Visualizing and Predicting Agricultural Production due to Population Migration and Climate Change

Team 25: Shadi Aslebagh, Lara Kassab, Alison Lui, Emiko Sano, Sama Shrestha

EXECUTIVE SUMMARY

In the year 2020, food insecurity was already a major problem in the world with over 2.3 billion people not having food available year-round. With the projected growing population, it is important to predict whether we would be able to produce enough food to meet this demand in order to address world food insecurity in the future. In addition, climate change has had a major effect on agriculture from increased droughts, floods, and other natural disasters. In other words, there are more mouths to feed than ever and our changing climate makes feeding them a matter of global security.

Maintaining and increasing production of cash crops like rice and maize are essential to feeding this rising global population. In the next fifty years, commodity crop traders expect to rely even more on predictive models of managing crop production and food supply chains. **In our work, we seek to use predictive models to understand the dependence of global annual rice production on country-specific environmental factors including temperature, precipitation, number of natural disasters, and area used for cropland.** World leaders must leverage the extensive historical temporal and spatial data at our hands to prepare for the effects of an increasingly erratic climate on global food supply.

In this project, we used linear and nonlinear regressive models to predict global rice production based on predictive environmental variables. The variables we considered were global population, number of natural disasters experienced, number of people affected by natural disasters, average annual precipitation, average annual temperature, and hectares of land demarcated as cropland. All of our data was demarcated by country and by year. We considered 48 years of data from 1961 to 2019 and x countries or territories.

Our first model used all country data to predict global rice production in tonnes. Our second model considered only the top 14 rice producing countries, which accounted for at least 90% of global rice production in all years in our dataset. In our third model, we first created a linear regression for each year in our dataset using the top 14 producing countries. Then, we averaged each of the 48 models (one for each year of data). Our fourth model used all country and year data to predict via a GAM [acronym] nonlinear model.

Our work lays the basis for an extensive and exciting study of crop production. First, we propose the addition of several more variables to improve the accuracy of our model. Some positive additions include maximum and minimum temperature and precipitation instead of the average annual variable. Second, we propose additional work in country or region-specific models grouped by similar geographies. Third, we propose adapting our model to predict variations instead of total production in

global or national rice production. This type of model would answer questions such as if maximum temperatures continue to increase, will this most likely create oversupply (causing waste or lower prices) or undersupply (causing stress on food supply chains and potential famine).

PROJECT DESCRIPTION

Overview

Climate change poses a considerable threat to global food security. Climate extremes, food insecurity, and migration in the region are interconnected and often the relationship is complex and non-linear in nature. The world population distribution is changing with climate change impacting migration patterns. Food production is a global trade problem and as the global population distribution changes, a major impact on global trade supplies is imminent.

Problem Statement

Due to climate migration and other factors, the world population distribution is changing and being pushed away from the equator. With rising population and the increased demand for food, how are these changes going to be impacting our global food supply? Can we identify and quantify the reasons for changes in the global food supply chain? Are changes in food production and import/export practices impacted by climate change? By identifying factors impacting the food supply chain, we can generate solutions to optimize our food supply and distribution. Predicting how much regional commodity production or prices would change would help countries make plans for shifting future production and/or trading practices.

Impact

Food production is a global trade problem and as the global population distribution changes. In the 21st century, global food systems face dual challenges of increasing food demand while competing for resources such as land, water, and energy that affect food supply. We all need to eat so addressing any issues in the food supply chain is important in securing food for the future. In the context of climate change and unpredictable shocks, such as a global pandemic, the need for resiliency in global food systems has become more pressing than ever.

Audience

Predicting cash crop supply is of national security and great interest to international merchants. We anticipate an engaged audience composed of national economists, agronomists focused on effects of climate change, and farmers overseeing extensive swathes of cropland.

These organizations can use this information to make policies that prevent further food scarcity in many regions in the world. By having the predicted production information, future crop planting may be adjusted.

Extensions of our model can be more easily translated to commodity prices and then could be of interest to international merchants. Countries predicting to produce more than what is necessary for the local population may be able to negotiate trade deals with areas expected to decrease production. Additionally, these adjustments along with policies may help in making sure that future food prices do not become out of reach for any population.

DATA SOURCING AND ANALYSIS

Data sources

Table 1. Data sources and descriptions

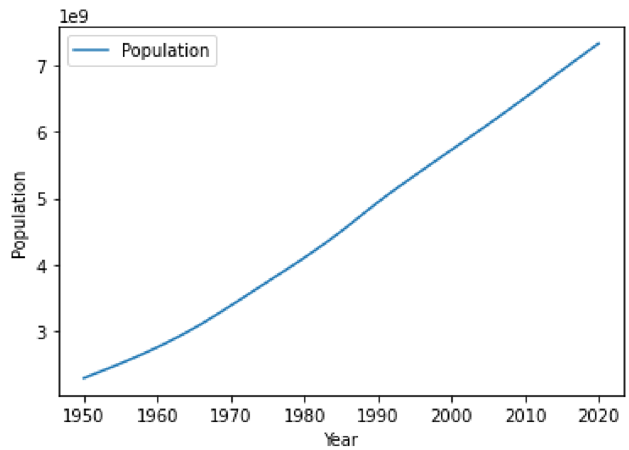
	Category	Name	Notes
1	Population	UN Data: Average annual rate of population change (percentage) Type: CSV	Annual rate of population change per country for 5 year segments from 1950-2019. Only the medium growth variant is considered.
2	Population	UN World Population Prospects: The 2019 Revision: Total population, both sexes combined (thousands)	Total population per country from 1950-2020. For years from 2021-2100, data includes prediction of annual rate of population change per country for different growth models.
3	Food Production and Trade	UN Food and Agriculture Organization (FAO) Database: Rice production per country per year by hectare Type: CSV	Total quantity of rice production per country per year (1961-2018)
4	Climate / Environment	World Bank Climate Data API Type: CSV	Annual data on precipitation (in mm) and temperature (in C°) from 1901-2012
5	Climate / Environment	UN Food and Agriculture Organization (FAO) Database: Arable Land Type: CSV	Hectares of Cropland per country per year (1961-2018)
6	Climate / Environment	UNSD Global Environment Statistics: Natural disasters Type: CSV	Number of natural disasters per country per year (1900-2020) and number of people affected. Each entry is subdivided into subcategories of disaster

Exploratory Data Analysis:

Population

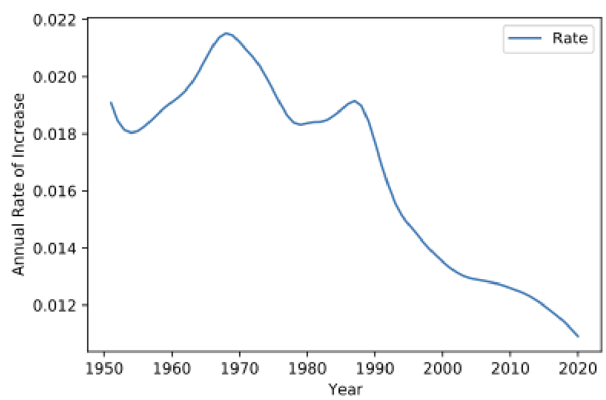
We explored the global population data from the UN database (#1 and #2 datasets). Dataset #2 had the total population per country (in thousands) for 282 countries or areas ranging from 1950 to 2020 (i.e. 71 years). We estimated the global population by adding the population of all available countries or area to visualize how the population has changed over time.

Figure 1. Population has been increasing linearly since 1950.



The population data per country was complete without any missing data. Dataset #1 contained the annual rate of population change per country in 5 year increments. Thus, to get a better picture of the year to year rate of increase, we estimated the annual rate of population increase from the total population dataset.

Figure 2. The population increase rate is declining.

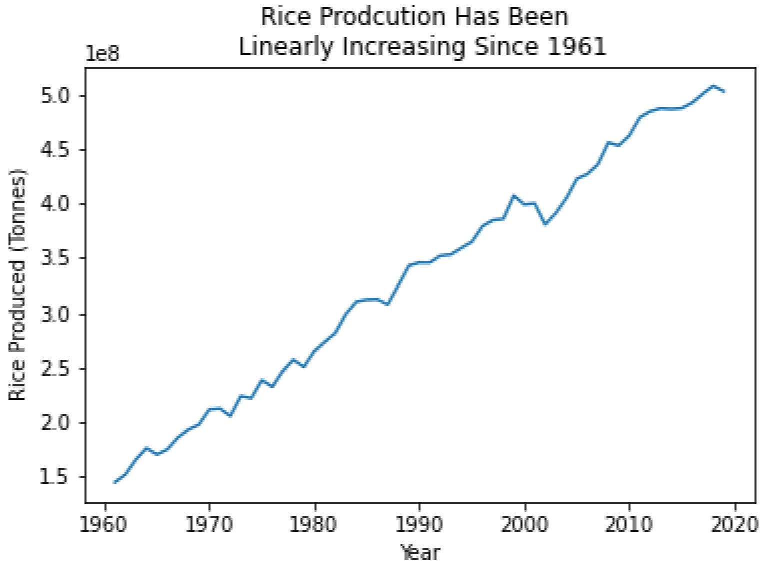


Food Production

We used rice production data (Rice, paddy (rice milled equivalent)) from the Food and Agriculture Association of the United Nations. This dataset included annual production in tonnes for 149 countries from 1961 to 2019. Out of the 7036 entries, we had rice production data for 6882 entries. For countries that no longer exist, ISO codes were mislabeled, so we assigned previously assigned codes to those countries.

