# Analysis of Climate Change and Water Access on the Production of Leading Agricultural Commodities in California

Gustav Hansen, Emily Jiang, Jason Yan

Modeling the Future Challenge 2020

**April 2020**

# Contents

# 1 Executive Summary

In recent years, climate change has been at the forefront of discussion regarding severe climate events in California. Between 2012 and 2016, California was mired in a drought of record duration and magnitude that devastated its agricultural industry, which accounts for 80 percent of state water usage. Total crop revenue losses during this time exceeded $45 billion [1]. Droughts and other causes of crop loss such as storms and wildfires are expected to rise in frequency as global warming accelerates. To understand how climate change and water access will impact Californian agriculture in the coming decades, our team seeks to provide mathematically-founded insights on this issue.

We specifically analyzed the production of almonds, avocados, and grapes, several leading state commodities. Using historical climate data from the National Oceanic and Atmospheric Administration, we selected nine factors: temperature, precipitation, atmospheric water vapor, snowfall, humidity, consecutive dry days, heat wave duration, and severe and moderate storm occurrences. We used these factors to build a random forest regression model that predicts crop loss as a proportion for every county in California, which we trained on crop loss data from the USDA Risk Management Agency. By inputting climate predictions from NOAA models, we projected yearly loss for almonds, avocados, and grapes from 2020 to 2100. We predict that while average crop loss will only slightly rise over time, losses in “outlier” years will increase in magnitude, aligning with the expectation that climate change leads to more drastic weather spikes. We also determined that heat wave duration, number of consecutive dry days (drought), and temperature most greatly influence crop loss. Exploratory factor analysis showed that our nine climate variables could be linearly combined into three uncorrelated components that together accounted for 77.4% of the climate model’s variability.

The results of our random forest model allowed us to conduct a regional risk analysis, which indicated that most counties can expect to witness increases in annual crop loss between 5 to 15% by 2100. Kern, Ventura, Santa Barbara, Stanislaus, and Los Angeles were particularly high-risk counties, with Kern projected to have the highest percent change in crop loss of 18.14%. Kern, Stanislaus, and Ventura had the greatest potential economic losses, of approximately $1.3 million, $350,000, and $250,000 per year respectively. We also identified several at-risk organizations: economies of predominantly agrarian communities, private insurers such as Global Ag, low socioeconomic status and minority farmers, as well as ancillary industries.

Based on our analysis, we outlined insurance and public policy recommendations to the RMA and other federal agencies. Specifically, we suggest the extension of several safety net programs to cover currently-excluded fruits and vegetables, including the 3 crops we studied. To improve affordability of crop insurance, we propose implementing a more progressive pricing model for subsidizing premiums that would also reduce government costs. We propose expanding programs that incentivize eco-friendly, water-conserving farming practices to encourage sustainable agriculture and greater crop insurance participation in California. As climate change becomes an increasingly formidable challenge faced by farmers nationwide, we believe our models and recommendations provide novel insight and the potential to benefit one of the United States’ most integral industries.

# 2 Background Information

For both developed and developing nations, agriculture has always been the backbone of civilization. The cultivation of crops and rearing of livestock not only feed an expanding population but also play a crucial role in a country's economy, providing jobs (nearly 1 million in the United States), income, and "homegrown security" when conflict with other nations restricts crop trade [2]. In 2017, California's 77,100 farms and ranches generated over $50 billion, with the leading commodities of dairy, grapes, and almonds [3]. For a state that produces 82% of the world's almonds, the importance of preserving and protecting its agricultural industry cannot be understated [4].

Given that output optimization in agriculture relies on a predictable, consistent climate, climate change is becoming an ever-increasing threat to crop production worldwide. Climate change has given rise to and accelerated global warming, droughts, rising sea levels, and severe weather events, which are expected to only worsen in the upcoming decades. Notably, climate change pushes weather to both opposite extremes, resulting in twofold consequences. Consequently, floods are waterlogging soil at the same time that heat and drought are destroying crops in other regions. Heat and drought also work in tandem with severe windstorms to blow away topsoil and expose the roots of crops. The US Department of Agriculture estimates that a staggering 90 percent of crop losses are related to extreme weather [5]. California in particular has been a hotspot for climate change-induced agricultural crises in recent years—subject to a 376-week-long drought from 2011 to 2019, rising sea levels along the coast, as well as heat waves, smog, and wildfires [6]. Still worsening matters is California's continued dependence on fossil fuels like natural gas and oil for energy, which release carbon dioxide, methane, and other gases to the atmosphere when burned and further contribute to the greenhouse effect [7].

Agriculture also is heavily dependent on freshwater access for irrigating crops and raising livestock; the industry accounts for approximately 80 percent of the United States' ground and surface water use and over 90 percent in many Western states such as California [8]. Freshwater access is not only affected by depletion and contamination through human extraction but is also exacerbated by climate change. Glaciers, which store about 69 percent of the world's freshwater, have been rapidly melting throughout the past century. If all land ice melted, the global sea level would rise approximately 70 meters [9]. This would not only contaminate groundwater supplies with saltwater but also flood coastal farms and communities. In addition, increased droughts caused by climate change dry out existing wells and necessitate the drilling of deeper and deeper wells to extract dwindling groundwater reservoirs, often resulting in permanent drops in the water table—an indication that the rate of groundwater extraction through wells is higher than the rate of replenishment via precipitation [10].

To protect farmers and ranchers against the effects of climate volatility, crop insurance is partially subsidized by the federal government and covers crop loss due to natural disasters as well as revenue loss due to declines in the sale prices of their output. Several major types of crop insurance are multiple peril crop insurance (MPCI), crop-hail insurance, and revenue insurance. MPCI policies are available for over 120 different crops, and cover losses caused by natural events such as hail, frost, wind, disease, drought, fire, flooding, and insect damage [11]. These MPCI policies are written and insured by private companies that have been authorized by the USDA Risk Management Agency (RMA) under the Federal Crop Insurance Program's public-private partnership. To make insurance more affordable, the federal government subsidizes the farmer-paid premiums. Meanwhile, crop-hail insurance is purchased from private insurers in regions experiencing frequent hailstorms and is not part of the Federal Crop Insurance Program. In California, where hail is a relatively rare event, this type of insurance is not significantly utilized. The final type, crop revenue insurance, pays farmers indemnities according to fluctuations in annual revenues compared to previous years in order to protect against drastic swings in crop prices [11].

More than ever, crop insurance is vital to California's agricultural sector, protecting 6.7 million acres and providing \$8.4 billion in liability protection in 2018. That year, private crop-hail insurance provided an additional \$27.1 million in liability protection in the state [12]. In California and other agricultural regions in the US, the most common causes of loss for federal crop insurance policies tend to be excess precipitation and flood, drought, heat and cold, wind, and hail [13]. While protecting crops is important for farmers in every state, it is in many ways even more crucial for Californian growers. California leads the nation in 75 crops and livestock commodities, and is the *sole national producer* (99 percent or more) of 14 crops, including almonds, artichokes, raisin grapes, kiwifruit, olives, and pistachios [3]. Thus, risks to California's leading commodities, if sufficiently severe, would cripple the entire nationwide supply of that crop, given that in many cases, there are no ”backup” states to supplement the Californian supply.

Recognizing the unique risks and concomitant challenges with being such a vital agricultural state, this report seeks to first analyze the impact of climate change and water access on the production of three of California's major commodities—grapes, almonds, and avocados—in the upcoming decades. Our model results can then be used to determine the organizations at greatest risk as a result of the predicted changes, and provide recommendations on how to adapt to them. California is a particularly crucial region of the United States to study because of its aforementioned vulnerabilities to the most damaging ramifications of climate change. As the state persists through its decade-long drought, temperatures and storm frequency climb steadily, and its freshwater supplies continue to dwindle, a deeper understanding of both the underlying climate change and its agricultural ramifications is necessary.

# 3 Model Development and Results

Numerous aspects of climate can be quantified, not all of which impact crop growth and agricultural loss/yield equally. In this section, we develop, implement, and test a mathematical model to project crop loss of grape, almond, and avocado farmland in California over the next decades using climate parameters.

