Predicting Crop Damages

RANDOM FORESTS VERSUS PANEL LINEAR PROBABILIITY MODELS

SHARFIYET IQBAL

Abstract

With increased frequency and severity of extreme weather events such as droughts due to climate change, crop yields are increasingly vulnerable to weather related damages, leading to potential risks of increased food insecurity and reduced agricultural productivity. It is essential to have measures against predicted hazards that may lead to substantial crop damages. This paper compares two predictive models; Panel Linear Probability Model (PLPM) and Random Forest (RF) for predicting crop damage events. Based on the analysis, RFs produce fewer errors, are more accurate and have higher recall than PLPMs.

Keywords: Classification, Panel Linear Probability Model, Random Forest, Prediction, Training, Testing, Climate Change, Temperature, Precipitation, County, Fixed Effects, Crop Damage.

Table of Contents

Introduction

Since the pre-industrial period (post 1850-1900), the earth's global surface temperature has increased by an average of 1^{o}C (NASA, 2022). The earth is currently warming at a rate of 0.2^{o}C per decade. This phenomenon has been coined Global Warming; "a long-term heating of the earth's surface temperature". Scientific consensus suggests that global warming is primarily due to human activities, particularly fossil fuel burning which has increased "heat-trapping greenhouse gases in the Earth's atmosphere" (NASA, 2022).

Global warming is a key indicator of climate change, which refers to a long-term change in the average weather patterns across "local, regional and global climates" (NASA, 2022). Scientists have used environmental data and computer models "to monitor past, present and future climate change". The data and applied models have produced key climate indicators that show global land and sea temperature increases, rising sea levels, melting of ice from earth's poles and mountain glaciers, changes in cyclical ocean patterns, an increased frequency and severity of extreme weather events such as droughts, and the changes to the coverage patterns of clouds and vegetation (NASA, 2022).

The Intergovernmental Panel on Climate Change (IPCC) on its Sixth Assessment Report presents the observed and projected impacts, risks, adaptation measures and enabling conditions of climate change, and strategies for climate resilient development (IPCC, 2022). Notably, the IPCC asserts that climate change and the related frequent and severe extreme weather events have led to natural and societal damages and losses of a greater magnitude than natural climate variability. The observed increase in frequency and intensity of extreme weather events, such as droughts and fires, extreme hot weathers in land and oceans, and heavier rainfall events have extensively impacted "ecosystems, people, settlements, and infrastructures". Some of these

impacts include increased tree mortality due to droughts, increased bleaching and mortality of warm-water corals, and increased human mortality due to extreme heat (IPCC, 2022).

Increased frequency and severity of extreme weather events due to climate change have impacted agriculture and food security across the world (IPCC, 2022). The rate of growth of agricultural productivity has declined in the last 50 years worldwide. Mid and low-latitude regions experienced negative impacts of climate change, but there were some positive impacts to agriculture and food resources in high-latitude regions. Notably, the IPCC reports with high confidence the negative impacts to global aggregate agriculture/crop production in Africa and Australasia, and mixed (negative and positive) impacts to North America, Europe, the Arctic, and Asia. Extreme weather and increased heat and dry conditions have caused loss of crop production in Europe and Africa, with the additional adverse impact of increased food insecurity in Africa (IPCC, 2022).

In addition to increased and more severe extreme weather events, an increase of concurrent extreme weather events has also been observed (IPCC, 2022). Increased concurrent drought and heat events have led to losses in crop production and tree mortality. This increase in concurrence of drought and heat events above 1.5^{o}C global warming is projected to cause an increase in the loss of maize production in major food-producing regions. The forecasted decrease in crop production is further intensified by a projected decrease in labor productivity due to extreme heat conditions. With reduced crop production, the supply of food would decrease, leading to increased pricing to curb food demand. Furthermore, low labor productivity and expenses would decrease household incomes. Therefore, there would be an increase in the risks of malnutrition and climate-related mortality with little to no means of adaptation, particularly in tropical regions (IPCC, 2022).

The adverse effect of climate change on agricultural production has also been extensively studied in scientific literature. Rosenzweig et al. (2013), through the use of multiple global gridded crop models (GGCMs), found strong negative effects on climate change on agriculture, particularly in areas of low-latitude and higher levels of warming. Majority of the models studied by the authors also indicated decreased agricultural yield in both high and low latitude regions. The authors stress that many agricultural areas across the globe are prone to future declines in crop yield due to climate change and additional issues such the water scarcity and soil degradation, and concerns regarding pests. This was exemplified by the 2012 drought in the United States (US), where declines of up to a quarter of total maize yields was observed, but the reduction in US maize exports was even greater. The authors also note that even though high-latitude regions may be better adapted to the effects of drought, other factors such as quality of soil may limit crop production (Rosenzweig et al., 2013).

To estimate the impact of climate change on the global agricultural market, Costinot, David and Smith (2016) aggregated data of 10 crops across 1.7 million agricultural fields in the world. The authors first generated estimates of agricultural productivity pre and post climate change. They then developed a general equilibrium model of trade between the 1.7 million fields. The authors estimated a 0.26% decrease in global GDP, allowing for adjustment of trade and production patterns. As the 10 crops in this study accounted for 1.8% of world GDP, the 0.26% decrease in global GDP accounted for one sixth of the "total crop value". The authors suggest that while trade adjustments may not significantly mitigate the decline in crop related GDP, production adjustments stemming from changing comparative advantages can substantially reduce climate change related impacts.

As discussed so far, the adversities of climate change have been documented thoroughly. Particularly for the case of agriculture/crops, the risks range from economic (reduced grain exports/imports and gross domestic product), social (increased food insecurity in Africa) and biological (higher risks of toxicity and soil quality due to reduced availability of fertile soil and increased pollution) (Rosenzweig et al., 2013; IPCC, 2022). These risks are further exacerbated by more frequent and severe extreme weather events such as droughts (IPCC, 2022). Therefore, it is essential that adaptation and planning measures are in place for climate change and related extreme events for effective intervention and reduction of adversities to the affected regions (Mann, Warner and Malik, 2019). Interventions can be achieved by means of policy and climate resilience strategies (IPCC, 2022). However, for a more targeted approach to solving real-time crises that are likely to be more frequent in the near future and moving forward, predictive modelling using real time climate data is crucial (Mann, Warner and Malik, 2019). This would entail acquisition of climate and geographic data, sub-setting part of the data to train econometric or machine learning models, generating predictions of interest (e.g. crop damage) on the remaining sub-set of data (test data), and lastly comparing the predicted events against the true events in the test data set (Mann, Warner and Malik, 2019).