Globally, rice production has been increasing linearly over time since 1961 (Figure 3).

Figure 3. Rice production has been linearly increasing since 1961.



The data was very right-skewed indicating that most countries produce less than 20 million tonnes per year. Total rice production during this time frame is also dominated by China and India accounting for 32% and 21% of the total global rice production, respectively.

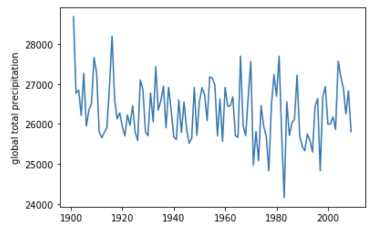
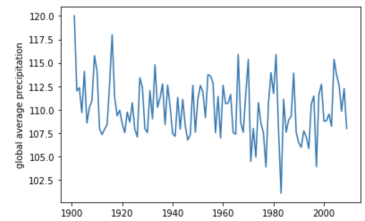
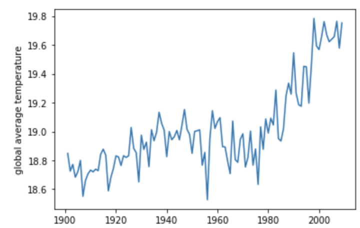
Climate data (Precipitation and Temperature)

The World Bank dataset (dataset #4) provides precipitation (in millimeters) and temperature (in Celcius) for 242 countries and territories for the years of 1901-2012.

These steps were performed in order to clean the data frame and make it analysis-ready:

- Set "year" column as an index column
- Checked how much data is missing for each country/territory (columns in the data frame) and decided to flag theses 2 countries/territories:
    - Wake Island (U.S.): 97.3% of data is missed.
    - Serbia and Montenegro: 100.0% of data is missed.
- By looking at the very end of the precipitation/temperature time series plots, we suspected there might be some missing data for some countries/territories for the last several years. We tested this hypothesis and found that for the years of 2010-2012 there are 39 (out of 242) countries/territories where no data is reported for precipitation and temperature. This is %16 of the countries/territories. Therefore, we will take extra caution for these years and decide later whether we want to include them in our analysis or we just focus on climate data before 2010. Here is the time series excluding the years after 2010.

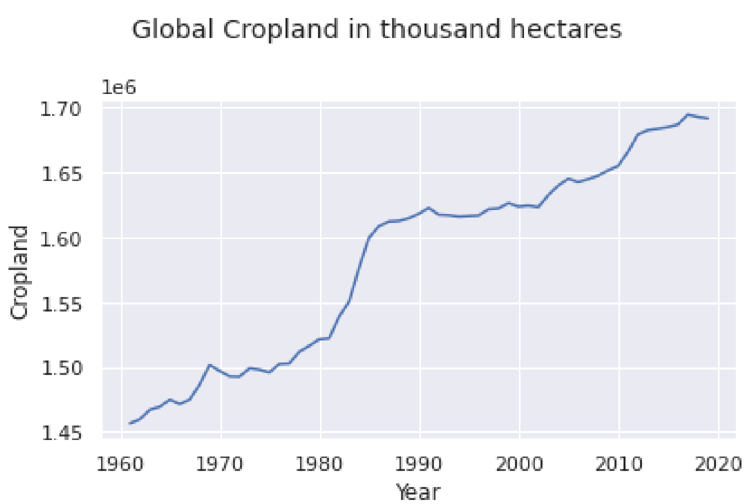
Table 2. Summary of global temperature and precipitation trends.

	total	average
precipitation		
temperature	NA	

Cropland and Country Area

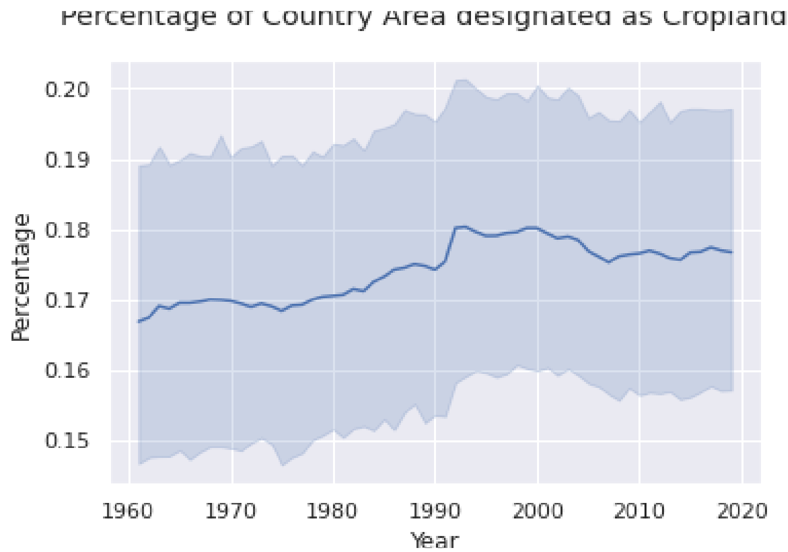
We extracted data from the Food and Agriculture Association of the United Nations quantifying the total area designated as cropland per country per year from 1961 to 2017. Cropland is defined as land used for cultivation of crops and includes all area designated for temporary crops, temporary meadows and pastures, land with temporary fallow, as well as land cultivated with long-term crops which do not have to be replanted for several years (such as cocoa and coffee), land under trees and shrubs producing flowers (such as roses and jasmine), and nurseries.

Figure 4. Available land for crop production is plateauing in recent years.



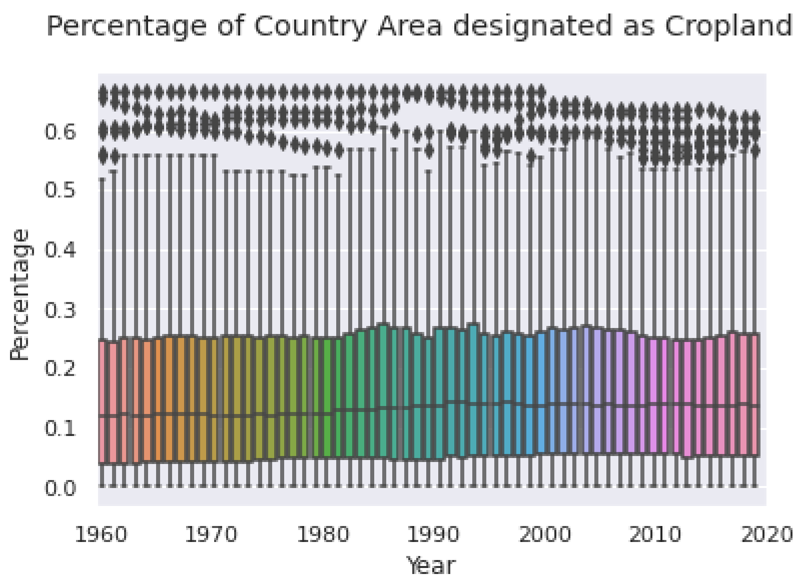
When we compute the percentage of total country area designated as cropland, we see that on average, countries are between 15% and 20% cropland.

Figure 5. Proportion of the total land dedicated to cropland has been relatively stable since 1961.