## 3.1 Data Methodology

Effective analysis of the impact of climate change and water access on agriculture in California requires comprehensive state climate and crop data covering at least several decades. We utilized data from NOAA's Geophysical Fluid Dynamics Laboratory (GFDL), which provides a database of climate summaries from land surface stations worldwide. GFDL creates global climate models (CM) containing data on many aspects of climate, including historical data from 1861 to the current year and predictions made by the laboratory up until 2100. Using data from these models was preferable to using NOAA's other climate datasets because it seamlessly transitions from past data to future projections, allowing us to train a model on historical climate and apply it directly to another period of time.

We utilize this pre-existing climate model to analyze the effects of climate change on the agricultural industry because GFDL is an esteemed laboratory in a national institute for oceanic and atmospheric research, responsible for the development of the first climate models to study global warming. Thus, their climate models are both comprehensive and highly reliable. From the GFDL database, we used the CM2.1 model, which is part of GFDL's contribution to the Coupled Model Intercomparison Project (CMIP3), an international endeavor headed by the World Climate Research Program [14]. The CM2.1 model contains data for each of many climate factors at given latitude-longitude coordinates, in increments of 200 km in both directions (approximately 1.5 latitude and 2.5 longitude increments) [15].

In order to construct a model that predicts agricultural yield from various climate factors, we need to train the model on existing loss data for the crops of interest. For this, we use federal crop insurance policy data from the USDA Risk Management Agency (RMA), which provides information on the total acreage per county of specific crops that had a loss each month due to various reasons [16]. RMA contains data from 1989 up until present day. From the RMA Report Generator tool, we can calculate the total number of insured acres per county [17]. As a government agency, the USDA dataset is reliable, and therefore useful in providing crop insurance loss data for the model.

## 3.2 Mathematics Methodology

### 3.2.1 Assumptions and Justifications

1. *The current trend of climate change progression remains constant.* We are unable to assume any major scientific or technological advances that would significantly reduce the effects of climate change, nor any major events that would significantly increase

them. Thus, we assume that current climate projections (e.g. of increasing temperatures, more droughts, and more severe storms) are accurate.

2. *This model analyzes the effects of climate change and water access on grape, almond, and avocado production in California.* These three major crops comprise a significant portion of the state's agricultural production [18]. While the remaining commodities are not negligible, we limit the scope of our investigation in order to make more specific predictions and recommendations relating to these industries.

3. *Earth is a perfectly spherical planet.* This assumption allows us to calculate distances using the Haversine formula for angular distances between two points on a sphere.

4. *Environmental conditions vary smoothly within 200 km increments.* It is unlikely that drastic changes in the values of climate variables will occur within these relatively small regions. This assumption facilitates the usage of an inverse distance weighting formula to estimate climate factors between discrete grid points. The increment value of 200 km was chosen because it is the width of the intervals at which the NOAA GFDL climate data is recorded.

5. *Only insured farms are considered.* As our data on historic crop loss is taken from the USDA's information on federal crop insurance policies, this assumption is necessary because we do not have data on uninsured farms. We also seek to provide insurance policy recommendations, which are only applicable to insured farms.

6. *Wildfire is not considered in our model.* According to the Los Angeles Times, "84% of U.S. wildfires were caused by human-related activity" [19]. Since wildfires are caused by human actions much more frequently than by purely natural causes, we decided not to incorporate wildfires into our model as we are primarily concerned with the impact of climate on the agricultural industry.

### 3.2.2 Model Development

We implemented a random forest model to predict crop loss when given a number of independent input variables, including temperature, precipitation, and other factors. A random forest regression model constructs an ensemble of regression trees (decision trees with continuous target variables) and outputs the mean value of the predictions made by the individual trees. Each regression tree in the forest takes a random subset of input variables with replacement to predict the output variable—in this case, the proportion of crops lost per acre. We utilize a random forest rather than a single regression tree to reduce overfitting and improve accuracy and robustness of the model.

As detailed in Section 3.1, we used the CM2.1 database from GFDL as our input data, which contains data for climate factors at given latitude-longitude coordinates in increments of 200 km on each side. However, the 200 km increments of environmental measurements are not precise enough to capture all of the counties in California, many of which are less than 200 km across. Since our intention was to analyze climate by county, we used the Python package geopy to convert each of the 58 counties in California to latitude and longitude coordinates according to the geographic center (centroid) of each. We then calculated the Haversine distance—the angular distance between two points on a sphere—to each of the four nearest coordinates at which the GFDL created predictions for. The Haversine distance is calculated as follows:

$$d = 2r \arcsin(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}) \quad (1)$$

We used the Python package Haversine to calculate the $\phi$ and $\lambda$ values from latitude-longitude coordinates. We then estimated the values of each climate variable at the county centroids by applying an inverse-distance weighting formula, which allowed us to compute a weighted average of the 4 surrounding grid points. The inverse-distance weighting formula is defined as follows:

$$x_{county} = \frac{\sum_{i=1}^{4} \frac{x_i}{d_i}}{\sum_{i=1}^{4} d_i} \quad (2)$$

where x is the value of any climate variable in the model and d is the calculated Haversine distance from the centroid to a given grid point. Figure 3.1.1 visualizes our climate estimation method using an outline of Alameda County as an example.

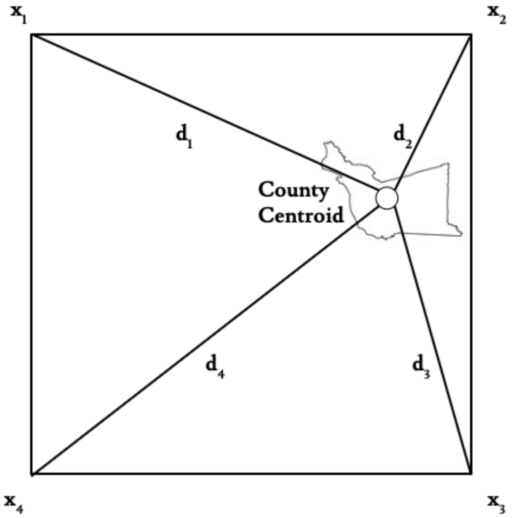


**Figure 3.2.1**: Diagram of Climate Estimation Method for Sample County (not to scale)

The GFDL database contained data in more factors than we deemed necessary to include in our model, so we selected elements that prior research showed to be the most impactful on crop output. Our final annual climate factors are listed in Table 3.2.1 below. The selected factors were chosen to incorporate both common meteorological parameters such as precipitation and average temperature as well as variables to represent extreme weather events that affect crop output.

**Table 3.2.1**: Annual Climate Variables for Random Forest Model

| **Symbol** | **Definition** | **Units** |
|---|---|---|
| TAS | Average daily temperature | °C |
| PR | Total annual precipitation (including snow and hail) | kg/m$^2$/$s$ |
| R95T | Average daily atmospheric water vapor content | cm |
| PRSN | Total annual snowfall | mm |
| HUR | Relative humidity, expressed as a decimal | N/A |
| CDD | Highest number of consecutive dry days | days |
| HWDI | Heat wave duration index | N/A |
| PRW | Percent of annual precipitation from large events, expressed as a decimal between 0 and 1 | N/A |
| R10 | Number of days with precipitation exceeding 10 mm | days |

While the input of the random forest model is climate data, taken from the CM2.1 database and preprocessed as aforementioned, the dependent variable is the percent crop loss of farmland, which is calculated by the formula:

$$Loss = \frac{NDA}{TA} \tag{3}$$

where *NDA* is the Net Determined Acres and *TA* is the total number of insured acres of the crop of interest. *NDA* is taken from the RMA cause of loss files, and is defined as the number of acres of loss due to damage after the insured's share is applied. *TA* is taken from the RMA report generator and is calculated as the sum of all acres covered by crop insurance policies, including those that did not have losses. For all available years (1989 to 2019), we computed percent loss per county for each of the three commodities of almonds, grapes, and avocados.

We wrote a Python program (see Code Appendix) to run our preprocessed data through a random forest regression model with 100 trees. Although the CM2.1 dataset contains climate data starting from 1861, the RMA database only contains loss data starting from 1989, so we were only able to use data for 1989 and later from the CM2.1 dataset. The historical data (1989 through 2019) were then randomly sorted and split into 2/3 for training the model and 1/3 for assessing accuracy. Since the CM2.1 dataset contains not only historical values but also predictions for all climate variables until the year 2100, we applied our random forest model to the predicted climate data in order to project future loss of the three crops.