Although both econometric and machine learning models can be used as predictive tools, machine learning has some advantages in the arena of prediction (Chen, 2021). Chen states that traditional statistical models, such as linear regression models can help researchers infer the signs and magnitudes of relationships between input and response variables. Using climate data as an example one can model the effects of climate variables, such as temperature and precipitation, on agricultural variables, such as crop yields. The significance of these effects can further be evaluated through statistical hypothesis tests, which include test statistics, confidence intervals,

and coefficient standard errors. However, Chen states that such models have limited predictive accuracy when data is high dimensional and has heteroskedastic error terms. Furthermore, statistical regression techniques often require the removal of outliers in data, which is in conflict with phenomena such as natural hazards. Hazards may be infrequent, but pose serious consequences to the environment and populations inhabiting it. In such cases where outliers provide valuable information to a model, machine learning can provide more accurate predictions (Chen, 2021).

Machine learning models fare better with hard to interpret, high dimensional data with heteroskedastic errors (Chen, 2021). Tree-based models such as random forests are robust to the presence of outliers and so are ideal for modelling impacts of natural hazards on crop damage. Furthermore, machine learning models can select control variables and, in some cases, interactions based on the data used for training to optimally fit models for predictions, and are better at error reduction than statistical models (Schrimpf, 2020). For statistical models, researchers set arbitrary set of controls and interactions e.g. polynomials, interactions and dummies. However, in situations where interpretability is important e.g. signs and magnitudes of coefficients, machine learning provides little information. At best, machine learning can act as a complement to statistical models for interpretability through providing information on importance of features (Chen, 2021). Therefore, it is important to understand the purpose of the modelling task; statistical models are superior when it comes to interpretability, but machine learning models are superior in prediction (Chen, 2021).

In their 2019 article, Mann, Warner and Malik (2019) state that current impact assessments of droughts on agriculture are “ad hoc, late or spatially imprecise”. The authors state that this fails to capture the large amounts of variability of these agricultural losses due to

drought at the village level. They proposed utilizing a hybrid of remote sensing and agricultural survey data to mitigate timing and scale related limitations of current impact monitoring/assessment practices. Utilization this type of data is particularly impactful for developing countries. For example, in Ethiopia, crop farming generally takes place in rainfed small hold agricultural lands. Only 2% of farms have irrigation systems in place. Droughts are especially harmful for crops in these lands, further exacerbated by the more frequent and severe droughts brought about by climate change (Mann, Warner and Malik, 2019).

As an adaptive measure against drought, remote sensing can monitor growth phases of crops and utilize this information around the midpoint of the growing season, which allows advanced prediction of substantial crop losses at time of harvest (Mann, Warner and Malik, 2019). The authors used agricultural survey data from 2010-2015 of Ethiopian sub-kebeles (villages with approximately 200 households, covering an area of 24km^2), combined with remote sensing from early growing season to forecast crop losses during harvest. Additional variables included precipitation, water for hydrological use and energy availability. The five major cereal crops harvested in Ethiopia were considered for this analysis; maize, wheat, sorghum, barley and teff (Mann, Warner and Malik, 2019).

The authors then trained a random forest model with remote sensing and agricultural data and the additional variables. This allowed for predictions in the test set using only remote sensing data. The authors predicted substantial crop losses using this model, which was defined as "crop losses of greater than or equal to 25% due to drought at a village level for five primary cereal crops". The most important features identified for predicting substantial crop losses using this random forest model were the time of maximum greenness and the initial greenup of crops. Results showed high predictive accuracy of substantial crop damage for all five cereal crops

investigated. Maize showed the best predictions for substantial crop damage - 81% correct predictions, followed by sorghum, wheat, barley and teff (75%, 65%, 58% and 57% correct predictions respectively) (Mann, Warner and Malik, 2019).

Given the current global situation with climate change, this essay draws inspiration from Mann, Warner and Malik's (2019) work on predicting substantial crop losses in Ethiopian Sub-kebeles. Using hazard data from the Spatial Hazard Events and Losses Database for the United States (SHELDUS), combined with annual US temperature and precipitation data, this study develops models to predict out of sample crop damages due to drought. Annual data from 1979 to 2018 at a county level for the US states California and Iowa is subset into an 80:20 split for training and testing respectively. California and Iowa are chosen as they were the top two states in the US in terms of agriculture production (USDA, 2021). A panel-linear probability model (PLPM) and a random forest (RF) are trained and then compared in terms of predictive performance. Results from this analysis shows that in predicting crop damages, the RF produced more accurate predictions with higher recall and fewer errors than the PLPM.

The remaining sections of this essay are as follows. The essay will briefly discuss RF and PLPM as predictive models. This will then be followed by the methods section, which introduces the data and variables, presents some summary statistics, and provides a note of caution on data leakage. This is then followed by the results of this analysis. Lastly, a concluding section discusses some caveats of this paper and recommends some future directions for predictive modelling in relation to climate change research.

Methods

Random Forests

This section summarizes Breiman's (2001) paper on random forests. In predictive modelling, random forests construct multiple decision trees. These decision trees have "nodes'

which represent thresholds for a particular input that branches the output variable into its possible classes. For example, using the variable temperature as a node in a hypothetical decision tree, if temperature exceeds a 25°C threshold, crop damages will occur, but not occur at 25°C or below. In this example, crop damages are classified as damage versus no damage. Breiman states that growing an ensemble of trees and letting these trees vote for the most popular class leads to significantly more accurate classification. Random vectors govern the growth of each tree in the ensembles. Each tree is grown using a training set and a random vector (independent from previously grown trees but identically distributed), which results in a classifier function h(x, random vector_k), where x is an input vector. Examples of ensemble methods include bagging, random split selection, new training set generation, and written character recognition.

Random forests are procedures where a large number of trees are grown, which then vote for the most popular class. Breiman formally defines this as:

> "A random forest is a classifier consisting of a collection of tree-structured classifiers {h(x, random vector_k), k = 1, . . .} where the {random vector_k} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x."

Breiman states that, based on the strong Law of Large Numbers, RFs always converge to a limiting value of a generalization error. This is shown by a margin function constructed from the classifier functions to measure how much the average number of votes for the correct class exceeds average votes for any other class, where a larger margin gives more confidence in the classification. As the number of trees increases, all sequences of random vectors converge to a limiting value of a generalization error (probability that the margin function is less than zero).

Therefore, random forests do not have issues related to overfitting if more trees are added as a model is trained, as the generalization error converges to a limiting value.

Accuracy of RFs depend on the strength of individual tree classifiers and a measure of dependence between them. The upper bound for the generalization error for random forests can be derived from the strength of individual classifiers and the correlation between them. The generalization error is less than equal to the ratio of correlation to the squared strength of accuracy.

Out of bag estimates (generalization error, classifier strength, and dependence estimates) are used determine the number of features (randomly) selected to determine splits at each node. There are two forms of random features - random selection of original inputs and random linear combinations of inputs. Results (classification accuracy) are insensitive to the number of features selected to split each node, thus usually 1-2 features can provide near optimal results.