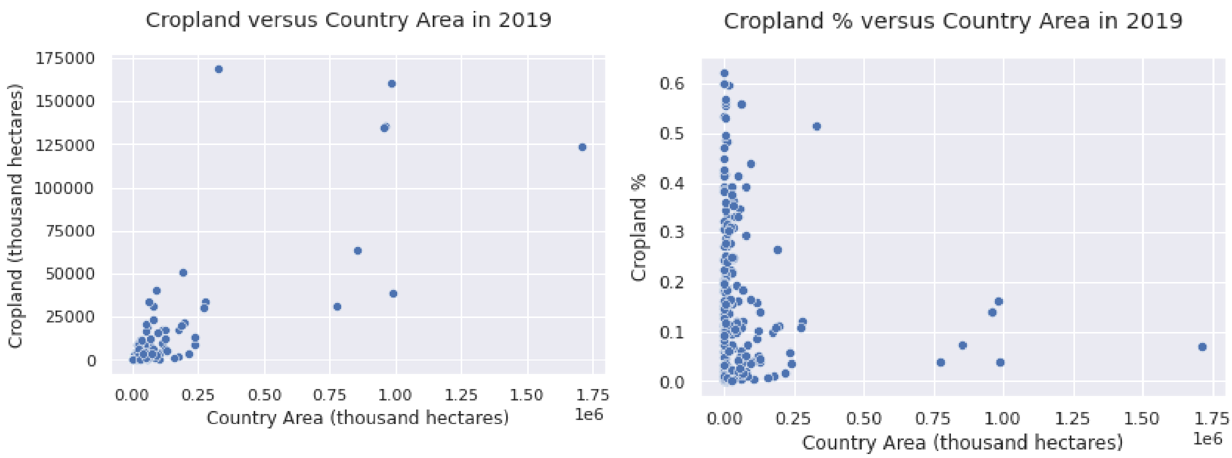
However, when we plot each country's Cropland % in our dataset individually, we see there is some variation and there are a handful of countries which have over 50% cropland.

Figure 6. Most countries dedicate about 0.1% of the total land for crop production.



In fact, when we consider just one year in our dataset and plot the area for total cropland or cropland % versus country area, we see that there does not seem to be a correlation between total cropland and total country area.

Figure 7. The size correlates to the amount of cropland but not to the percent dedicated to cropland.



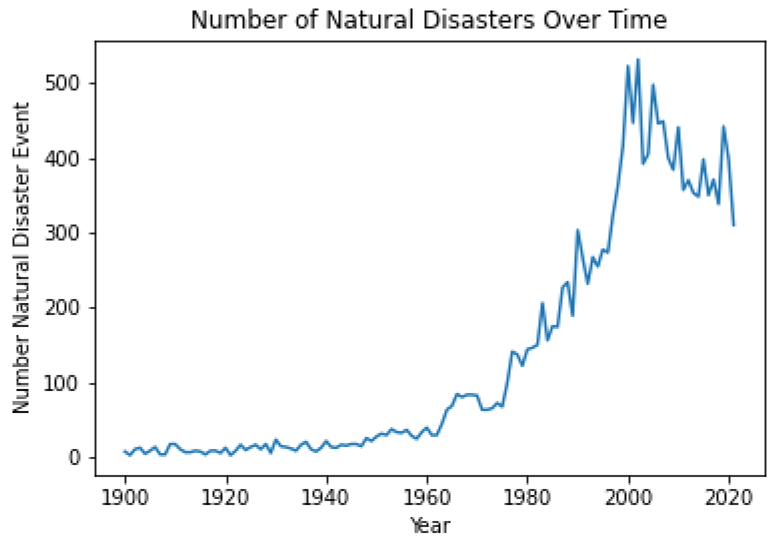
There are a few outlier countries that are significantly larger than others. These few countries have cropland % between 0.5% and 0.2%. But because of their large size, the total cropland may be a better predictor for our model.

Natural Disasters and Number Affected

We downloaded global natural disaster data from the Centre for Research on the Epidemiology of Disasters (CRED)'s Emergency Events Database (EM-DAT). We collected data on all events in the 'Natural Disasters' category. Natural disasters are divided into six different subgroups: biological (epidemic, insect infestation, and animal accident), climatological (drought, glacier lake outburst, and wildfire), extra-terrestrial (impact and space weather), geophysical (earthquake, volcanic activity, and mass movement), hydrological (flood, landslide, and wave action), and meteorological (storm, extreme temperature, and fog). The dataset also includes location (e.g. country, continent, etc.), magnitude of the event (where applicable; e.g. duration, area affected), human impact measures(e.g. the estimated number of affected people (injured, homeless, total number deaths, etc.), economic impact measures (e.g. total estimated damages), and disaster impact (e.g. sector affected). The dates range from 1900 to 2021, but it does not include COVID-19 epidemic data.

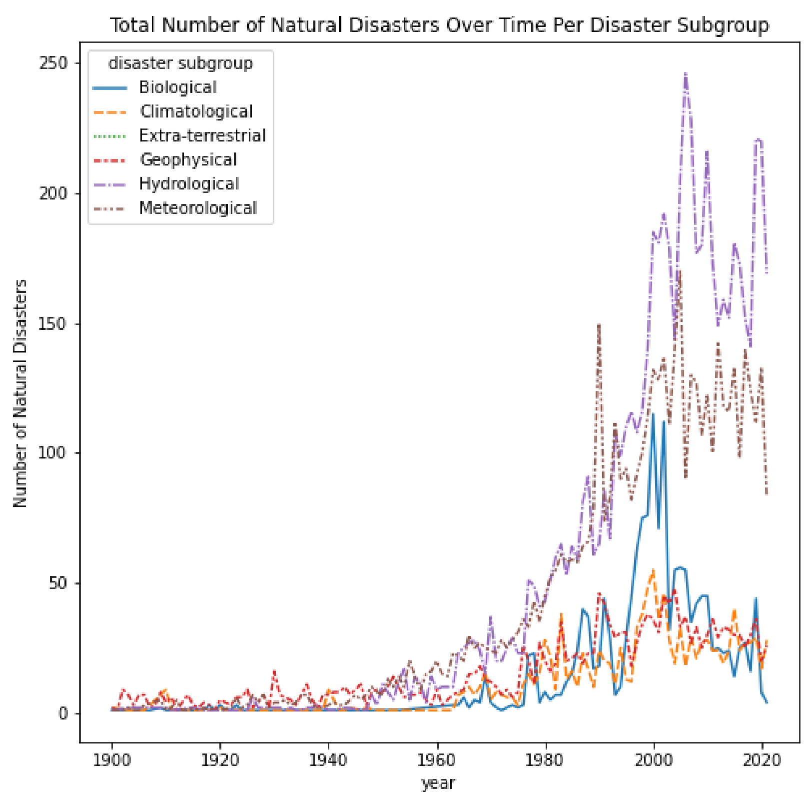
Data on geographical details and duration of an event is sparse, so we will not use this information for future analysis.

Figure 8. The number of natural disasters increased sharply from 1960 to the early 2000s, but it has been in slow decline in the last two decades.



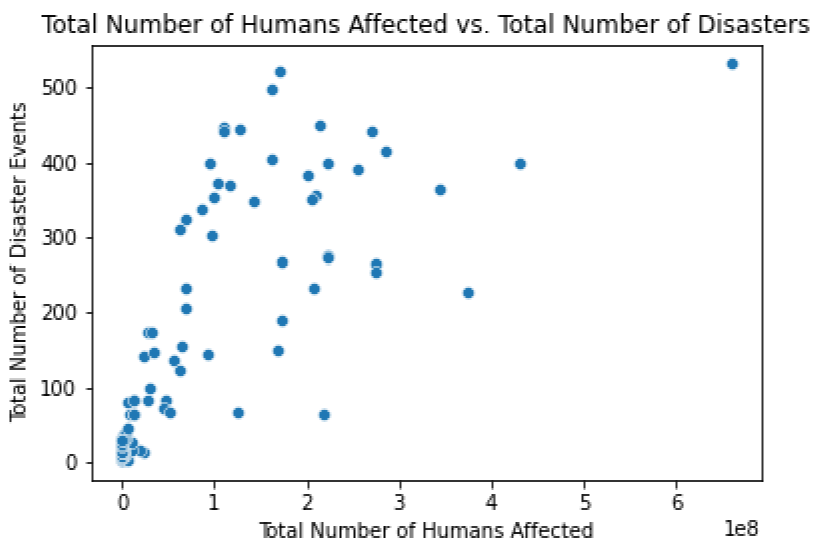
At the global level, the total number of disaster events has been increasing steadily since around 1960. It is not clear that the reported data prior to 1960 is reliable. However, it does appear that the steady increase peaked around the early 2000s, at which point, there is a downward trend for the total number of reported events. For all the disaster subgroups, hydrological events have the largest number of events reported.

Figure 9. Hydrological events are the dominant natural disaster type globally.



Of all the reported events,the United States has the largest number of events reported, followed by China and India. These top 3 countries are some of the largest countries in the world, so there may be a land area component affecting the result. Of the top 10 countries, many of them are smaller countries within the ring of fire (The Philippines, Indonesia, Japan, Mexico). On first impression, this may be attributed to the number of earthquakes in these areas, but all of the top incidents are of type hydrological or meteorological.
For the human impact, all distributions are right-skewed and indicated that most events had less than 1000 casualties (all non-deaths affected and deaths). There was no obvious trend for the total numbers over time. In order to have one representative number for the human impact, we added the total number of people affected and the total deaths and referred to this here as 'total casualties.' The total casualties have some positive linear correlation with the total number of disasters (R2 = 0.6, P = 0.0). For modeling, we are keeping both the number of events and the total number of casualties for modeling.

Figure 10. The total number of disasters and the total number of casualties have weak positive correlation.