## 3.3 Results and Model Analysis

### 3.3.1 Model Output

Executing our Python program generated 100 regression trees that individually predicted the proportion for crop loss in each county. Figure 3.3.1 diagrams the flow of logic for a sample regression tree in the random forest. Each box except for those in the last row contains a conditional statement; the result of those conditional statements dictate how the algorithm gradually modifies its prediction to arrive at the final value on the bottom row. For example, in this sample tree, if the first conditional statement (HWDI $\leq 2.05$) evaluates to True, the model modifies its prediction from 0.12 to 0.29, and so on until a box on the final row is reached.

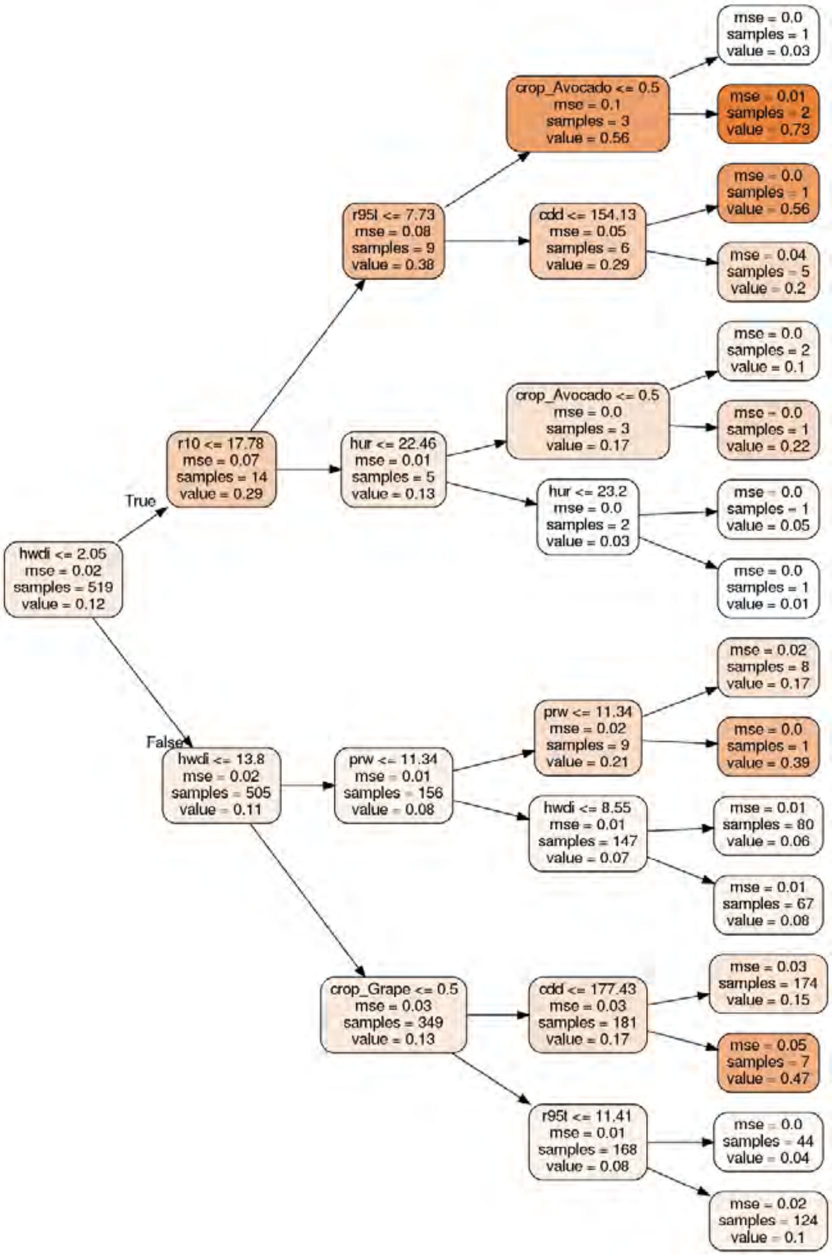


**Figure 3.3.1**: Visualization of Sample Single Regression Tree in Random Forest

Testing the trained model on the reserved 1/3 of historical data resulted in a relatively low mean squared error of 0.08. We then inputted the CM2.1 predicted values to project future crop loss by county. Figures 3.3.2, 3.3.3, and 3.3.4 show scatterplots of our random forest's projected average proportion of crop loss (across all counties) from 2020 to 2100.