Adaboost (Freund & Schapire, 1996), another ensemble classifier method, grows an ensemble of trees by successively reweighing the training set but has no random elements. Current weights of the Adaboost training set depend on how previous ensembles were formed. RF with random features provide favorable results to Adaboost, whereas other random feature models (e.g. bagging) are not comparable to Adaboost in terms of performance.

Algorithms that use bagging, random split selection, or those that introduce random noise into the outputs consistently have shown lower generalization error, but none of these perform as well as Adaboost or other training set reweighing algorithms. Therefore, for accuracy comparable to Adaboost, randomly selected inputs or combinations of inputs to grow each tree are introduced. To improve accuracy, the randomness should minimize correlation between features while maintaining strength. This class of procedures have desirable characteristics such

as having accuracy as good (sometimes better) than Adaboost, being relatively robust to outliers and noise, faster computation speeds than bagging and boosting, providing useful internal estimates of error, strength, correlation and variable importance, and being simple and easily parallelizable.

For cases with many input variables (e.g. medical diagnosis and document retrieval), a single tree classifier only slightly improves accuracy over random choice. However, combining trees using random features can improve accuracy. Computing internal estimates of variable importance and binding these together by reuse reruns help understand the mechanism of the RF "black box".

The sections below provide additional details on out of bag estimates and the two types of random features used in RFs.

Using out of bag estimates to monitor error, strength, and correlation

Breiman's (2001) experiments use bagging in along with random feature selection. Each new training set is drawn, with replacement, from the original training set. Trees are then grown using random features in the new training set, and are not pruned. Bagging is utilized for two reasons; Bagging using random features enhances accuracy and bagging produces ongoing estimates of generalization error, strength, and correlation of the ensemble of trees that are done out-of-bag.

An out of bag error estimate is as accurate as using a test set the same size as the training set. Therefore, using out-of-bag error estimates eliminates the need for a set aside data for a test set (Breiman, 2001). Additionally, out-of-bag error estimates are unbiased, unlike methods like cross validation. For each (bootstrapped) training set generated, one-third of the instances are left out. This fraction of the training sets is combined, and the error rate decreases as the number of

combination increase, potentially overestimating the out-of-bag error rate. Therefore, it is necessary to go beyond the point where the test set error converges to get unbiased out-of-bag estimates i.e. running as many iterations until the training error goes beyond test error.

Random forests using random input selection (Forest-RI)

The simplest random forests grow trees to maximum size without pruning using classification and regression trees and at each node, randomly selecting a small group (fixed number) of input variables for splitting (Breiman, 2001). Breiman's experiments used 100 trees to get reliable out of bag estimates, as out-of-bag estimates are based on 1/3 of trees as in forest. Furthermore, growing 100 random forests with random features is much faster than growing 50 trees using all features in AdaBoost.

Breiman (2001) found that test set errors of Adaboost versus RF were comparable. In smaller datasets, difference in error between single input RF versus multiple inputs was negligible, but more pronounced in larger datasets. Additionally, for the datasets tested, a single random input produced smaller test errors than multiple random inputs in some cases, whereas multiple inputs had slightly smaller test errors in other datasets. Therefore, in RF, a single random input to split each node could produce good accuracy. Lastly, computation time using random input selection is faster than Adaboost and bagging when there is a single input. Breiman suggests that this may scale to multiple inputs.

Random forests using linear combination of inputs (Forest-RC)

Breiman (2001) states that if there are only a few inputs, using a substantial subset of inputs may increase strength but also lead to higher correlation between inputs. Alternatively, additional features can be generated by random linear combinations of several input variables. Weighted inputs are added with coefficients that are uniform random numbers on [-1,1]. This can

increase the strength without a large increase in correlation. For the random combinations generated, the best split is searched. This compares better with Adaboost than Forest-RI (Breiman, 2001).

Breiman (2001) states that sometimes, the selection model produces a smaller test error than single input Forest-RI or 2 combination Forest-RC, as the out of bag estimate selects inputs/combinations at random when error rates between the two feature choices are close. He conjectures that on large datasets, as more features are added, strength keeps rising while correlation becomes asymptotic more quickly. Adding more features with forest RC can reduce errors, but if there is no change, additional features using forest RI can reduce the error rate.

Panel Linear Probability Model (PLPM)

Linear probability models are multivariate regressions with a binary or categorical dependent/output variable (Wooldridge, 2012). The coefficient estimate of a regressor in a LPM with a binary dependent variable represents the change in probability that an event will occur (for example, change in probability that crop damage will occur), in response to a change in that specific regressor (for example, a change in precipitation by 1mm, holding other regressors constant).

A PLPM is the application of an LPM on panel data. As this dataset includes year and county information, techniques of panel regressions were applied to LPM for this study. Specifically, the PLPM model took the form:

$$P(Y_{i,t} = 1 | X_{i,t}) = a_i + B_1 X_{1,t} + B_2 X_{2,t} + \cdots + B_n X_{n,t} + u_{i,t}$$

Here, the probability that $Y_{i,t}$ is true conditional on the probability of the regressors $X_{i,t}$, is a linear function of the regressors X and the county fixed effects a_i. The coefficient estimate on B_1 is the change in probability of the occurrence of Y (Y = 1), given a marginal change in the regressor X_i, holding the effects of counties (a_i) and other X_{-i} regressors constant.

For predictive modelling, a PLPM with individual fixed effects, as specified above, was implemented using the PLM package in R (Croissant et al., 2022). A subset of data was used to train the PLPM, and then the trained model was used to make predictions of the occurrence of crop damage on a test dataset. Time fixed effects were excluded, as the future years that are only present in the testing set in an out-of-sample forecast cannot be estimated in the training dataset, which only contains data on existing years.

Predictions on the test set produced predicted probabilities of the occurrence of Y. The occurrence of Y is then determined by an arbitrary threshold of $P(Y = 1|X) > 0.5$, at probabilities of 0.5 and below the event Y is classified as "not occurring".

Data

Aggregated loss data from SHELDUS was linked with annual temperature and precipitation data by US counties. As the SHELDUS data was in monthly format, it was summarized to an annual frequency to match the frequency of the temperature and precipitation data. Table 1 contains the variables considered for this study.

Table 1 Variables extracted from the SHELDUS, Temperature and Precipitation datasets.