For subsequent analysis, only the number of disaster events and total casualties were used for generating the master dataset (see Summary section below).

Summary
We combined all datasets on country (ISO code) and year to generate a master dataset. This master dataset is used for all subsequent analysis and modeling. This was further limited to years 1961-2019 when rice production data was available and also limited to only countries (population data included . Our dataset had 14306 rows and 15 columns. Missing values from the natural disasters' dataset were assumed to be zero and imputed as such. We also dropped any countries missing rice production data.

Out of 147 countries, we found that rice production is dominated by only a handful. In fact, 14 countries account for 90% of global rice production since 1961. Therefore, we decided to streamline our analysis and consider only these major producers. These countries include China, India, Indonesia, Bangladesh, Vietnam, Thailand, Myanmar, Japan, The Philippines, Brazil, The United States, South Korea, Pakistan, and Cambodia. This final dataset was then used for our subsequent models. The final dataset had 826 rows and 15 columns. The columns include: year (1961-2019), iso (alpha-3 country code), country_or_area (country or territory name), region (categorical region classification of the country), continent (continent in which the country belongs to), number_of_disasters (total number of disasters in the specified year and country), temperature (average temperature of the specified year and country in ℃), precipitation (total precipitation of the specified year and country in mm), cropland (total area dedicated to cropland specified year and country in 1000 hectares), country_area (total area of the specified year and country in 1000 hectares), cropland_pct (percent land dedicated as cropland for specified year and country), variant (model variant used for population projections; only applicable to population in the future), population (population of the specified year and country), and rice_production_in_tonnes (rice production of the specified year and country in tonnes). We further narrow the features used for our models which are described below.

STATISTICAL ANALYSIS AND MODELING

Our model is designed on the following variables:

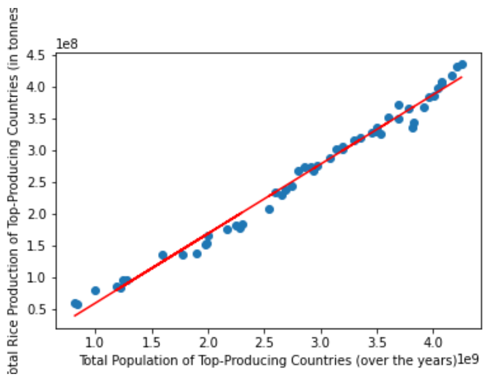
Table 3. List predictor and response variables for available for modeling.

Response Variable	Annual global rice production
Predictor Variables	Population
	Number of disasters per country per year
	Total people affected by the disasters per country per year
	Total annual precipitation
	Average annual temperature
	Total cropland area
	Total country area
	Percentage of country area dedicated to cropland (cropland area / country area)

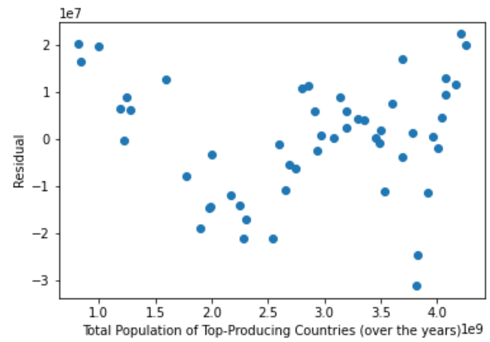
We started by summing all data across countries to compute a single global value for each variable for each year in our dataset. In a simple 2D correlation matrix, we found immediately that rice production (rice_production_in_tonnes, last row in figure) is highly correlated with population (0.95). It also has positive correlation with cropland and country area which can be intuitively described since larger countries can grow a large amount of crop. There are moderately large negative correlations with temperature and precipitation (-0.22 and -0.26). Finally, there are large positive correlations with the number of disasters and total humans affected (0.48 and 0.56) which were unexpected trends.

Table 4. Population and rice production have a strong positive correlation.

Graph A: Population vs Rice Production for Linear Regression Model



Graph B: Population vs Rice Production Linear Regression Residual Plot (Observed - Predicted)



As shown in figure #-A above, we found that population alone has a strong positive correlation with rice production. Because we are interested in understanding the effect of climate change to rice production, we removed population as a predictor variable so as to not confound our models and findings.

Model 1: Linear Regression Including All Countries

Table 5. List of predictor and response variables used for Model 1.

Response Variable	Annual global rice production
Predictor Variables	Number of disasters per country per year
	Total people affected by the disasters per country per year
	Total annual precipitation
	Average annual temperature
	Total cropland area

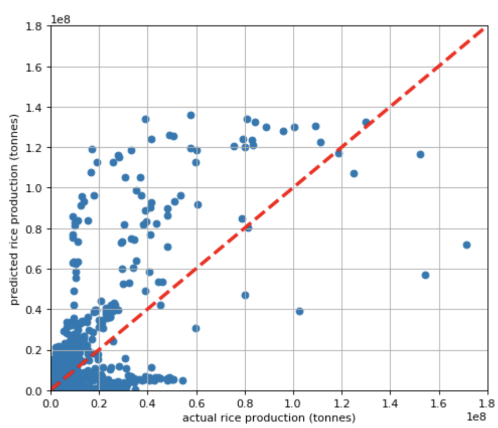
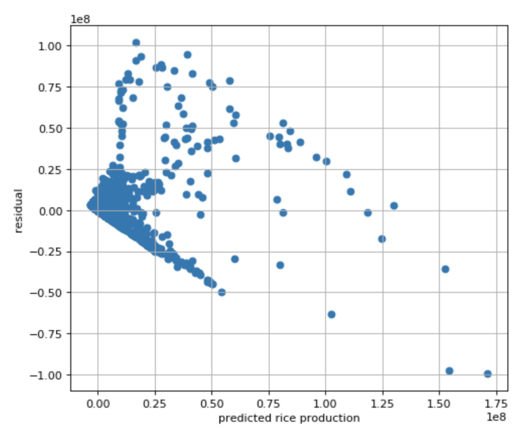
In our first model, we included data for all countries for all the available years (1961-2019). However, we had to exclude the data for years 2013-2019 in the modeling processes due to missing precipitation and temperature data for those years. For modeling we focused on ordinary least square linear regression analysis with predictor variables being: number of disasters per country per year, people affected by each disaster per country per year, total annual precipitation, average annual temperature, and total cropland area and the response variable being annual global rice production. Linear regression fits a linear model to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.

Table 7 shows the comparison between the predicted rice production and the actual rice production (bothe in tonnes) in the left panel and the residual (the difference between the predicted and actual value) plot on the right panel. From this analysis, we can see that we're incorrectly weighted by low and zero rice-producing countries. Thus, we decided to move to only the top rice-producer countries.

Table 6. List of the R^2 value and the estimated coefficients from the linear regression model.

R^2	0.5188995620724093
Coefficient 1 (number_of_disasters)	9.62052748e+05
Coefficient 2 (total_human_affected)	3.90892356e-01
Coefficient 2 (temperature)	9.07242248e+04
Coefficient 3 (precipitation)	9.50051845e+03
Coefficient 5 (cropland)	1.18414093e+02

Table 7. Summary of Model 1.

<table>
<tr><td>Graph A: Predicted vs Actual Plot</td><td>Graph B: Residual Plot</td></tr>
<tr><td>
</td><td>
</td></tr>
</table>

Model 2: Linear Regression for Top Producing Countries

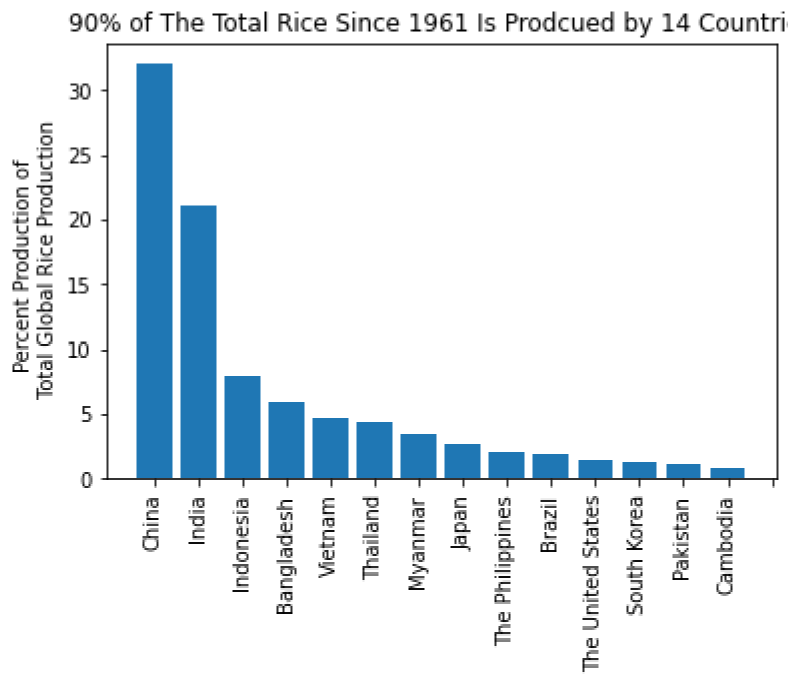
Table 8. List of predictor and response variables used for Model 2.