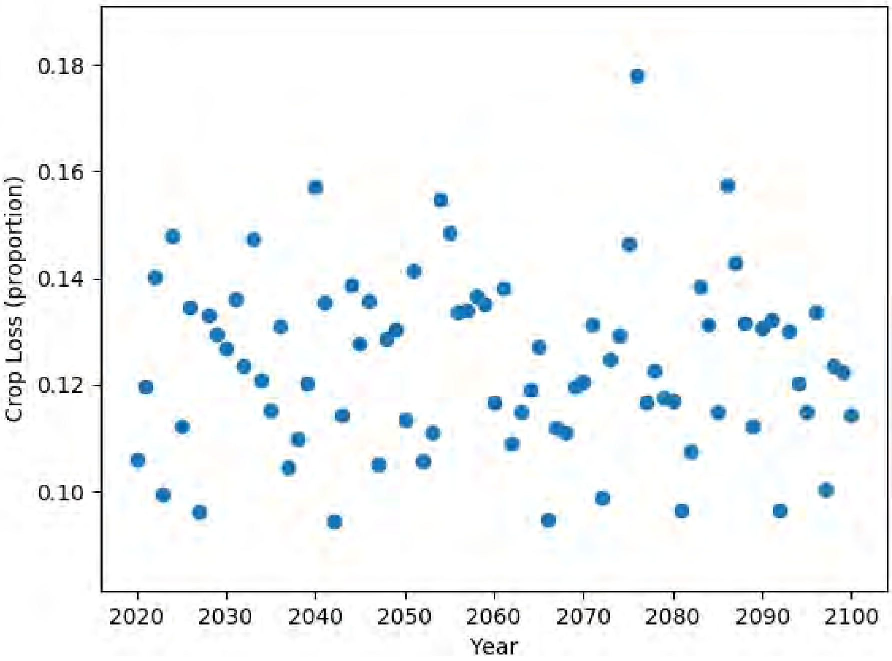


**Figure 3.3.2**: Projected Average Proportion of Crop Loss for Almonds Over Time

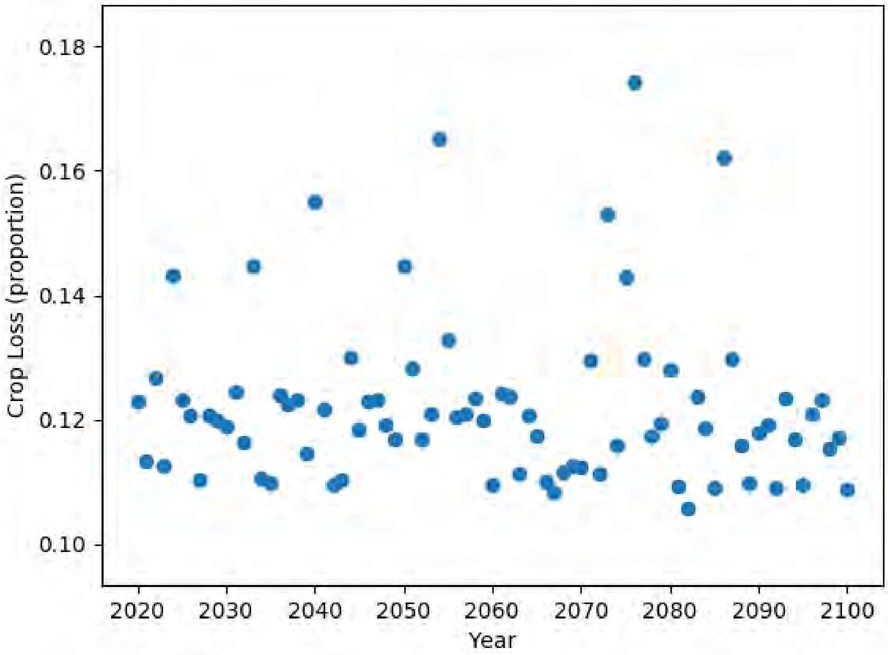


**Figure 3.3.3**: Projected Average Proportion of Crop Loss for Avocados Over Time

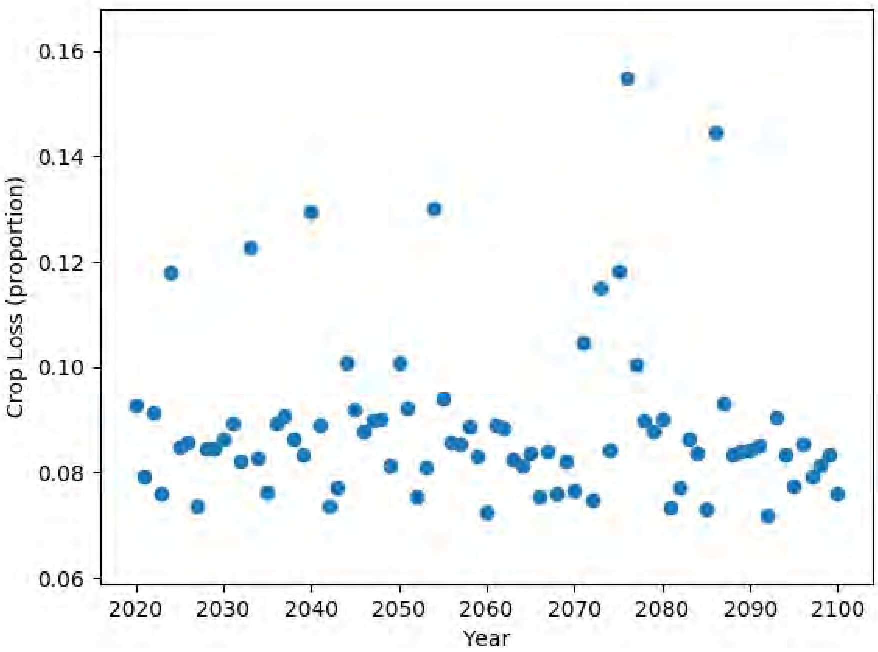


**Figure 3.3.4**: Projected Average Proportion of Crop Loss for Grapes Over Time

We also calculated and recorded in Table 3.3.2 the unusual values, data points that fell outside of 2 standard deviations of the mean, in each of the above three figures. For each crop, the amount of crop loss in ”unusual” years generally rises over time in value. Although there are several unusual years that are in common among the three crops, the years still differ for each crop because they respond differently to changing climate conditions. These unusual or ”outlying” points represent the years that are projected to cause the most harm to farmers. For instance, almond and avocado farmers may lose nearly 18% of their crops in 2076.

**Table 3.3.2**: Unusual Data Points in Crop Loss Predictions

| Year | Almond | Avocado | Grape |
|---|---|---|---|
| 2033 | | | 0.1226 |
| 2040 | 0.1570 | 0.1551 | 0.1296 |
| 2054 | | 0.1650 | 0.1301 |
| 2073 | | 0.1530 | |
| 2076 | 0.1780 | 0.1744 | 0.1549 |
| 2086 | 0.1574 | 0.1621 | 0.1446 |

After obtaining the prediction results, our model returned the feature importances of the nine climate factors, as reported and ranked in Table 3.3.3. Our model indicated that heat wave duration index ($HWDI$), highest number of consecutive dry days ($CDD$), and average daily temperature ($TAS$) are the three most influential features in determining loss of the selected crops, while average daily atmospheric water vapor ($R95T$), total annual precipitation ($PR$), and total annual snowfall ($PRSN$) are relatively the least influential variables out of the nine factors we analyzed. From this, we infer that our model is sound, as it is logical for daily temperature as well as excessive heat and dryness to most greatly impact crop loss—these factors relate directly to a crop's basic requirements for survival. Furthermore, it is relatively unlikely that water vapor content and snowfall to significantly impact crop loss, especially since crops are not typically grown in the wintertime when snowfall most often occurs; the same reasoning applies to total annual precipitation since this variable includes snowfall.

**Table 3.3.3**: Importance of Climate Features in Random Forest Model

| Symbol | Feature Importance | Importance Rank |
|---|---|---|
| HWDI | 0.1264 | 1 |
| CDD | 0.1138 | 2 |
| TAS | 0.1079 | 3 |
| PRW | 0.1034 | 4 |
| R10 | 0.1031 | 5 |
| HUR | 0.0981 | 6 |
| R95T | 0.0962 | 7 |
| PR | 0.0894 | 8 |
| PRSN | 0.0772 | 9 |

### 3.3.2 Analysis of Climate Factors

Recognizing that the graphs of predicted crop loss (Figures 3.3.2, 3.3.3, 3.3.4) exhibit no apparent trend, we sought to analyze the most impactful variables in our model. As aforementioned, heat wave duration index (*HWDI*), highest number of consecutive dry days (*CDD*), and average daily temperature (*TAS*) are the three most influential features. To investigate how these variables were individually influencing predicted crop loss, we graphed the predictions for each of them in Figure 3.3.6, using predictions for 2020 through 2100 taken directly from the GFDL CM2.1 model, as these were the inputs in our random forest model. While all three plots trend slightly upward, there is a significant amount of scatter that suggests only a very weak correlation between time and each of the three variables. However, these predictions do align with current scientific notions regarding the progression of climate change—while global warming is certainly cause for concern and threatens the delicate equilibrium of ecosystems, temperature and drought increases throughout the century are still slow.

Since the graphs of our three most important model factors in Figure 3.3.6 all have a large amount of scatter and only weak trends over time, we conclude that our crop loss predictions in Figures 3.3.2, 3.3.3, and 3.3.4 are reasonable with respect to available data. We also observed that the high outliers in the loss predictions for almonds, avocados, and grapes appeared to increase in magnitude over time. This observation is in agreement with the expectation that climate change will increase the frequency and severity of severe weather events, such as hurricanes, floods, heat waves, and droughts, that threaten crop growth. Finally, the predictions for all three crops appear to be very similar from their scatter plots, which suggests that between almonds, avocados, and grapes, none of the three crops are significantly more/less resistant to climate change than the others.