Variable	Definition	Type	Database
Crop Damage (D)	A value of 0 indicating US$0 in crop damages for a particular observation, 1 if damages exceed US$0	Binary	SHELDUS
Year (t)	Time dimension, spanning from 1979 to 2018	Time	SHELDUS
County (i)	Geographical index, counties selected for the two states; California (n = 58) and Iowa (n = 99)	Categorical	SHELDUS
Hazard (H)	Variable describing natural disaster/extreme weather event. For this study, drought was selected.	Categorical	SHELDUS
Temperature (W)	Temperature for a particular county in a particular year, measured in degrees Celsius.	Numeric	ERA5
Precipitation (P)	Precipitation for a particular county in a particular year, measured in millimeters.	Numeric	ERA5

The loss dataset from SHELDUS only included observations during the presence of hazards. As a result, years or counties without any corresponding hazards were missing from this dataset. To fill in these missing years and counties, the temperature and precipitation data (hereafter weather data) were merged with the SHELDUS loss data such that all observations from the temperature and precipitation data was retained (left-merging of tables, using year and county to link the 3 datasets). This essentially joined the weather information to any corresponding loss data observation by year and county, and joined any non-corresponding weather data to empty rows of observations (as a successful merge would require datasets to be of the same number of rows). The empty rows for year, state, and county was then filled with the corresponding value in the weather datasets, empty rows for Hazard was filled with "No hazard", and empty rows for crop damage was filled with zero. Lastly, the merged dataset was checked for duplicates, which were removed upon detection. This resulted in a panel dataset with a total of 6,280 observations, where the time spanned from 1979 to 2018, and included 157 counties (58 counties in California and 99 counties in Iowa).

Due to hardware limitations, data for two states was considered for the years between 1979 and 2018. California and Iowa were selected for this predictive exercise, as they were the largest crop producers in the US in 2021 (USDA, 2021). In 2021, California and Iowa had 11.8% and 8.0% share of US receipts for all US agricultural commodities. In 2022 dollars, these shares would be US$54 billion for California, and US$37 billion for Iowa of a total national production of US$463 billion. Being the largest producers of US agriculture, drought prediction would benefit the aforementioned states to prepare for shortages, as well as allow local and national governments to strategize in case of emergencies.

Table 2 and Figure 1 show the number of drought events in California and Iowa over the timespan considered for this study. In this timespan, while Iowa had 687 drought events, California had only 9. However, upon looking at drought events in these states by year, some additional observations are made. California experienced droughts in the years 2000, 2009 and 2014. Across these years there were 1, 3, and 5 droughts respectively. Although only 3 "Drought" years were identified, the frequency of droughts increased with each of the 3 occurrences. Iowa experienced over 90 droughts per "Drought" year in the late 1980s and early-middle 1990s. After this, number of droughts per "Drought" year in Iowa peaked at 72 droughts in 2003, fluctuating between 20 to 51 droughts per "Drought" year.

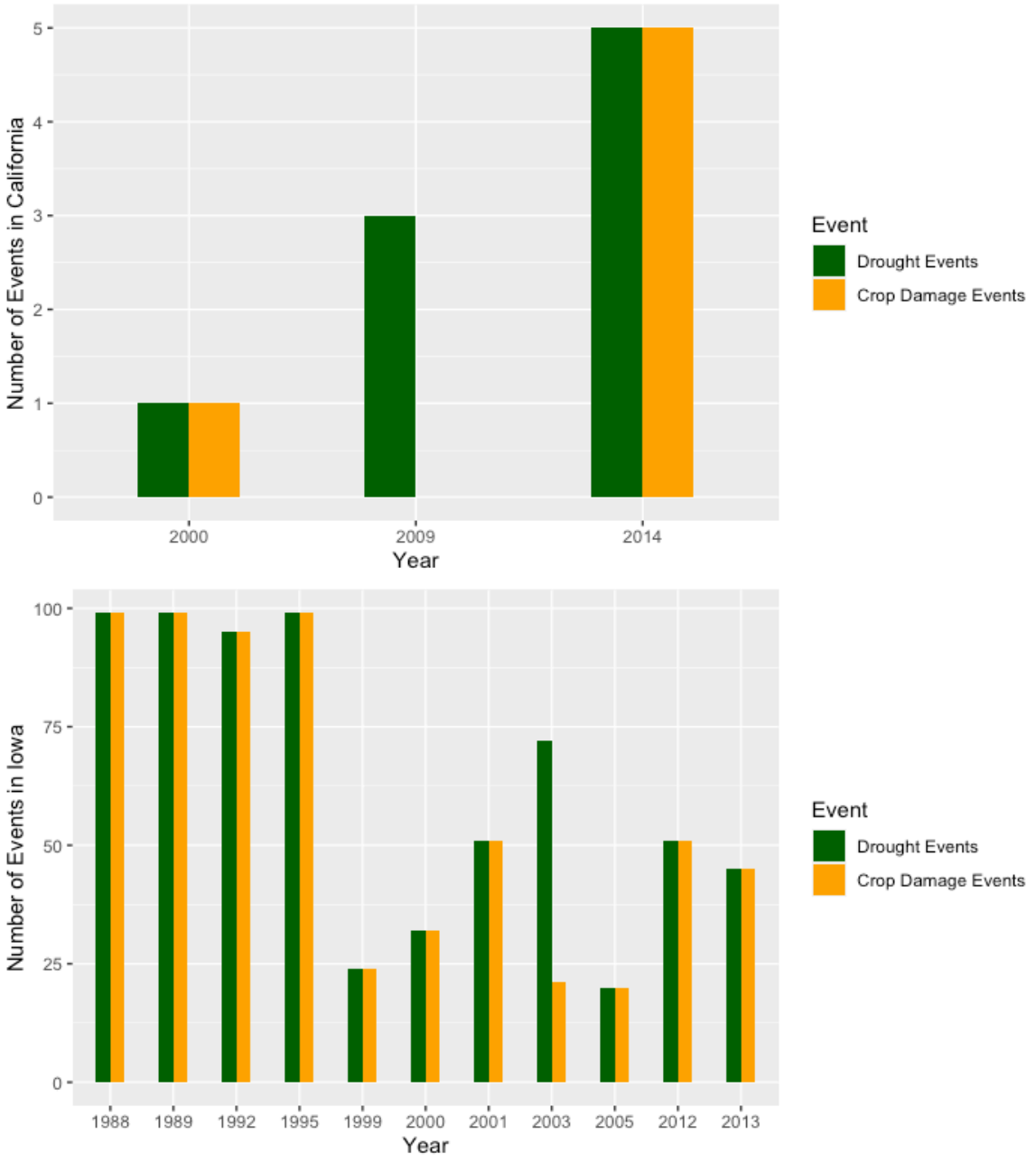
It is also observed that although a minority, not all drought events were associated with crop damage. The 3 droughts during the year 2000 in California were not associated with crop damage. In the case of Iowa, only 21 of 72 droughts had were associated with crop damage in the year 2003.

Table 2 (Top) Total Number of Droughts in California and Iowa between 1979 and 2018. (Bottom) Total Droughts and Crop Damage Events Disaggregated by Year.

State	Droughts
California	9
Iowa	687
Total	**696**

Year	State	Droughts	Crop damage
2000	California	1	1
2009	California	3	0
2014	California	5	5
1988	Iowa	99	99
1989	Iowa	99	99
1992	Iowa	95	95
1995	Iowa	99	99
1999	Iowa	24	24
2000	Iowa	32	32
2001	Iowa	51	51
2003	Iowa	72	21
2005	Iowa	20	20
2012	Iowa	51	51
2013	Iowa	45	45

Figure 1 Comparison of the Number of Drought Events and Crop Damage Events During Drought Years in; (Top) California and (Bottom) Iowa California.