Response Variable	Annual global rice production in top 14 producing countries
Predictor Variables	Number of disasters per country per year
	Total people affected by the disasters per country per year

	Total annual precipitation
	Average annual temperature
	Total cropland area

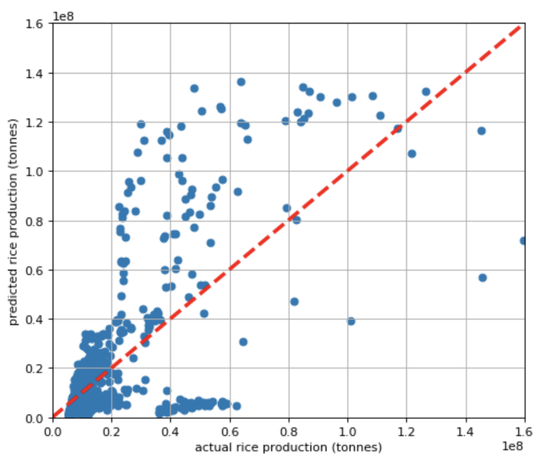
When calculating percentage production of each country (of the total global rice production from 1961 to 2019), we found that about 90% of total global production is accounted for by 14 countries (Figure 11).

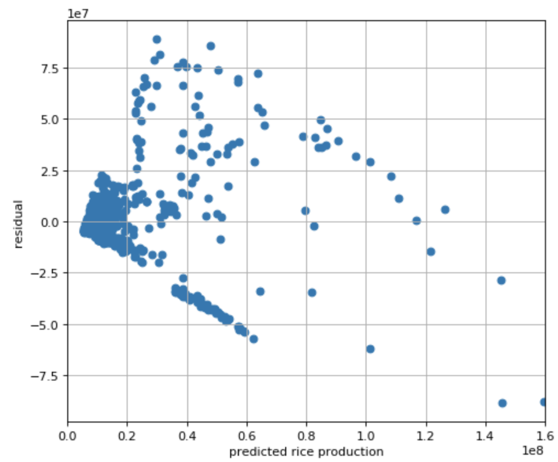
Figure 11. Top 14 countries account for 90% of the total global rice production.



For the second step of modeling we focused on top rice-producer countries and performed the same regression analysis for this subset of countries. We picked the countries that produce 90% of the global rice corp including 14 countries listed below: China, India, Indonesia, Bangladesh, Vietnam, Thailand, Myanmar, Japan, The Philippines, Brazil, USA, South Korea, Pakistan, and Cambodia. Figure # shows the comparison between the predicted rice production and the actual rice production (bothe in tonnes) in the left panel and the residual (the difference between the predicted and actual value) plot on the right panel for the top 14 rice-producer countries (90% of the global rice production).

Table 9. Summary of Model 2.

<table>
<tr><th>Graph A: Predicted vs Actual Plot</th><th>Graph B: Residual Plot</th></tr>
<tr><td>


</td><td>


</td></tr>
</table>

Table 10. List of the R^2 value and the estimated coefficients for the linear regression model 2.

R^2	0.46401770672920356
Coefficient 1 (number_of_disasters)	7.49365699e+05
Coefficient 2 (total_human_affected)	3.38064202e-01
Coefficient 2 (temperature)	-1.80659164e+05
Coefficient 3 (precipitation)	5.73058513e+03
Coefficient 5 (cropland)	1.43387850e+02

Model 3: Individual Year Models

Table 11. List of predictor and response variables used for Model 3.

Response Variable	Annual global rice production in top 14 producing countries
Predictor Variables	Number of disasters per country per year
	Total people affected by the disasters per country per year
	Total annual precipitation
	Average annual temperature
	Total cropland area

Figure 12. Population has a strong positive correlation with rice production

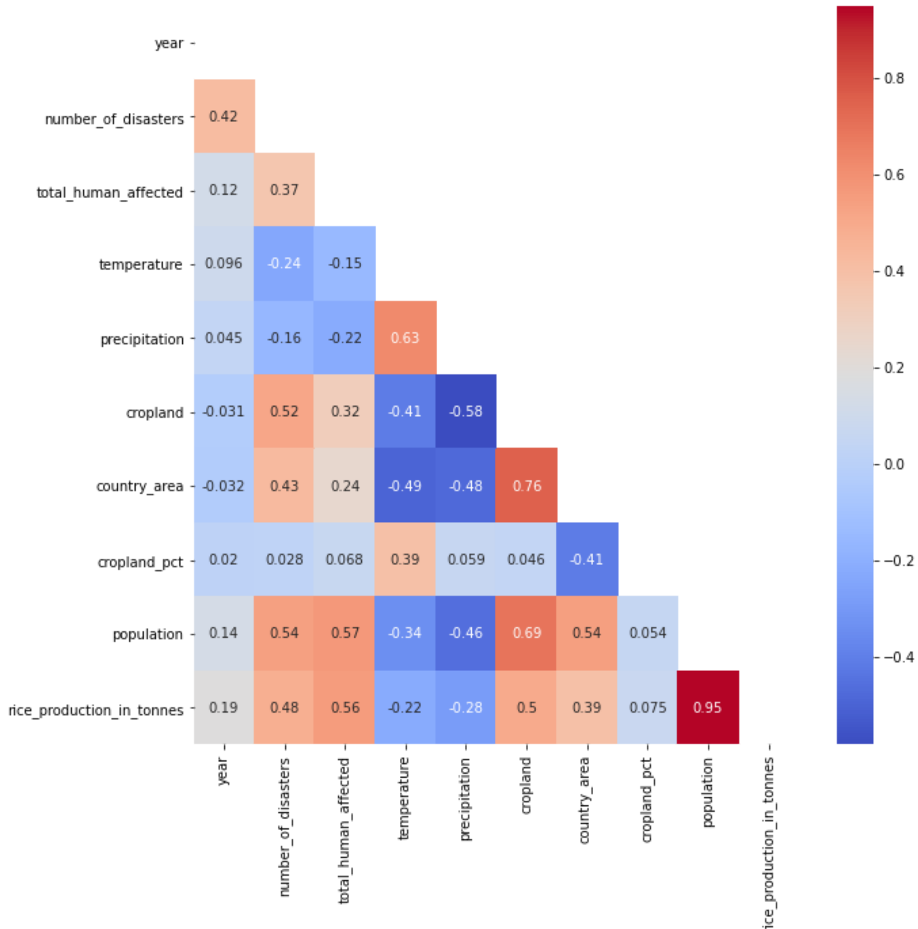
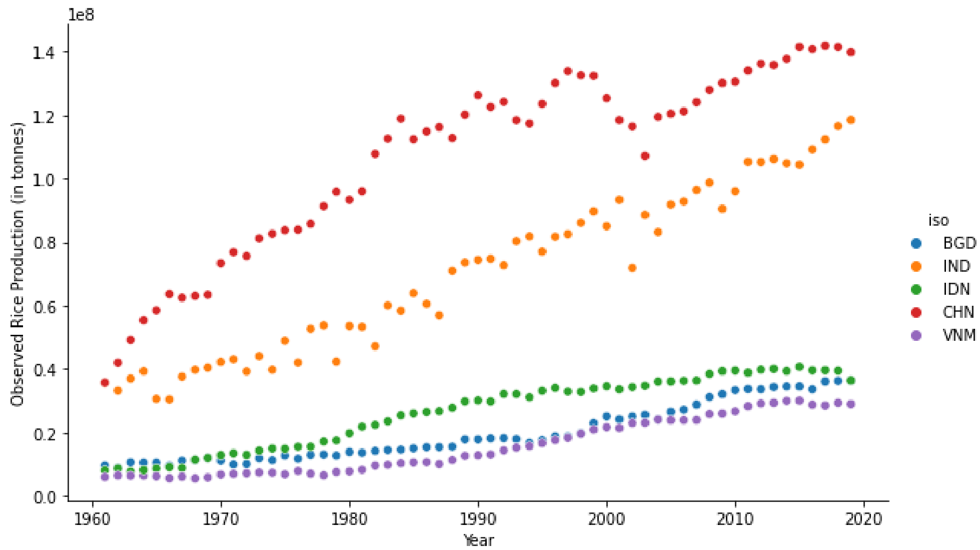