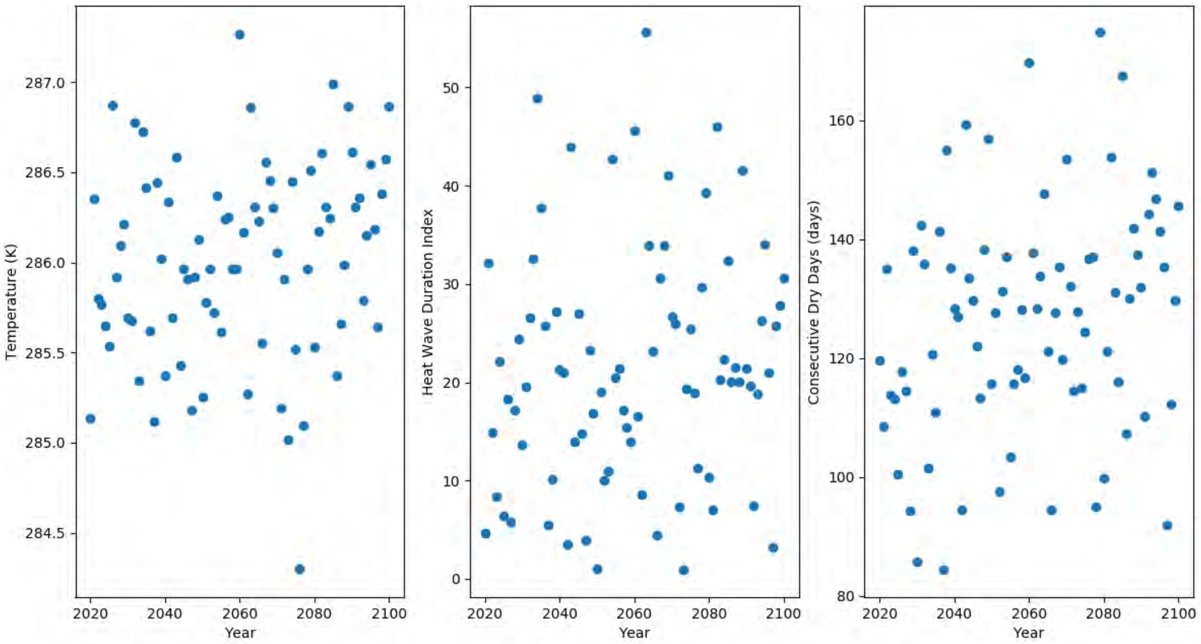


**Figure 3.3.6**: GFDL CM2.1 Predictions of *TAS*, *HWDI*, and *CDD* Over Time

### 3.3.3 Exploratory Factor Analysis of Climate Model

In creating our model, we implemented the CM2.1 dataset without considering possible underlying relationships between the climate variables; this was because random forest models do not require independent data to make reliable predictions. However, uncovering the underlying structure of our model's inputs can assist in the creation of legislation to target different aspects of climate with minimum redundancy. To this end, we ran an exploratory factor analysis (EFA) to identify any defining climate features in our model that influence the others. The first step of EFA is to run a principal component analysis (PCA). PCA is an operation that reduces the dimensionality of data while still accounting for most of its variability. The principal components obtained by PCA form an uncorrelated basis set that fully describe the data. However, it should be noted that principal components, as linear combinations of the input features, have no intrinsic meaning and are unitless.

First, we standardized each feature to have a mean of 0 and standard deviation of 1 because PCA is sensitive to the scale of the data. The covariance matrix $A$ was created by computing the covariance, as defined below, between every combination of factors:

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(Xi-\overline{x})(Yi-\overline{y}) \tag{4}$$

We then computed the eigenvalues of matrix A and sort the eigenvectors by their eigenvalues in decreasing order. In order to choose the correct number of principal components that can explain the data, we use the Kaiser criterion: the number of principal components that have an eigenvalue greater than 1. This resulted in three principal components:

*PC1*, *PC2*, and *PC3*. Then, we can create a transformation matrix $W$ with the eigenvectors as columns, and compute the new principal components as follows:

$$y = W^T \times x \tag{5}$$

We then obtain factor loadings, which represent the correlation of each variable with the principal components, by multiplying each eigenvector in the transformation matrix $W$ by the square root of its corresponding eigenvalue. However, these loadings do not have much separation and are all close to around 0.5. To obtain high factor loadings for a few variables (i.e. to obtain the underlying features that can describe the entire data), we apply a varimax rotation, which maximizes the sum of the variance of the squared loadings.

For each variable, the rotated loading with magnitude greater than 0.5 was assigned to its corresponding principal component. Finally, we obtained the communalities, which represent the percent of variance that can be explained by the principal components. These were obtained by squaring the loading of our chosen principal components for each variable, as shown below.

$$C_{var} = \sum_{i=1}^{n}(PC_i^2) \tag{6}$$

If for each variable we squared all the principal components, not just those that we chose, we would obtain 1. Our exploratory factor analysis results are summarized in Table 3.3.3, including the rotated loading factor, the communality of each feature, and the variance explained by each principal component.

**Table 3.3.4**: Loading Values and Communalities of Climate Factors

| **Variable** | **PC1** | **PC2** | **PC3** | **Communality** |
|---|---|---|---|---|
| TAS | **0.95** | -0.21 | 0.04 | 0.95 |
| PRW | **0.91** | 0.16 | -0.2 | 0.9 |
| PRSN | **-0.9** | 0.28 | 0.05 | 0.88 |
| CDD | **0.56** | -0.25 | 0.39 | 0.53 |
| PR | -0.47 | **0.81** | -0.13 | 0.91 |
| R10 | -0.43 | **0.81** | -0.11 | 0.85 |
| HUR | -0.04 | **0.75** | -0.26 | 0.64 |
| R95T | 0.16 | **0.67** | 0.31 | 0.57 |
| HWDI | -0.11 | -0.05 | **0.85** | 0.74 |
| **PC variance %** | **45.8** | **20.3** | **11.3** | |

The results of EFA suggest that the climate variables we chose can be primarily grouped into three uncorrelated components that together describe 45.8 + 20.3 + 11.3 = 77.4% of the model's variability. From Table 3.3.3 above, we can see that the most influential component PC1 is primarily comprised of the average temperature, the atmospheric water vapor content, and snowfall precipitation, all having loading magnitudes greater than 0.90. The cumulative dry days variable is more evenly spread across the three components with a low communality of 0.53, indicating that it is not so well described by a single principal component. Overall, *PC1* generally describes temperature and atmospheric water content, through water vapor and snow. *PC2*, which generally describes precipitation, consists of precipitation, days of more than 10 mm of rainfall, relative humidity, and percentage of precipitation from large events, each with relatively high communalities. Finally, *PC3* only contains the heat wave duration index, and by itself describes extreme heat events.

These component classifications indicate that *PC1*, *PC2*, and *PC3* each contain uncorrelated, independent data. The EFA results allow us to conclude that the key uncorrelated factors affecting variability in climate data are temperature and atmospheric water content, precipitation, and extreme heat events. In context, this is useful in informing policymakers of which factors independently affect climate; such knowledge points to areas of focus for legislation to create maximum impact. The relatively low cumulative PC variance of 77.4% indicates, as expected, that climate is a complex phenomenon, and that a large number of principal components would be required to fully represent it. However, the fact that we were able to account for a majority of climate variability with only three independent components validates our choice of the nine climate variables in our random forest model, suggesting that our manual selection was effective.

### 3.3.4 Summary of Model and Analysis Results

We first obtained our model results by running the random forest regression model built in Section 3.2 with 100 trees. This resulted in three graphs for projected average crop loss per year from 2020 to 2100, one for each of the crops almonds, avocados, and grapes. For each crop's graph, we calculated and discussed the physical significance of the statistically unusual points, and evaluated the sensibility of the model output, determining that the results were reasonable. We also analyzed time-based trends in the three most influential climate variables (average daily temperature—TAS, heat wave duration index—HWDI, and highest number of consecutive dry days—CDD); we noted that the very slight upward trends in these three climate variables over time corresponded logically to the relatively small magnitude of trends in our random forest model output. Then, we performed exploratory factor analysis (EFA) with principal component analysis (PCA) to analyze any underlying relationships, or covariance, between our input variables. EFA resulted in 3 principal components that each contained uncorrelated data, which allowed us to conclude that the key independent factors affecting climate data variability are temperature/atmospheric water vapor, precipitation, and extreme heat events. Collectively, the model results, conclusions, and analysis of input (climate) variables highlight meaningful targets for agricultural and climate legislation.

### 3.3.5 Strengths and Weaknesses

Our model is strong and outputs sensible results, as described previously. In particular, implementing a random forest model mitigates the tendency of single regression trees to overfit to the data set on which they are trained; this supports both the robustness and the resilience of our model, as the final prediction outputted by the forest is the mean of many (100) individual regression tree predictions. We were also able to determine the relative impact of the nine climate factors that we considered via the feature importances obtained from the model. Such information is useful in risk analysis because it highlights the aspects of climate that have the greatest influence on crop yield and should therefore be addressed using insurance and public policy changes, as discussed in future sections. Our per-county breakdown of data also enables regional analysis. Furthermore, the data on which the random forest model is trained is sourced from reliable and unbiased organizations—specifically, the USDA Risk Management Agency and the National Oceanic and Atmospheric Administration's Geophysical Fluid Dynamics Laboratory.

However, the model is somewhat limited in its scope of analysis. As per assumption 5, we only analyze insured farmland, since the crop loss data from the Risk Management Agency is only available for crops with insurance policies taken out on them. Historically, crop insurance participation in California has been relatively low. For example, although raisin grapes have a Federal Crop Insurance Program participation rate of 80 percent, almonds have a participation rate of only 34 percent [20]. In addition, the nine climate factors analyzed in our model were hand-selected from the CM2.1 database based on prior research to likely have impact on crop growth. Narrowing our pool of factors was necessary due to the large size of the dataset and the high computational intensity of both the data preprocessing and the random forest model. Given more time and computing power, a greater number of factors recorded in the database could be investigated. While our current model had a program runtime of 30 minutes, 6 to 7 hours would be required to account for every climate variable in the CM2.1 database.