The dataset above for California and Iowa, between the years 1979 to 2018 was then split into training and testing sets using an 80:20 train-test split ratio. The training data used a subset of data from the years 1979 to 2008 (n = 4710), while data from 2009 to 2018 was withheld for

the test set (n = 1570). Two models were then fit with the training data – a panel linear probability model (PLPM) and a random forest (RF). PLPM was used as a baseline predictive model, which was then compared with the RF for performance, similar to the work done by Chen (2021) which compared linear regression with machine learning.

The PLPM model was specified as below:

$$P(D_{i,t} = 1|W_{i,t}, P_{i,t}) = \alpha_i + \beta_1 P_{i,t} + \beta_2 W_{i,t} + \beta_3 W_{i,t}^2 + \beta_4 W_{i,t-1} + \beta_4 P_{i,t}^2 + \beta_5 P_{i,t-1} + \beta_6 W_{i,t} * P_{i,t} + u_{i,t}$$

Where the dependent variable $D_{i,t}$ stands for the binary crop damage variable and α_i are the county fixed effects. Variables denoted with W and P stand for temperature and precipitation, and their squares, lags, and an interaction is included in this model. The PLPM only considers county/individual fixed effects, as all counties appear in both training and testing sets. Hence, intercepts for each county can be calculated in the training set and used to predict crop damages in the testing set. A threshold of $P(D_{i,t} = 1) > 0.5$ was set to predict the occurrence of crop damages, whereas no crop damage was predicted at probabilities of 0.5 or below. Time fixed effects are excluded as the training set only includes years up to 2008, and the test set contains data from year 2009 to 2018. In a true out-of-sample prediction, time fixed effects would not be available as the data would be from the future (with no data available on input variables). As a result, any intercept calculated for time fixed effects would not apply to the withheld years of data available in the testing set.

The RF model was specified as follows:

$$D_{i,t} = f(i, W_{i,t}, W_{i,t}^2, W_{i,t-1}, P_{i,t}, P_{i,t}^2, P_{i,t-1}, W_{i,t} * P_{i,t})$$

Where the binary crop damage is a function of county, temperature, precipitation, and the lags, squared polynomials and interaction of temperature and precipitation. As stated in Breiman (2001), each tree that is grown either randomly selects either of the input variables to classify crop damage, or uses a random weighted linear combination of the 8 input variables, with weights picked from uniform random numbers between [-1,1].

One way of including time fixed effects would involve time demeaning the dataset used. However, this type of manipulation has the potential to influence predictions in the testing set. In this case, as all observations would be time demeaned, the same manipulation would occur in the training and testing data, and therefore lead to incomplete separation of the training and testing sets. As all data would be time demeaned, this would essentially "train" the test dataset regarding time fixed effects. This manipulation of test data before prediction cause issues with prediction when the model is used to predict new data which has not been time demeaned. Kapoor and Narayan (2002) extensively outline this issue as a type of Data Leakage, where "leakage" of information from the training set to the testing set caused reproducibility issues in 329 scholarly articles about civil war prediction. The authors found that once leakage was controlled for using model information sheets, the predictive performance of logistic regression and several machine learning models were not significantly different, contradicting the analysis of these papers investigated (Kapoor and Narayan, 2022).

Kapoor and Narayan (2022) identified 8 sources of data leakage, which they recommended checking for and documenting in model information sheets. The 8 types of data leakage, and the strategies used to counter them in this paper are outlined below:

Table 3 Model Information Sheet for the Predictive Models Used in this Paper

Type of leakage	Definition	Strategy
No test set	How training and testing sets are split on all steps of the modelling process.	Set a random number seed at the very beginning of the script. Split data into training and testing sets prior to fitting model and making predictions.
Preprocessing on training and test set	How training and testing data sets are separated during pre-processing and selection. This is to address leakage due to incorrect imputation.	Similar to above, all pre-processing is done for the training data, and predictions are made on separate unprocessed test data.
Feature selection on training and test set	How data is split into training and testing during feature selection.	Features strongly correlated with crop damages, such as occurrence of drought, duration of drought event and number of records were not selected as they lead to near perfect collinearity in the model.
Duplicates in datasets	How duplicates in data have been handled, if present.	Duplicates were scanned for and deleted.
Use of illegitimate features	Precisely explain that features and proxy variables are legitimate for the particular modelling task.	Features include US county code, precipitation and temperature levels in each county for the given timespan. No proxy variables were used. Temperature and precipitation patterns, and inclusion of county information legitimately provide information on rainfall and temperature across the US counties studied, which are relevant to study the extent of crop damages in the presence or absence of droughts.
Temporal leakage	Train and test timestamps must be separate, and test date should be at a later timestamp.	Training data takes a subset of the dataset from 1979 to 2008, which testing data subsets data from 2009 onwards. There is no overlap in timestamps between training and testing data.
Dependencies in training and test data	Reason and address dependencies that exist in data.	As mentioned in "Feature selection on training and testing set", variables highly correlated to crop damages were excluded from modelling. The top 8 features by predictive strength were included for modelling.

Sampling bias in test data distribution	Reason about selection bias, how rows were selected for analysis, and how test set matches the distribution about the scientific claims that are being made.	Selection bias may have been present if only observations during droughts were used. However, the dataset was infilled to include non-drought observations for the counties studied across the years, including the corresponding precipitation and temperatures.

Model Evaluation

A confusion matrix presents combinations of actual and predicted values of crop damage to assess the performance of a model as a classification model (Zhou & Liu, 2021). These components of the confusion matrix and measures that are functions of these components are used in this paper to compare the predictive performance of the two models tested (PLPM and RF). The following table presents the information organized on a confusion matrix.

Table 4 Components of a Confusion Matrix

		Actual Class	
		Crop Damage	**No Crop Damage**
Predicted Class	**Crop Damage**	True Positive (**TP**)	False Positive (**FP**) Type I Error
	No Crop Damage	False Negative (**FN**) Type II Error	True Negative (**TN**)

Each of the four quadrants present the correctly predicted crop damage events (True Positive) and non-crop damage events (True Negative), and the incorrectly predicted crop damage events (False Positive) and non-crop damage events (False Negative). The values on these four quadrants can be used to calculate the following model performance measures (Zhou & Liu, 2021):

Accuracy – the ratio of correct predictions to all predictions, used to measure the proportion of correct predictions by the model:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

Error rate – the ratio of incorrect predictions to all predictions:

$$\text{Error rate} = \frac{FP+FN}{TP+FP+TN+FN}$$

Precision – the ratio of true positive predictions to all positive predictions:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall (Sensitivity) – ratio of true positive predictions to the sum of correct positive predictions and incorrect negative predictions:

$$\text{Recall} = \frac{TP}{TP+FN}$$

Specificity – ratio of correct negative predictions to the sum of correct negative and incorrect positive predictions:

$$\text{Specificity} = \frac{TN}{TN+FP}$$

F1 – the harmonic mean of precision and recall, used as a combined measure of both of these metrics. Compared to arithmetic and geometric mean, the harmonic mean puts more emphasis on small values.

$$\frac{1}{F1} = \frac{1}{2}\left(\frac{1}{P} + \frac{1}{R}\right)$$

Results and Discussion

This section presents model summaries of based on the training data, predictions on the test data, and evaluates the two models that were fitted for prediction on the test set. Firstly, trained model summaries and confusion matrices for the Panel Linear Probability model are presented, followed by the Random Forest. Model performance based on predictions on the test data is then presented for both models.