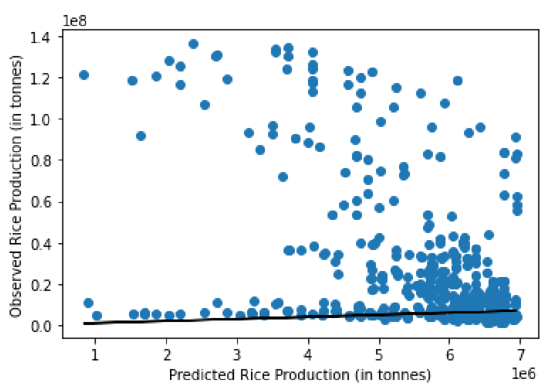
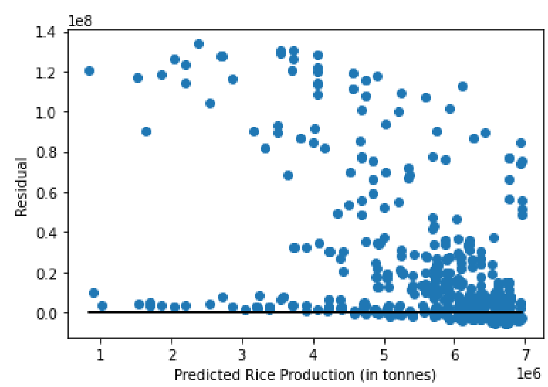
Figure 13 shows the rice production (in tonnes) over the years 1961 to 2019 for the top 5 producing countries. We observe how the production of rice has been on the increase for these countries but at rate appears to vary. Therefore, Instead of grouping all years data together, we decided to consider separating by year, producing a multiple linear regression model for each year, and then taking the average linear regression model over the years 1961-2012.

Figure 13. Rice production has been increasing at varying rates for the top 5 countries.



We perform regression analysis for all the top-producing countries for years 1961-2012 using the average linear model. The residual plot suggests that we are missing some features. It is overestimating the rice production for the predicted low-producing countries and underestimating some of the predicted high-producing countries.

Table 12. Summary of Model 3.

Graph A: Predicted vs Actual Plot	Graph B: Residual Plot (Observed - Predicted)
	

Model 4: Panel Analysis

Table 13. Predictor and response variables used for Model 4.

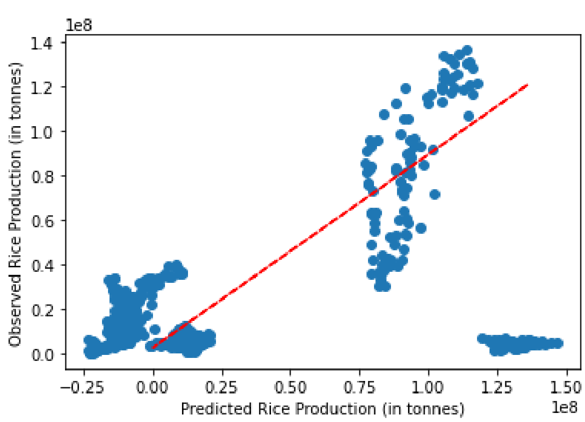
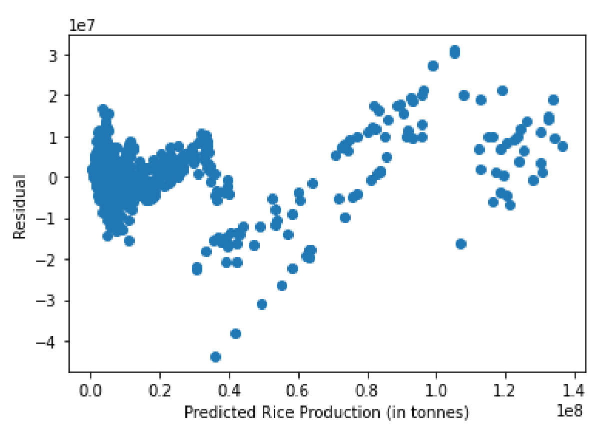
Response Variable	Annual global rice production in top 14 producing countries
Predictor Variables	Number of disasters per country per year
	Total people affected by the disasters per country per year
	Total annual precipitation
	Average annual temperature
	Total cropland area

Since we have panel data where we have repeated measurements from the same countries over time, we fit a fixed effects model with both country level and time level effects. Our model is of the form:

$$y_{it} = X_{it}\beta + \alpha_i + \epsilon_{it},$$

where α_i is the country specific effect, and t denotes the time dependence (in years).

Table 14. Summary of Model 4.

<table>
<tr><td>Graph A: predicted vs actual</td><td>Graph B: Residual Plot</td></tr>
<tr><td>
</td><td>
</td></tr>
</table>

Adding the country specific parameter, α_i , increases the complexity of our model and it does a poor job fitting our data ($R^2 = 0.30$). Specifically, for the lower producing countries, we get negative predictions for rice productions implying model misspecification. Note that this model performs worse than the linear regression in Model 2 which is equivalent to a pooled panel analysis. This implies accounting for country specific clusters wasn't hugely beneficial to the model.

Model 5: Generalized Additive Model (GAM) Analysis

Table 15. Predictor and response variables used for Model 5.

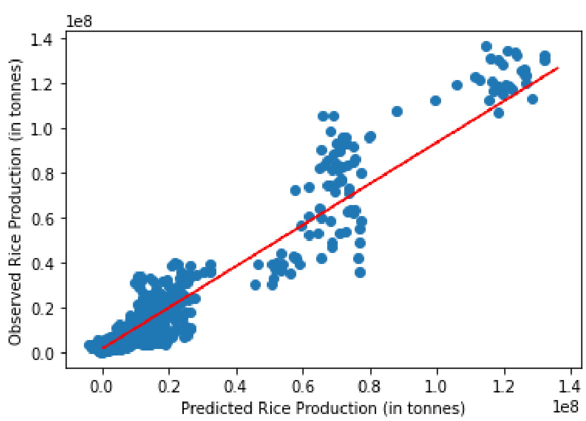
Response Variable	Annual global rice production in top 14 producing countries
Predictor Variables	Number of disasters per country per year
	Total people affected by the disasters per country per year
	Total annual precipitation
	Average annual temperature
	Total cropland area

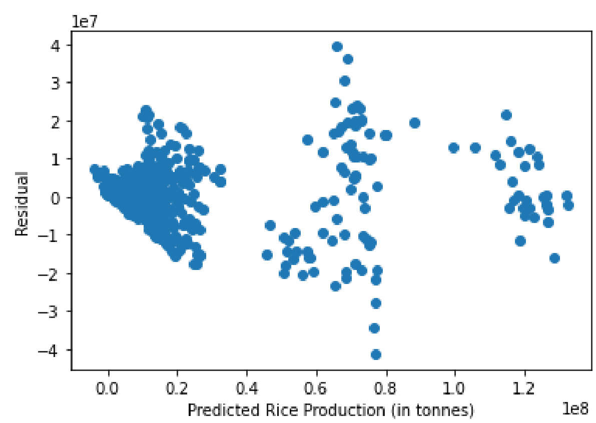
In order to explore whether there was a nonlinear relationship between the predictors and the dependent variable (Y), we fit a GAM model to our data. GAM is a type of statistical modeling where a target variable is roughly represented by an additive combination of a set of different functions of the predictors or independent variables. It is represented by the following formula:

$$g(Y) = f_1(X_1) + f_2(X_2) + f_3(X_3) + f_4(X_4) + f_5(X_5) + \varepsilon,$$

where g is a link function and f are functions of the predictors. For our model, we used the identity function as a link function and considered Y to follow a normal distribution.

Table 16. Summary of Model 5.

<table>
<tr><td>Graph A: predicted vs actual</td><td>Graph B: Residuals</td></tr>
<tr><td>