Overall, our model is reproducible, flexible, and can easily be applied to other crops and US states as long as appropriate data is available. As aforementioned, it can also be expanded to incorporate a larger set of climate factors for a more comprehensive analysis.

# 4 Analysis and Conclusions

In this section, we utilize our model predictions to determine which organizations will be negatively affected by the anticipated trends and quantify the risk. We then provide recommendations on how these identified risks can be countered or combatted, including insurance and public policy changes.

## 4.1 Risk Analysis

### 4.1.1 Regional Risk Analysis

Although our climate data was recorded by county, we primarily analyzed the model results in terms of average trends across the entire state. In this section, we performed a regional risk analysis to identify the counties with the highest risk for crop loss as a result of climate change in the coming decades.

Due to disparities in growing conditions between different regions of California, we aimed to quantify risk for each county in relation to its own current status quo. Thus, we defined risk of future crop loss as the percent change from the mean loss in years 2000–2030 to the mean loss in years 2070–2100. A greater percent change would indicate a higher risk, since over the next several decades that county would lose a higher percentage of their usual profits compared to counties with lower risk. We computed these percent changes and colored a map of the counties of California accordingly, where darker colors represent more positive percent changes in loss (and therefore higher risk). This is shown in Figure 4.1.1; see Appendix Table 6.2.1 for all values used in our calculation.

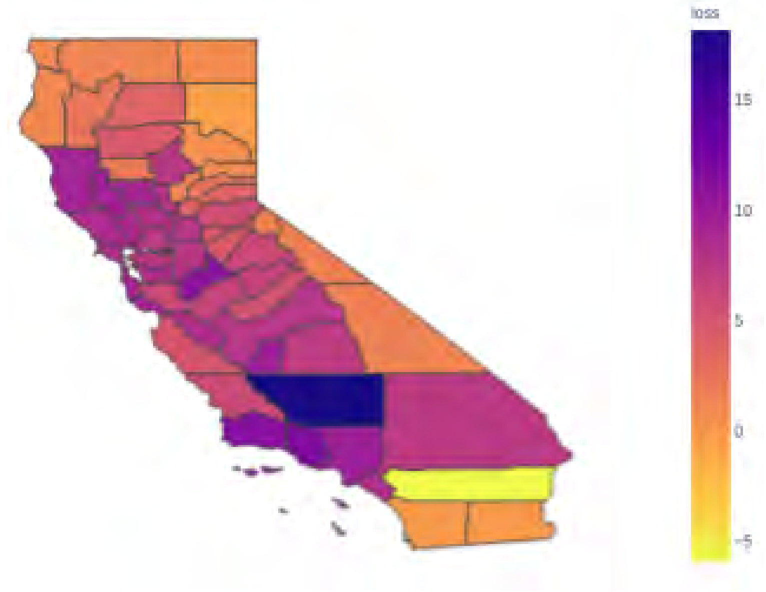


**Figure 4.1.1**: Counties of California Colored by Risk of Crop Loss (% Change in Loss)

From these results, we concluded that most counties in California are expected to witness increases in crop loss of around 5 to 15% by the end of the century. Notably, both the northernmost and southernmost counties appear to maintain relatively constant proportions of crop loss over time. The top 5 high-risk counties are Kern, Santa Barbara, Ventura, Stanislaus, and Los Angeles County; their percent changes are listed in Table 4.1.1. We then multiplied the percent change in crop loss by each high-risk county's annual crop production value [21], obtaining values for annual loss of agricultural value in 2070–2100 compared to 2000–2030. Our results are also reported in the table below.

**Table 4.1.1**: Crop and Economic Loss Projections of High-Risk Counties

| **High-Risk County** | **% Change in Crop Loss** | **Total Annual Production Value ($)** | **Potential Annual Loss ($)** |
|---|---|---|---|
| Kern | 18.14% | 7,187,938 | 1,303,892 |
| Ventura | 11.84% | 2,110,187 | 249,846 |
| Santa Barbara | 11.30% | 1,426,662 | 161,213 |
| Stanislaus | 10.76% | 3,261,412 | 350,928 |
| Los Angeles | 9.79% | 192,519 | 18,848 |

As expected, Kern County is predicted to have the highest potential annual loss because it has both the highest total annual production value as well as the greatest risk for crop loss. Stanislaus, Ventura, and Santa Barbara Counties are also at risk for significant financial loss.

### 4.1.2 Direct Risks to the Agricultural Industry

From our quantitative analysis, several risk groups or organizations are identifiable. The primary persons at risk are farmers, whose livelihoods are directly impacted by crop failure and resulting revenue loss. Logically, farmers in areas most prone to crop losses—namely, the counties listed in Table 4.1.1—are at the most risk overall. However, the socioeconomic status of farmers plays a large role in whether, and how well, they can adapt to changes. Crop insurance, managed by the USDA RMA, is available for many common crops and subsidized at a variable rate, ranging from 38 to 80 percent of the premiums [22]. Despite this, farmers of low socioeconomic status are still often unable to afford insurance. Without insurance, these disadvantaged farmers are always susceptible to revenue loss from crop damage and thus do not have a safety net to provide them peace of mind, unlike their wealthier counterparts. Over time, this situation may evolve into a cycle in which poorer farmers fall further and further behind middle-class and wealthy agricultural producers due to lack of resources and greater vulnerability to catastrophic financial blows.

Currently, the Federal Crop Insurance Program is a public-private partnership in which the USDA RMA, as a government agency, authorizes 15 private insurance companies to write MPCI (multiple peril crop insurance) policies: AFBIS, ARMtech, Country Financial, Crop Risk Services, Diversified Crop Insurance services, Farmers Mutual Hail, Global Ag, Great American, Hudson, NAU Country, ProAg, Precision Risk Management, Rain and Hail and RCIS [23]. As a result of our projected increase in crop losses over time, we can expect the financial assets of these private insurers to take a hit as a greater number of insurance claims will most likely be filed, and more indemnities paid out overall.

Specifically in our state of interest, Global Ag faces a unique risk as it is the first crop insurance provider based in California and the only company in the Federal MPCI program focused primarily on specialty crops. Specialty crops are defined legally as “fruits and vegetables, tree nuts, dried fruits and horticulture and nursery crops, including floriculture,” which California’s farms produce a large proportion of the nation’s supply of [24]. However, since the welfare of approved private insurers is intertwined with RMA regulation, which controls the rates that can be charged to farmers, it remains relatively unlikely that these companies will suffer irrecoverable damage. Not only does the government subsidize farmer-paid premiums, it also reimburses the private insurance companies to offset their operating and administrative costs [23]. Overall, individual farmers are still the group that faces the greatest risk as a result of our model’s projected agricultural changes.

#### 4.1.3 Ancillary Risk and At-Risk Subgroups

We centered our analysis on grapes, almonds, and avocados, for all of which California is one of the country’s major producers. Consequently, there is significantly greater risk to the industries that rely on the production of those crops than industries linked with more universally-grown commodities, such as wheat and corn. For example, companies responsible for packaging, transporting, and distributing almonds, which California produces 82% of the world’s supply of, would be significantly affected by changes in output across the state. In addition, if a county has a primarily agrarian local economy, its residents are also subject to greater risk. Of the high-risk counties identified in Section 4.1.1, Kern was the state’s #1 agricultural county based on 2017 gross production value, while Stanislaus and Ventura also fell within California’s top ten [25]. Since its economy is predominantly agrarian, and the region is also anticipated to have the greatest increase in crop loss over the next few decades according to our model, we consider Kern to be an especially high-risk county.