Panel Linear Probability Model (PLPM)

As stated in the previous section, the PLPM used temperature, precipitation, as well as the first lags, polynomials and interactions of these two variables. Entity (county) level fixed effects were included in this model, while time fixed effects were excluded to prevent data leakage. The table below presents the model summary for the PLPM based on the training data.

Table 5 Summary statistics of PLPM model based on training data

	Dependent Variable
	Crop Damage (Binary)
Precipitation	-0.715***
	(0.043)
Temperature	-0.228***
	(0.027)
Temperature^2	0.004***
	(0.001)
Temperature (1st lag)	0.088***
	(0.006)
Precipitation^2	0.030***
	(0.004)
Precipitation (1st lag)	-0.050***
	(0.008)
Temperature × Precipitation	0.036***
	(0.003)
Observations	4,553
R2	0.154
Adjusted R2	0.122
F Statistic	113.811*** (df = 7, 4389)
Note:	*p<0.1; **p<0.05; ***p<0.01

The regression summary, such as input variable coefficient signs, magnitudes and standard errors, and related test statistics provide some context on the features of the PLPM model. From the regression table above, all inputs are significant at the 1% level ($p<0.01$), suggesting that each input variable is significantly related to crop damage. The F statistic of the training model is also significant at the 1% model, suggesting that at least one of the input variables significantly explains the variation in crop damage, rather than the baseline assumption that the dependent variable fluctuates randomly around a mean value. Taking the model as a whole, the inputs explain 15.4% of the variation in crop damage based on R^2, and 12.2% if penalizing for excessive regressors (Adjusted R^2).

Precipitation, its first lag, and temperature all have negative signs on their coefficients, suggesting a negative relationship between these variables and crop damage. This suggests that in this model, an increase in precipitation in the current or previous year, or an increase in temperature in the current year would lead to no crop damage. The negative relationship between precipitation and its lag with crop damage is intuitive, as a lack of moisture, particularly during droughts can be damaging to crops (Mann, Warner and Malik, 2019). However, it is surprising that an increase in temperature in the current year is not associated with crop damage in this model. Some potential explanations for this are that increased temperature on its own in the current year is not associated with the occurrence of crop damage, preventative measures against the effects of temperature fluctuations on crop damage in California and Iowa, or due to sample size issues or unobserved heterogeneity.

The squared polynomials of precipitation and temperature, the first lag of temperature, and the interaction between precipitation and temperature all show positive coefficients in this model. This suggests that for temperature and precipitation, there is a certain threshold beyond which increases and temperature and precipitation lead to the occurrence of crop damage. The positive coefficient on the first lag of temperature suggests that increased temperature in the previous year is associated the occurrence of crop damages in the following year. Lastly, the interaction between temperature and precipitation is also associated with crop damage occurrence, suggesting that concurrent rises in temperature and precipitation is positively related with crop damage.

The trained PLPM model was then tested on data withheld from the training set to obtain predictions of crop damage. The table below presents the confusion matrix for predicted values based on the PLPM. Out of 1,570 observations, there were 57 (3.63%) correctly predicted crop

damage events, and 741 (47.20%) correctly predicted events without crop damage. Furthermore, there were 44 (2.80%) false positive and 728 (46.37%) false negative events.

Table 6 Confusion Matrix based on PLPM

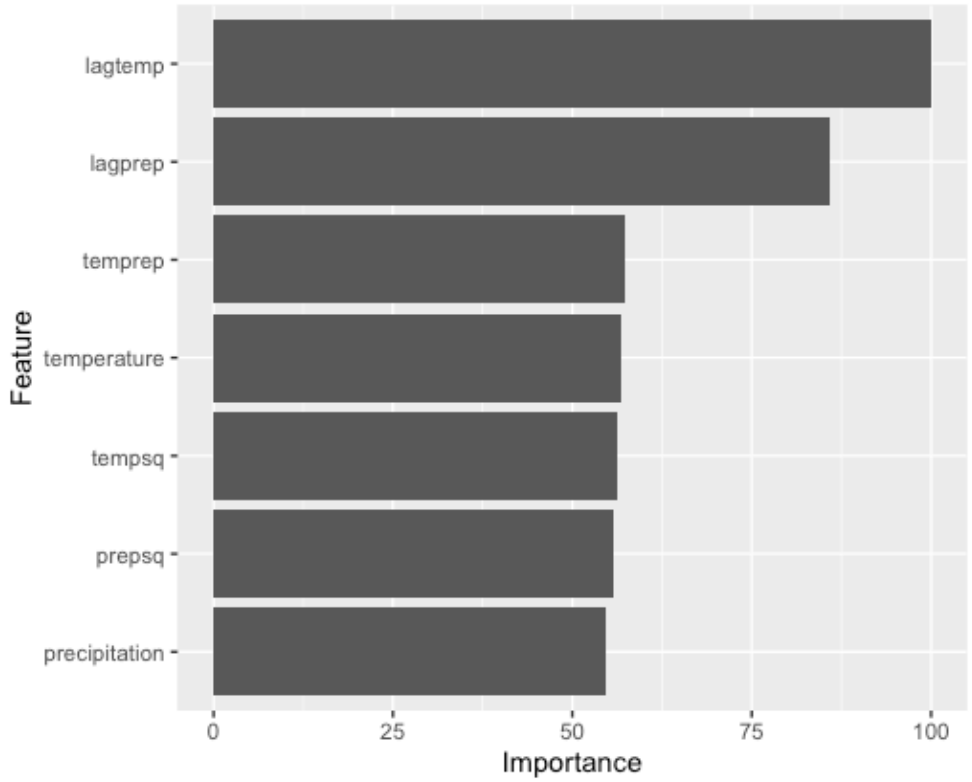
		Actual Class	
		Crop Damage	**No Crop Damage**
Predicted Class	**Crop Damage**	57	44
	No Crop Damage	728	741

Random Forest (RF)

Similar to the PLPM above, the random forest was trained to predict the occurrence of crop damage using inputs of county, temperature, precipitation, and the lags, squared polynomials and interaction of temperature and precipitation. The RF model was implemented using the CARET (Classification and Regression Training) package developed for the R programming language (Kuhn, 2022). Upon training, the importance of each input (feature) was obtained for the RF (Figure 1). The first lag of temperature was deemed the most important feature in classifying crop damage events, followed by the first lag of precipitation, the interaction of temperature and precipitation, temperature, temperature squared, precipitation squared, and lastly precipitation.

