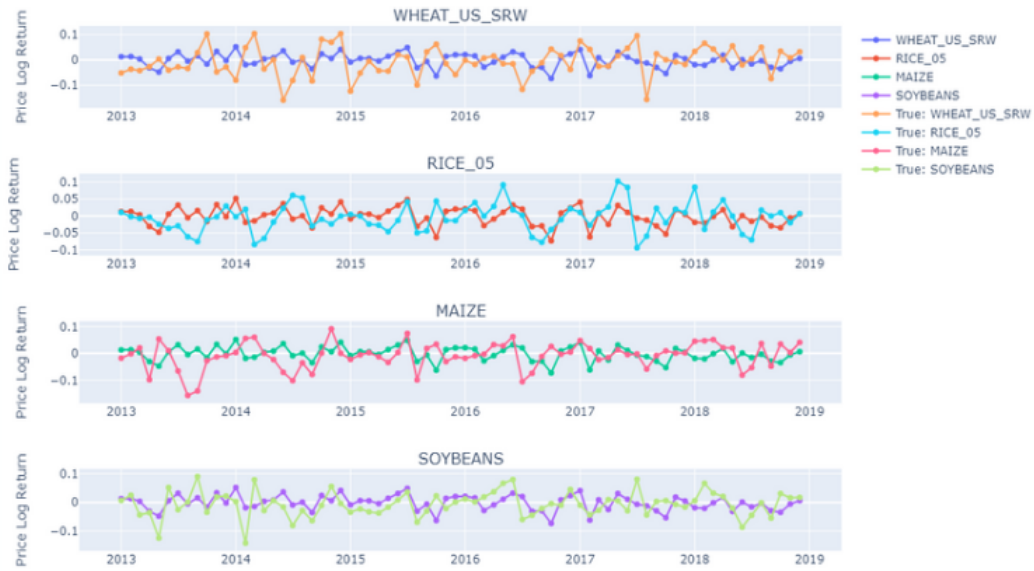


Figure 6: SARIMAX model with AO, IOD, NHTELE, PNA, PSA and SAM index from JRA55

indexes as exogenous variables.

To ensure the validity and reliability of our models, we must rigorously assess them for serial correlation and heteroskedasticity. This step is pivotal in identifying the optimal model, as it guarantees that the residuals of the model possess the desirable characteristics of independence and identical distribution. By consider the presence of serial correlation and heteroskedasticity, we can refine and validate our model, thereby enhancing its forecasting capabilities.

Project Timeline

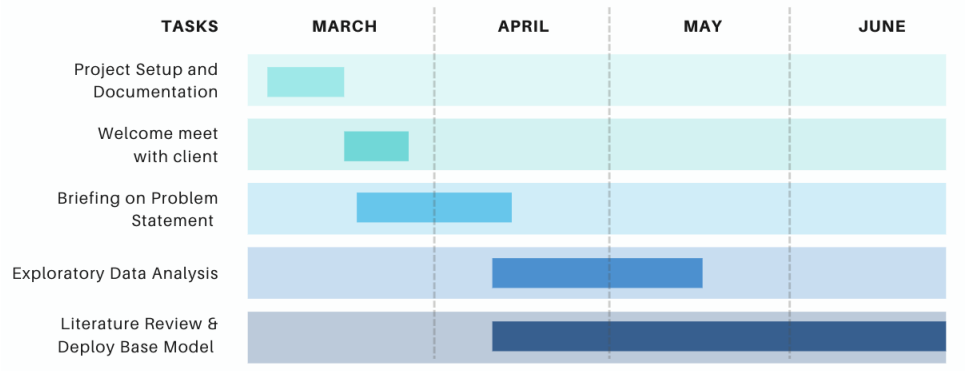


Figure 7: Pipeline Activities - (March to June)

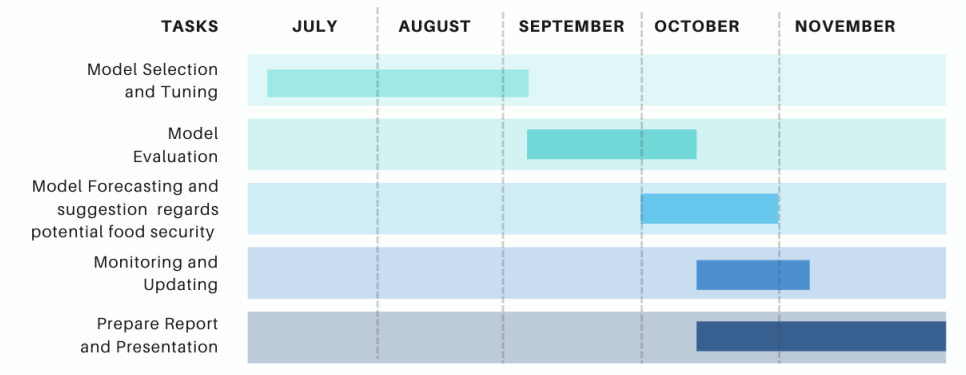


Figure 8: Pipeline Activities - (July to November)

Proposal for semester 2

In our current project, we're working to understand how different climate teleconnections can change the prices of commodities like coconut oils, wheat, rice, maize, and soybean.

Our starting point was a research paper given to us by the client. This paper showed how coconut oil prices can be influenced by ENSO (El Niño-Southern Oscillation) factors. We used this information to perform an exploratory data analysis (EDA). Our goal by the end of June is to create a basic model for predicting commodity prices, drawing from the techniques described in papers by Kitsios [11] and Priyanga [13].