A specific subset of farmers who face particular risk from climate change and dwindling water access is historically underserved farmers. The Agriculture Improvement Act of 2018 (2018 US Farm Bill) defines and includes provisions to accommodate the concerns of historically underserved producers, which include socially disadvantaged, beginning, limited resource, and veteran farmers and ranchers [26]. Resource inequity is important to address everywhere, but particularly so in California, which has a relatively high Hispanic population: the state ranks third in the US in concentration of Latinx farmers, and 12% of all California farms are operated by Latinx farmers, compared to the national average of 3 percent. On average, Latinx and other farmers of color are of lower socioeconomic status and receive 36% less government funding than their white counterparts. Furthermore, the Sustainable Economies Law Center stated that “while 35% of non-Hispanic farmers acquire crop insurance… only 10% of Hispanics are enrolled in USDA insurance programs, and just 1% in Monterey County” [27]. Given that crop insurance is one of the most reliable and well-established means of risk management in agriculture, populations with lesser or inferior access to such programs are more susceptible to financial devastation. Although Hispanics are the largest minority group of Californian producers, Asian and Native American/Alaskan producers also comprise non-negligible portions of the industry [28].

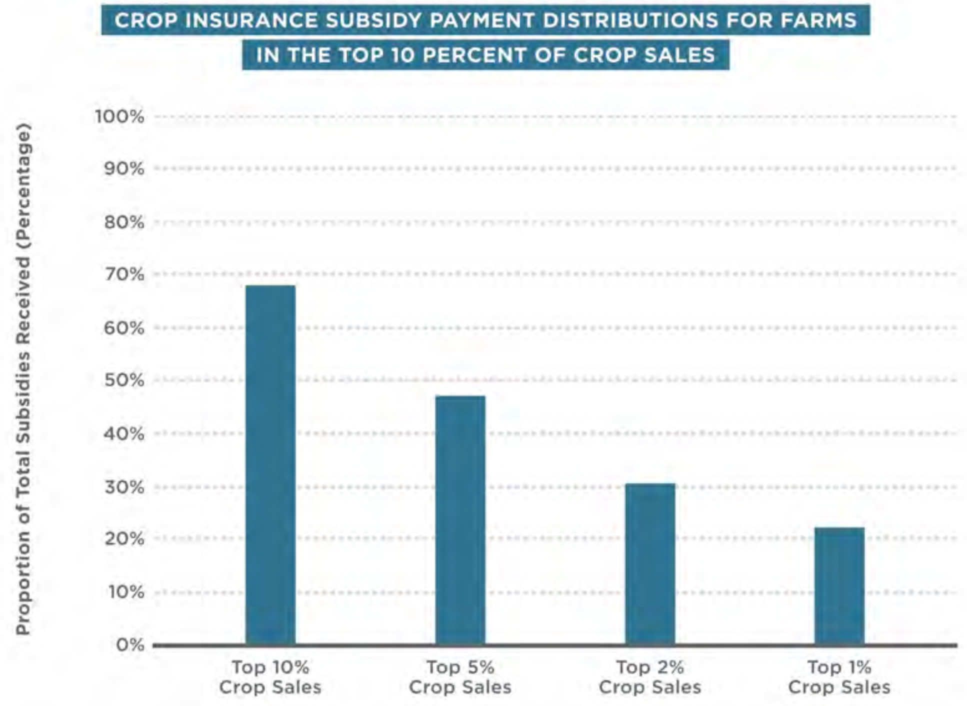
## 4.2 Recommendations

### 4.2.1 Insurance Recommendations

Formed in 1938 as part of President Franklin D. Roosevelt's New Deal as a result of the devastating impact of the Dust Bowl on American farmers, the Federal Crop Insurance Corporation is a government program that provides U.S. farmers and agricultural entities with crop insurance protection. Today, the FCIC, as a part of USDA's Risk Management Agency, oversees the implementation of crop insurance and risk management provisions in the US Farm Bill [29]. The Farm Bill is the federal government's primary agricultural and food policy tool; its most recent iteration is the 2018 Farm Bill, which was signed into effect in December of that year. Regarding insurance, the legislation was intended to increase the availability, affordability, and integrity of crop insurance programs for US farmers. The results of our mathematical modeling and risk analysis have led us to outline the following recommendations for crop insurance programs.

The Federal Crop Insurance Program, through government-subsidized Multiple Peril Crop Insurance policies written by authorized private companies, provides coverage for hundreds of crops, livestock, organics, dairy, and other agricultural products. Under the 2018 Farm Bill, the USDA Farm Service Agency provides a supplementary safety net that farmers and ranchers can use in addition to FCIP coverage through the Agricultural Risk Coverage (ARC) and Price Loss Coverage (PLC) programs. These both protect farm revenue from changes in market conditions. Notably, none of three crops analyzed in our model are covered by the programs. Therefore, we propose the Farm Service Agency to extend its coverage for ARC and PLC to common vegetables, fruits, and tree nuts, which are currently excluded from the programs' benefits. Given California's status as the nation's largest producer of vegetables, this insurance policy change would be a noteworthy first step toward mitigating the struggles of fruit and vegetable farmers in the state, whose crops face comparable risk from climate change yet are not offered the same insurance options as other commodities. ARC and PLC benefits are provided by farm, which is more in the interest of individual farmers than by-county benefits. Furthermore, since the largest proportion of spending projected by the 2018 Agricultural Improvement Act were for nutrition programs, we believe this extension of popular safety net programs is in line with the priorities of the federal government.

One of the major issues with the current state of the Federal Crop Insurance Program is the disparity between the lowest and highest subsidies received by farmers. Unlike subsidies for commodities, which cannot be paid to farm couples with over $1.8 million in gross income, there is no income restriction on eligibility for crop insurance subsidies. Consequently, large and already successful farms take up a vast majority of the total insurance premium support fund (Figure 4.2.1)—the top-selling 10 percent of farmers receive nearly 70 percent of all subsidies [30]. To address the inequitable distribution of subsidies, we propose that the Durbin-Grassley amendment be passed in the next Farm Bill, which would reduce subsidies for farmers with an Adjusted Gross Income (AGI) over $700,000. Though this amendment has passed the Senate twice, it ultimately never became law. However, this policy change not only would affect less than one percent of farmers (those with sufficient AGI) but is also estimated by the Congressional Budget Office to save more than $490 million over ten years [31]. Given that both the insurance and public policy recommendations outlined in our report are dependent upon increased funding, reducing federal costs with this amendment is a beneficial course of action.



Source: Bekkerman, A, Belasco, E., Smith, V. Where the Money Goes: The Distribution of Crop Insurance and Other Farm Subsidy Payments. American Enterprise Institute. 2018

**Figure 4.2.1**: Crop Insurance Subsidy Payment Distributions for Farms in the Top 10 Percent of Crop Sales [30].

Another significant area of concern is addressing historically underserved agricultural producers. Section 4.1.3 notes that Latinx farmers in particular—but also Asian and Native American/Alaskan farmers—tend to be of lower socioeconomic status on average while also receiving 36% less government funding than white farmers. In addition, poorer farmers may either be unable to afford crop insurance premiums (even after subsidies) or would experience significant financial setback if they paid it. Since insurers are legally required to write MPCI policies for any eligible applicant, regardless of race, we propose first focusing on improving affordability of insurance. To this end, one possible measure would be for the RMA to create a more progressive pricing model for subsidizing premiums, in which farmers with a lower gross income would be given subsidies closer to 80% of the premium price (the upper end of the current range of subsidy rates) while farmers with higher gross income would be given subsidies gradually closer to 40% (the lower end of the range). Future studies would have to be conducted after the implementation of these changes to assess their efficacy in making crop insurance more accessible across the board. Then, if data suggests that minority groups are still receiving significantly less funding, further steps should be undertaken to investigate potential racial bias among crop insurers.

### 4.2.2 Public Policy Recommendations

Addressing the agricultural risks of climate change at their roots requires strengthening and expanding US environmental policy. As shown in Figure 4.2.2, total water withdrawals in the US are the lowest they have ever been since the 1970s, despite a national population that has increased by over 50 percent during that time [32]. This statistic is a strong indication that water conservation efforts have resulted in tangible success, and attests to the capacity of legislation to effect change. Though changes to most of the policies below are outside the scope of the USDA RMA, we nevertheless recommend that the federal government prioritize climate preservation in future iterations of the Farm Bill and other environmental legislation.