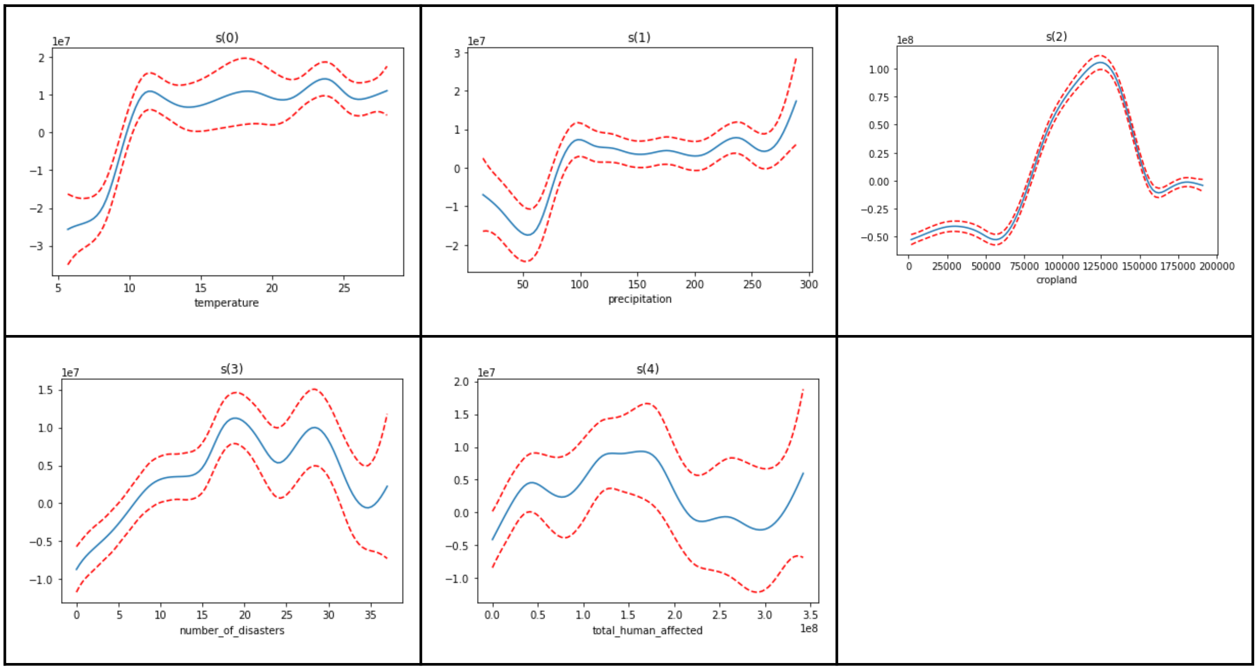
</td><td>


</td></tr>
</table>

The linear GAM model fits our data well ($R^2 = 0.93$) as seen by the observed vs predicted plot where the data track fairly close to the y=x line. The residual plot on the right doesn't show too much heteroscedasticity with the magnitude difference in the top producing countries arising from the data.

We can decompose and inspect the contribution of each feature to the overall prediction of our model. The plot below shows the function terms in our GAM model showing nonlinear functions for each predictor.

Figure 13. The GAM appears to be over-fitting.



These graphs help us visualize a potential non-linear relationship that may exist between our predictors and the dependent variable. However, GAM models have a potential to overfit data and this may explain why our GAM model does a great job at predicting the outcome. Our model may underperform with new testing data if the relationship between the predictors and the dependent variable is not as nonlinear as implied by the splines above. Nevertheless, this model does show that the relationship between predictors and rice production may not all be linear.

CONCLUSIONS

We set out to predict rice production from features that relate to climate change. We extracted data from a variety of sources to include climate variables such as the number of natural disasters, temperature, and precipitation. From the combined dataset, we predicted the global rice production from the available variables and produced five models. In all of our models, we observed a negative correlation between rice production and both temperature and precipitation. We observed a positive correlation between rice production and both natural disasters and cropland.

However, we are not confident in these observed correlations being very informative. For example, for the observed temperature and precipitation results were not always statistically significant. Additionally, for our observed positive correlation between rice production and the number of natural disasters, this is because it appears that record-keeping for natural disasters may not be as reliable as it is in recent decades.

Based on the success of the panel model with respect to the first three models, we do recommend clustering by country-level effects and time or year-level effects in future analyses. It would be interesting to bring back in all countries to this style of model.
The nonlinear GAM model was the most successful but was evidently overfitting some parameters. We expect it would not be robust to new test data and we recommend refitting when additional data as described in the Future Work section becomes available.

FUTURE WORK

From our various models, we found that linear models could not sufficiently predict crop production and that even our nonlinear model relied on complex relations and therefore likely overfit the data. Inspired by our analysis, we propose the following future directions.

Additional Variables

In several of our models, we saw additional patterns in the model residuals which imply that there must be additional variables yet omitted. One such obvious variable is technological proficiency of the farming practices. Improved technology in irrigation and crop cover can protect against extreme temperature and precipitation. New seed technology and breeding advancements can similarly protect against drought or overwatering. We considered using national GDP as a marker for technological advancement, however, this is naturally correlated to crop production – especially in the major cash crop producing countries that we are considering. Another variable we wished to include were the maximum and minimum temperature and precipitation instead of the annual averages. We expect that these variables have greater year to year variance and could better explain the total or variance in total annual crop production.

Country-specific models instead of a global prediction

Correlations between climate and crop production vary per region. To increase predictive accuracy, regional models should be created. The bulk of rice production occurs in East and Southeast Asia. Other new models could be built for major producers like Brazil and the United States.

Predicting Variation instead of Absolute Production

Our model and datasets could be quickly adapted to answer a different global question: can we predict years of production which are over or under the historical trends. In other words, if we assume that global production continues to increase linearly on average, can we predict years of surplus or subpar production years which in real life result in oversupply and waste or undersupply and famine.

Alternative Models

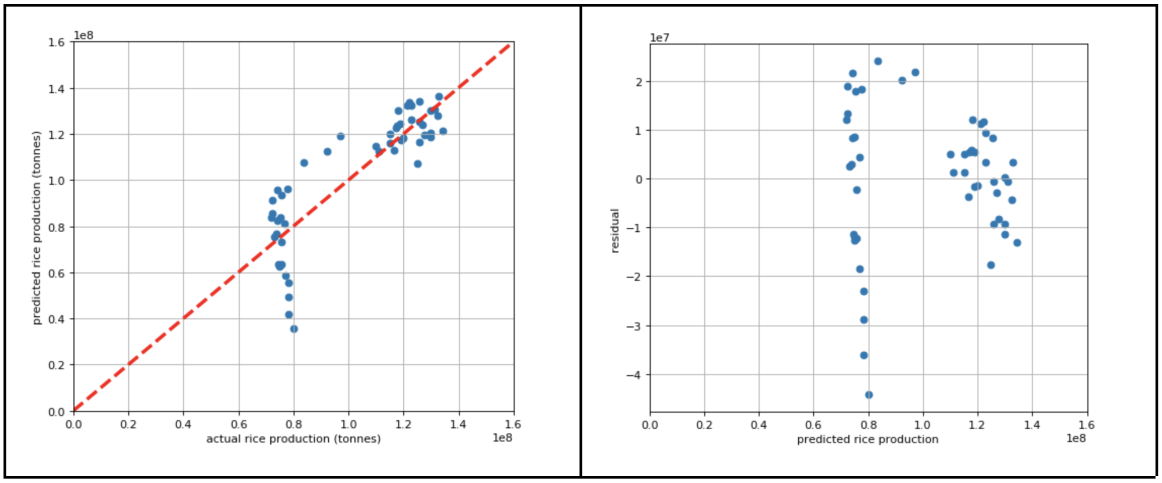
While linear regressions could be improved with additional variables as mentioned above, we also have great interest in more advanced models. While beyond the scope of this project, we propose multivariate time series modeling and vector autoregression as alternatives. We expect that these both will be able to better take into account historical trends in cropland use and previous years' farming productivity to model the future.

APPENDIX

Linear regression for Country-by-Country case study

In order to further investigate the performance of the linear regression model, we ran the model for some country-by-country cases. Out of the 14 top rice-producer countries we presents the results for the first top 3 countries (China, India and Indonesia) and the last 3 countries on the list (Korea, Pakistan and Cambodia) in figures #-#, respectively. Apparently, the model performs better (higher R^2 values) when considering each country separately compared to the cases where we aggregated all countries together. This initiated the idea of "country-specific models instead of a global prediction" which we presented as one of the suggested future work in our report.

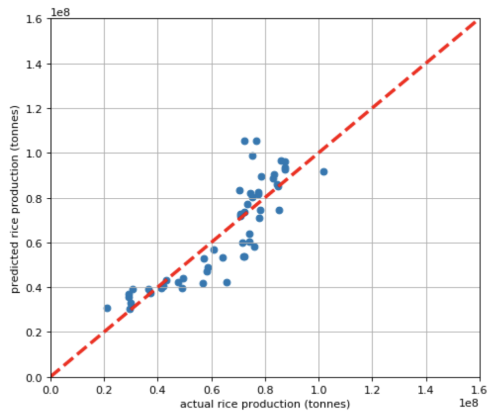
China

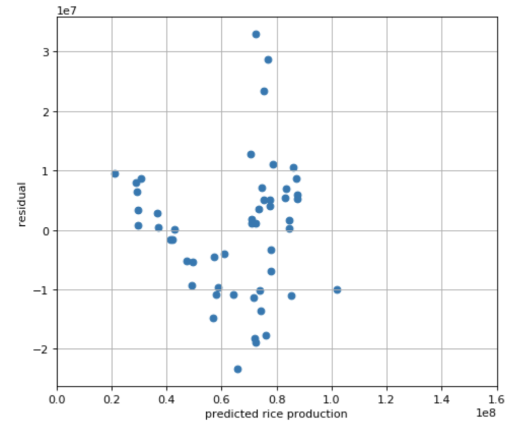


This table lists the R^2 value and the estimated coefficients for the linear regression model.