Once we have this basic model, we'll aim to improve it by adjusting its settings, a method used by Darekar [4]. This will help us see how the model behaves with different settings. If our adjustments improve the model, we'll then start to explore other methods and data models that could work even better.

We're interested in some models that use a Bayesian approach, such as Dynamic Bayesian Networks, Bayesian Structural Time Series, and Long Short-Term Memory. These models, which are also used by CSIRO [3], should help us get more reliable results. We also think there might be unknown connections between different weather variables - 'Granger causal relationships'. Dylan and Terence's work [9] suggests these relationships are important when studying climate patterns. We will consider this when we work on improving our model.

After we've built and fine-tuned our model, we will test it and interpret the results. If all goes well, we should be able to use our model to make predictions about food security by October. We will then monitor the prices of commodities and weather patterns to check if our model's predictions are correct. We plan to compile all our findings into a report and a presentation, which we aim to finish by the end of November.

Looking ahead, we want to find ways to improve our predictions and spot any weaknesses in our current model. We might be able to improve our SARIMA model by changing certain parameters, like the loss function, moving average order, and autoregressive order. This could help us make better predictions.

Team member information

Team Member	Roles
Jiaqi Hu	Data Analyst
Jihang Yu	Data Analyst
Ritwik Giri	Data Analyst, Project Coordinator, Meeting Facilitator, Scribe
Sunchuangyu Huang	Data Analyst, Documentation Specialist, Meeting Organizer
Xiangyi He	Data Analyst, Meeting Organizer, Scribe

Table 2: Team Roles

References

[1] William Anderson et al. "Trans-Pacific ENSO teleconnections pose a correlated risk to agriculture". In: *Agricultural and Forest Meteorology* 262 (2018), pp. 298–309.

[2] Stephen Bartos. *IMPACTS OF CLIMATE CHANGE ON OUR FOOD SUPPLY Fork in the Road A report on current and growing risks and vulnerabilities in Australia's food supply chain arising from climate change*. 2022. URL: https://farmersforclimateaction.org.au/wp-content/uploads/2022/03/Fork-in-the-Road_V5.pdf.

[3] CSIRO. *Causal Inference in Complex Multiscale Systems*. https://research.csiro.au/ai4m/causal-inference-in-complex-multiscale-systems/. Accessed 2023.

[4] Aditya Darekar and A. Ashok Reddy. "Forecasting wheat prices in India". In: *Wheat and Barley Research* 10.1 (2018), pp. 33–39.

[5] João Dias and Helder Rocha. "Forecasting wheat prices based on past behavior: comparison of different modelling approaches". In: *Computational Science and Its Applications–ICCSA 2019*. Vol. 19. Springer. 2019, pp. 167–182.

[6] Global Citizen. *4 ways climate change is affecting food security right now*. 2015. URL: https://www.globalcitizen.org/en/content/4-ways-climate-change-is-affecting-food-security-r/ (visited on 05/19/2023).

[7] Luciano Gutierrez. "Impacts of El Niño-Southern Oscillation on the wheat market: A global dynamic analysis". In: *PloS One* 12.6 (2017), e0179086.

[8] Luciano Gutierrez, Francesco Piras, and Maria Grazia Olmeo. "Forecasting Wheat Commodity Prices using a Global Vector Autoregressive model". In: *2015 Fourth Congress, June 11-12, 2015, Ancona, Italy*. 207264. Italian Association of Agricultural and Applied Economics (AIEAA). 2015.

[9] Dylan Harries and Terence J. O'Kane. "Dynamic Bayesian Networks for Evaluation of Granger Causal Relationships in Climate Reanalyses". In: *Journal of Climate* 33.10 (2020), pp. 4239–4264. DOI: 10.1175/JCLI-D-19-0476.1.

[10] Varsha Jadhav, R. B. CHINNAPPA, and Gangadhar M. Gaddi. "Application of ARIMA model for forecasting agricultural prices". In: *International Journal of Agricultural Science and Research* 7.3 (2017), pp. 103–108.

[11] Vassili Kitsios, Lurion De Mello, and Richard Matear. "Forecasting commodity returns by exploiting climate model forecasts of the El Niño Southern Oscillation". In: *Environmental Data Science* 1 (2022), e7.

[12] Kiseok Nam. "Investigating the effect of climate uncertainty on global commodity markets". In: *Energy Economics* 96 (2021), p. 105123.

[13] V. Priyanga et al. "Forecasting coconut oil price using auto regressive integrated moving average (ARIMA) model". In: *Journal of Pharmacognosy and Phytochemistry* 8.3 (2019), pp. 2164–2169.

[14] Tyson A. Schillerberg and Di Tian. "Changes of crop failure risks in the United States associated with large-scale climate oscillations in the Atlantic and Pacific Oceans". In: *Environmental Research Letters* 15.6 (2020), p. 064035.

[15] The Nature Conservancy Australia. *Climate Change: Frequently Asked Questions*. 2020. URL: https://www.natureaustralia.org.au/what-we-do/our-priorities/climate-change/climate-change-stories/climate-change-frequently-asked-questions/ (visited on 05/19/2023).

[16] Jan Verschuur et al. "Climate change as a driver of food insecurity in the 2007 Lesotho-South Africa drought". In: *Scientific Reports* 11.1 (2021), p. 3852.