Breiman (2001) states that random forests either use randomly selected inputs or linear combinations of random inputs to split the ensemble of trees. RF aggregates ensemble votes over several random inputs and random linear combinations of inputs to determine feature importance. As a result, random selection of the first lags of temperature and precipitation, or a random linear combination of the top important features allowed for more accurate classification.

Figure 2 Features Ranked by Importance in RF



The confusion matrix below presents the predictions of the RF model in the withheld testing data. Notably, there were 34 (2.17%) correctly predicted crop damage events, fewer than the PLPM model. However, the RF predicted 1,418 (90.32%) non-crop damage events correctly, almost double of what was correctly predicted by the PLPM. The RF predicted 51(3.25%) false positive events, 7 (0.45%) more than the PLPM, while it only predicted 67 (4.27%, 42.10% fewer than PLPM) false negative events compared to the 728 (46.37%) false negatives predicted by PLPM.

Table 7 Confusion Matrix Based on RF

		Actual Class	
		Crop Damage	**No Crop Damage**
Predicted Class	**Crop Damage**	34	51
	No Crop Damage	67	1,418

Comparing Performance Metrics of PLPM and RF

The table below provides the results of the model performance metrics for both the PLPM and RF models. Upon making predictions on the test data, the RF made more accurate predictions with a smaller error rate, higher recall and marginally higher specificity than PLPM. The precision for PLPM was surprisingly higher than RF, but this was mainly due to the fact that RF produced fewer true positive predictions and slightly more false positive predictions than PLPM, leading PLPM being 16.44% more precise than RF. The F1, which is the harmonic mean of precision and was also higher for RF than PLPM, suggesting that RF had better performance when considering recall and precision simultaneously. Based on the performance metrics calculated, the RF fared better in making predictions compared to PLPM on five out of six metrics. The results from the performance metrics support the findings of Chen (2021) that machine learning models are superior to linear regressions in predictive modelling.

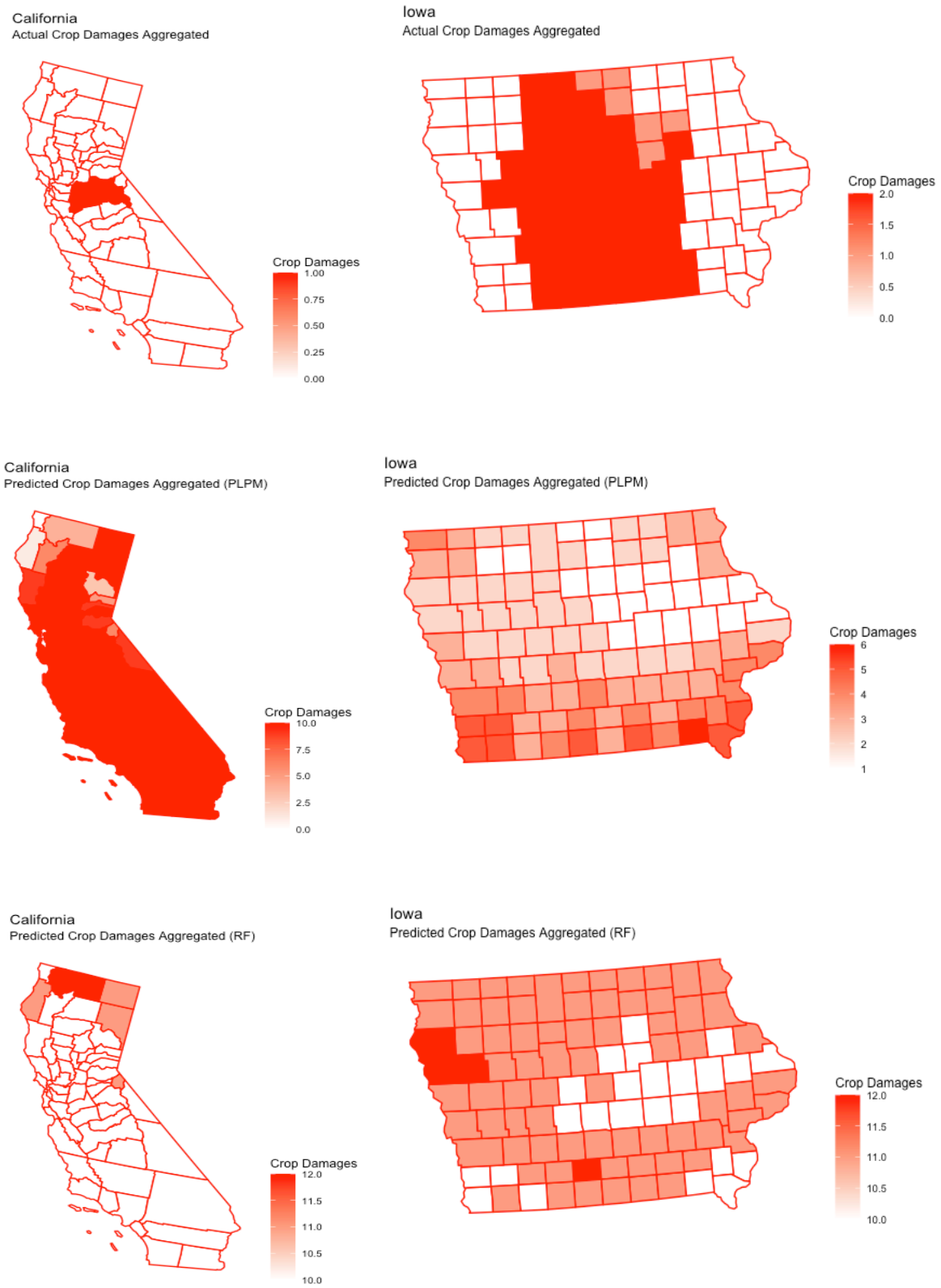
Table 5 Comparison of Model Performance Metrics of PLPM versus RF

Metric	PLPM	RF
Accuracy	50.83%	92.48%
Error rate	49.17%	7.52%
Precision	56.44%	40.00%
Recall	7.26%	33.66%
Specificity	94.39%	96.53%
F1	12.87%	36.56%

To visualize test set predictions for California and Iowa, heatmaps with actual and predicted crop damages are presented in Figure 2. Each map presents the total number of crop damage events in each state for the entire time span included in the test set. In the case of California, PLPM predicted more crop damage events across several states in California, which otherwise did not observe crop damages in reality. However, PLPM did predict crop damages in areas that did actually experience crop damage. RF did not predict such widespread crop damage events across Californian counties, but failed to predict crop damages in counties that actually experienced them. These observations can be explained by the lack of actual crop damage events in California, leading to RF predictions failing to precisely predict these events. Furthermore, the accurate predictions from PLPM can be attributed to chance, as it predicted crop damage across almost all counties in California.