R^2	0.730240885045064
Coefficient 1 (number_of_disasters)	4.95545523e+05
Coefficient 2 (total_human_affected)	-1.44604564e-02
Coefficient 2 (temperature)	3.07140008e+06
Coefficient 3 (precipitation)	3.47527212e+05
Coefficient 5 (cropland)	1.22966237e+03

India

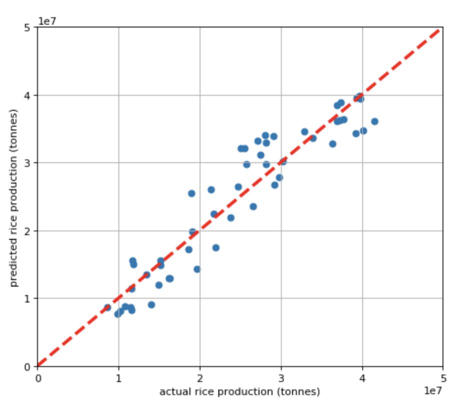


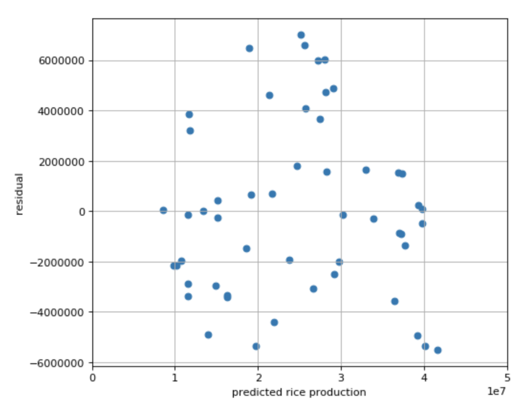


This table lists the R^2 value and the estimated coefficients for the linear regression model.

R^2	0.7505455217629496
Coefficient 1 (number_of_disasters)	1.23288936e+06
Coefficient 2 (total_human_affected)	-3.81166706e-02
Coefficient 2 (temperature)	2.51788702e+06
Coefficient 3 (precipitation)	3.51610077e+05
Coefficient 5 (cropland)	4.82133630e+03

Indonesia

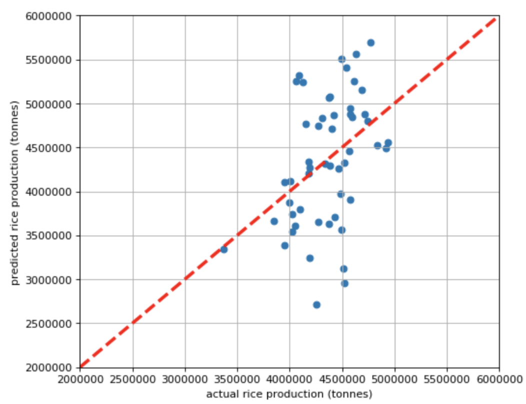


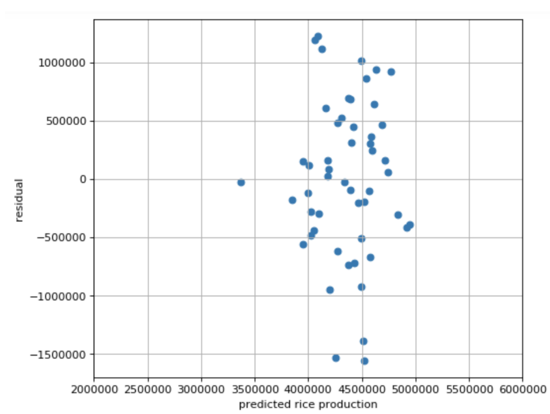


This table lists the R^2 value and the estimated coefficients for the linear regression model.

R^2	0.8953430520006394
Coefficient 1 (number_of_disasters)	3.01123271e+05
Coefficient 2 (total_human_affected)	2.59380643e-01
Coefficient 2 (temperature)	1.67813159e+07
Coefficient 3 (precipitation)	5.35474015e+03
Coefficient 5 (cropland)	7.77272076e+02

Korea

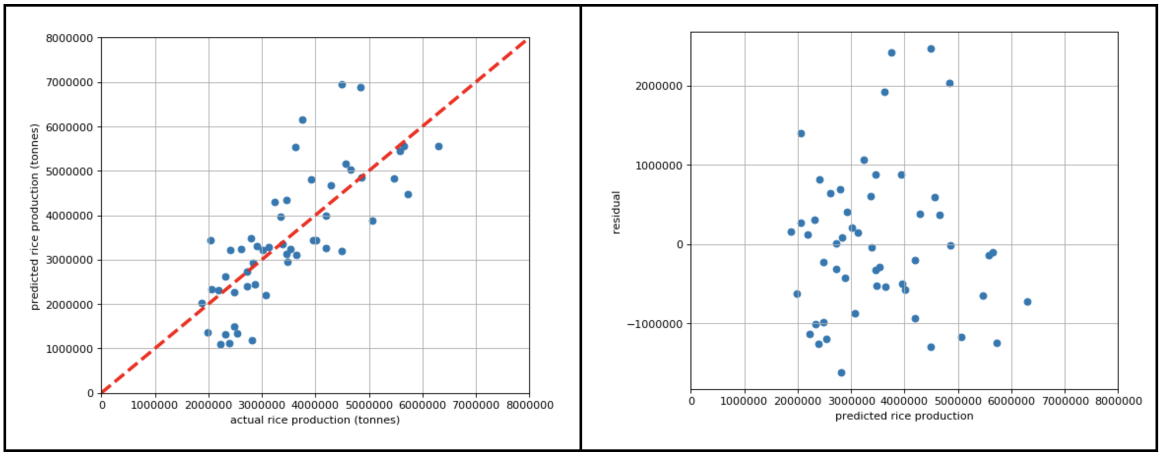




This table lists the R^2 value and the estimated coefficients for the linear regression model.

R^2	0.1678781645691764
Coefficient 1 (number_of_disasters)	7.95593276e+04
Coefficient 2 (total_human_affected)	-4.31977155e-01
Coefficient 2 (temperature)	4.32436000e+05
Coefficient 3 (precipitation)	-4.23126511e+03
Coefficient 5 (cropland)	1.02820465e+03

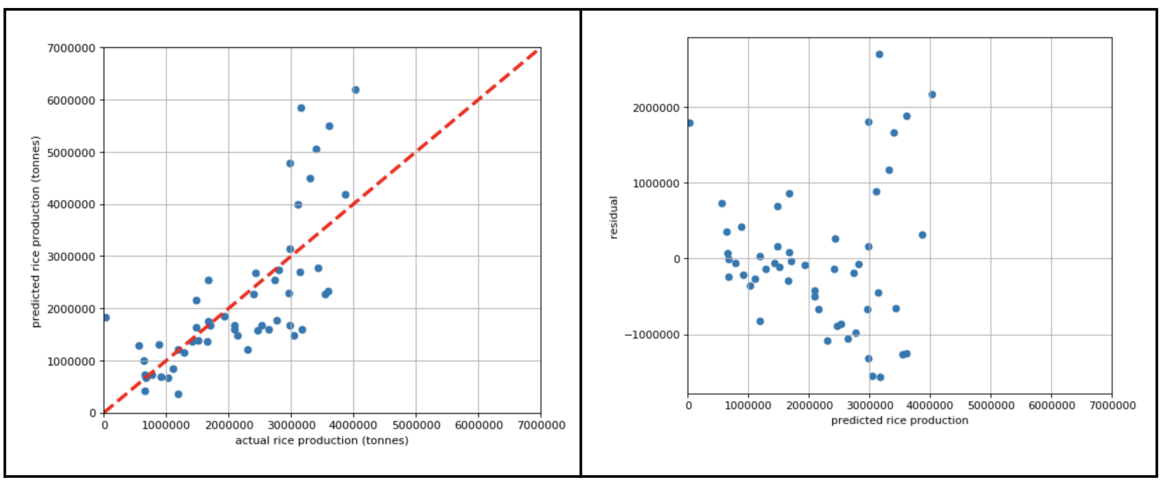
Pakistan



This tablelists the R^2 value and the estimated coefficients for the linear regression model.

R^2	0.5834483467919691
Coefficient f1 (number_of_disasters)	2.50073724e+05
Coefficient 2 (total_human_affected)	-2.04248528e-02
Coefficient 2 (temperature)	1.16951466e+06
Coefficient 3 (precipitation)	5.86875187e+04
Coefficient 5 (cropland)	-1.24701835e+02

Cambodia



This table lists the R^2 value and the estimated coefficients for the linear regression model.

R^2	0.5399876927607575
Coeff1 (number_of_disasters)	1.43098192e+05
Coeff2 (total_human_affected)	-2.84185043e-01

Coeff2 (temperature)	7.30112869e+05
Coeff3 (precipitation)	1.36446595e+04
Coeff5 (cropland)	9.93100353e+02