In the case of Iowa, majority of crop damage events were found in high frequency in the central north and south of the state. PLPM predicted more frequent events in the south-east and south-west of Iowa, but predicted either less frequent events or no events in the central region. RF on the other hand did predict frequent floods in one of the central-south counties in Iowa as seen in the actual data, but also incorrectly predicted frequent crop damage events in north-west of Iowa.

Figure 2 Aggregated Crop Damages for California (Left) and Iowa (Right). Top: Actual Occurrences, Middle: PLPM Predictions, Bottom: RF Predictions



Conclusions and Future Directions

In this work, the RF predicted crop damages with a smaller error rate (7.52% versus 49.17%), more accuracy (92.48% versus 50.83%), and higher recall (33.66% versus 7.27%) with marginally higher specificity (96.36% versus 94.39%) than the PLPM. Although RF produced less precise predictions (40% versus 56.44%), the higher F1 score of RF (36.56% versus 12.87%) suggests that RF fared better when considering precision and recall simultaneously. Based on better performance in five of six metrics and by rationalizing the lower precision by way of the F1 score, RF outperformed PLPM in this exercise. These results agree with Chen's (2021) finding that machine learning models fare better in predictive modelling. However, the faster computation speed, higher precision and similar specificity of PLPM to RF also gives PLPM some merit as a predictive model.

The recall of the RF used in this model (33.66%), was lower than the recall for crop damage of teff (57%) in Mann, Warner, and Malik's (2019) paper, which was the cereal crop with the lowest recall. This can be attributed to several factors. This study does not include remote sensing data and data on biological characteristics of the crops. Furthermore, the dataset used in this study excludes information on the type of crops that were faced with droughts and experienced crop damages. Inclusion of these features would potentially improve recall of the predictions in the data subset for testing.

This work is an initial approach to predicting damages associated with natural disasters. Classification models were compared for their predictive performance. By definition, binary classification predicts either the presence or absence of a particular event. In this paper, crop damages encoded as "crop damage" versus 'no-crop damage" based on whether or not dollar damages were recorded in SHELDUS data. However, this type of predictive model is limited in predicting the extent of crop damages, which would be predictions of exact dollar amounts to

compare with actual dollar amounts of damage. Thus, a logical next step would be to either create a crop damage classifier with more than two levels (multi-class). A further step would be to design regression models to predict dollar amounts of damages instead of levels.

Due to hardware limitations, this study was limited to testing a panel linear probability model with a random forest model for prediction. Furthermore, only two states with high agricultural production was considered. This work can be expanded to test additional models, such as logistic regression, Adaboost, and other model along with all US counties and states for a more comprehensive study. With more data and additional models to test, further comparisons can be made in terms of predictions across different regions and more comprehensive tests against data leakage. In addition, the reproducibility of this model can be tested by using data from other countries and regions.

Lastly, additional inputs and dimensions would be beneficial in modelling crop damages. This essay excludes remote sensing data, visual inputs, and time series components seen in related works such as Mann, Warner and Malik (2019). Remote sensing data and visual inputs can provide additional information on crop characteristics such as quality, growth stage and visual anomalies that can be useful to detect potential crop damages. With regards to time, increased frequency, such as daily, weekly, or monthly data can provide information on repetitive trends (seasonality) or other fluctuations within a particular growth cycle of a crop.

References

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Cash receipts by Commodity State Ranking. USDA ERS - Data Products. (2021). Retrieved December 29, 2022, from https://data.ers.usda.gov/reports.aspx?ID=17844

Chen, J. M. (2021). An introduction to machine learning for panel data. *International Advances in Economic Research*, *27*(1), 1–16. https://doi.org/10.1007/s11294-021-09815-6

Climate change 2022: Impacts, adaptation and vulnerability. Intergovernmental Panel on Climate Change. (n.d.). Retrieved December 29, 2022, from https://www.ipcc.ch/report/ar6/wg2/

Costinot, A., Donaldson, D., & Smith, C. (2016). Evolving Comparative Advantage and the Impact of Climate Change in Agricultural Markets: Evidence from 1.7 Million Fields around the World. *The Journal of Political Economy*, 124(1), 205–248. https://doi.org/10.1086/684719

Croissant, Y., Millo, G., Toomet, O., Kleiber, C., Tappe, K., Zeilis, A., Henningsen, A., Andronic, L., & Schoenfelder, N. (2022, August 16). Linear models for panel data [R package PLM version 2.6-2]. The Comprehensive R Archive Network. Retrieved January 9, 2023, from https://cran.r-project.org/web/packages/plm/

Freund, Y., & Schapire, R. E. (1996). Experiments with a New Boosting Algorithm. In *icml*, 96, 148-156.

Kuhn, M. (2022, August 9). *Classification and regression training [R package caret version 6.0-93]*. The Comprehensive R Archive Network. Retrieved January 9, 2023, from https://cran.r-project.org/web/packages/caret/index.html

Mann, M. L., Warner, J. M., & Malik, A. S. (2019). Predicting high-magnitude, low-frequency crop losses using machine learning: An application to cereal crops in Ethiopia. *Climatic Change*, *154*(1-2), 211–227. https://doi.org/10.1007/s10584-019-02432-7

NASA. (2022, November 11). *The effects of climate change*. NASA. Retrieved December 29, 2022, from https://climate.nasa.gov/effects/

NASA. (2022, September 26). *Overview: Weather, Global Warming and climate change.* NASA. Retrieved December 29, 2022, from https://climate.nasa.gov/global-warming-vs-climate-change/#what_is_climate_change

Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., Boote, K. J., Folberth, C., Glotter, M., Khabarov, N., Neumann, K., Piontek, F., Pugh, T. A., Schmid, E., Stehfest, E., Yang, H., & Jones, J. W. (2013). Assessing agricultural risks of climate change in the 21st century in a global gridded Crop Model Intercomparison. *Proceedings of the National Academy of Sciences*, *111*(9), 3268–3273. https://doi.org/10.1073/pnas.1222463110

Schrimpf, P. (2022). *Machine Learning in Economics*. QuantEcon DataScience. Retrieved December 29, 2022, from https://datascience.quantecon.org/applications/ml_in_economics.html

Wooldridge, J. M. (2012). *Introductory econometrics: A modern approach*. Retrieved January 9, 2023, from https://economics.ut.ac.ir/documents/3030266/14100645/Jeffrey_M._Wooldridge_Introductory_Econometrics_A_Modern_Approach__2012.pdf

Zhou, Z.-H. (2021). Model Selection and Evaluation/ Performance Measures. In S. Liu (Trans.), *Machine Learning* (pp. 32–41). essay, Springer.