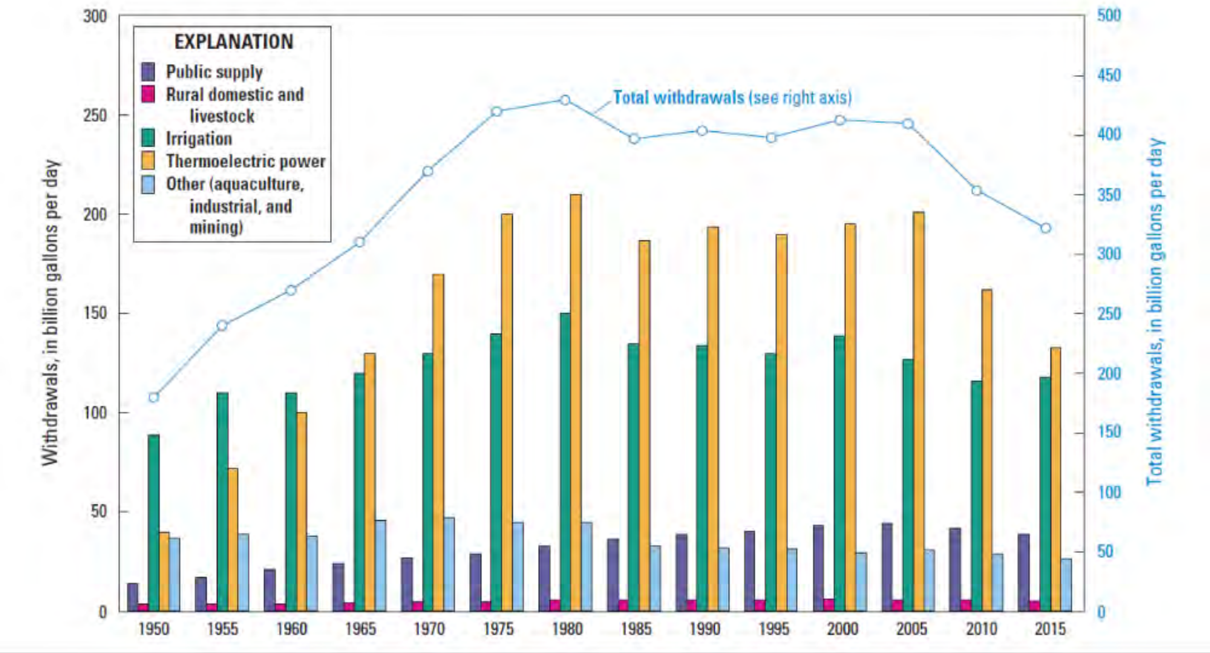


**Figure 4.2.2**: Trends in Total US Water Withdrawals by Category, 1950–2015 [33].

Specifically, we recommend the continuation of the Environmental Quality Incentives Program (EQIP), which was created by the 1996 Farm Bill to provide cost-sharing assistance for farmers to adopt more environmentally-friendly agricultural techniques. These techniques are expected to be more sustainable in the future but require greater initial investments, revenue losses, and lower crop yields. The 2018 Farm Bill already expanded eligibility, allowing the National Resources Conservation Service to enter into EQIP contracts with water management entities to support water conservation and irrigation efficiency [33]. Since our model has indicated that drought (or consecutive days without precipitation) is one of the most influential factors in determining crop loss, we suggest that the NRCS continue their water conservation efforts.

In addition, current EQIP policy dictates that in 2020, states may identify and increase payment rates for up to ten high-priority practices; eligible high-priority practices include water conservation to mitigate drought and address declining aquifers, habitat restoration, and natural resource concerns [33]. As we predict agriculture in California to be threatened by heat and drought, we suggest that its State Conservationists indicate water conservation and drought mitigation, as well as decreasing fossil fuel dependence, as high-priority practices beginning in 2020. The latter will provide greater incentive—via financial assistance—for farmers to invest in renewable energy in order to reduce carbon emissions. We highlighted these priorities based on the results of our principal component analysis in Section 3.3.2: water conservation and drought mitigation efforts corresponds to PC2 (precipitation), while decreasing fossil fuel dependence/curbing global warming corresponds to PC1 and PC3 (temperature; heat waves).

The Natural Resources Conservation Service, an agency of the USDA, offers both EQIP as well as the Agricultural Management Assistance (AMA) to help producers manage financial risk through diversification, marketing, and natural resource conservation practices. AMA provides up to 75 percent financial assistance of the cost of installing conservation practices, such as water management and irrigation structures, soil erosion control, and integrated pest management. It is currently available in sixteen states where participation in the Federal Crop Insurance Program has been historically low, but notably, California is not one of them. For the crops studied in our model, wine grapes and table grapes were 35 percent and 19 percent insured respectively by acreage, almond farmland was 42 percent insured, and avocados were 8 percent insured in 2012 [34]. (This final number should be viewed with discretion, as crop insurance for avocados was piloted fairly recently, in 1998 [35].) Given that 86 percent of eligible acres were insured under FCIP in 2015, California's participation in the program is still relatively low. Consequently, we recommend that NRCS and RMA make Agricultural Management Assistance available to California, as this would have the potential to significantly benefit local economies, such as in major agriculture counties Kern and Stanislaus, as well as the livelihoods of farmers.

### 4.2.3 Concluding Remarks

Overall, our model and analysis conclude that climate change, especially the effects of droughts, heat waves, and changing average temperatures, will introduce risk to producers of grapes, almonds, and avocados in California in the upcoming decades. To address projected risks, both economic and social, we recommend that the federal government expand safety net programs such as Agricultural Risk Coverage and Price Loss Coverage to cover fruits and vegetables, as a majority of California's agricultural economy depend on these "specialty crops." We also suggest incorporating the Durbin-Grassley amendment in the next Farm Bill to reduce insurance subsidies for farmers with gross annual income over $700,000, and implementing a more progressive premium subsidy system to address the unique risks faced by historically underserved and impoverished farmers. Ultimately, we recommend that the federal government prioritize climate preservation—for more than just the agricultural industry—by strengthening and expanding incentive programs for sustainable farming, resource conservation, and renewable energy use.

# 5 References

1. National Geographic Society. (2019). The California Drought. Retrieved from https://www.nationalgeographic.org/media/california-drought/.

2. Chou, B. (2016). Floods, Droughts and Agriculture. Retrieved from https://www.nrdc.org/experts/ben-chou/floods-droughts-and-agriculture.

3. California Department of Food and Agriculture. (2018) California Agriculture Statistics Review. Retrieved from https://www.cdfa.ca.gov/statistics/PDFs/2017-18AgReport.pdf.

4. Los Angeles Times. (2014, January 12). California farms lead the way in almond production. Retrieved from https://www.latimes.com/business/la-fi-california-almonds-20140112-story.html.

5. Foerster, J. (2019). Today's Extreme Winter Weather Can Impact Tomorrow's Crop Farming. Retrieved from https://www.forbes.com/sites/jimfoerster/2019/02/15/todays-extreme-winter-weather-can-impact-tomorrows-crop-farming/6cc7483fbcee.

6. National Integrated Drought System. (2019). California. Retrieved from https://www.drought.gov/drought/states/california.

7. Energy Upgrade California. (2019). Climate Change in California: Facts, Effects and Solutions. Retrieved from https://www.energyupgradeca.org/climate-change/.

8. United States Department of Agriculture. (2019). Irrigation  Water Use. Retrieved from https://www.ers.usda.gov/topics/farm-practices-management/irrigation-water-use/.

9. National Snow and Ice Data Center. (2020). Facts About Glaciers. Retrieved from https://nsidc.org/cryosphere/glaciers/quickfacts.html.

10. National Geographic Society. (2019, July 30). Water Table. Retrieved from https://www.nationalgeographic.org/encyclopedia/water-table/.

11. Insurance Information Institute. (2020). Understanding Crop Insurance. Retrieved from https://www.iii.org/article/understanding-crop-insurance.

12. Crop Insurance. (2020). California. Retrieved from https://cropinsuranceinamerica.org /california/.

13. USDA Southwest Climate Hub. (2018). AgRisk Viewer. Retrieved from https://swclimatehub.info/rma/rma-data-viewer.html.

14. Geophysical Fluid Dynamics Laboratory. (n.d.). GFDL's Contribution to the Coupled Model Intercomparison Project. Retrieved from https://www.gfdl.noaa.gov/cmip/.

15. Geophysical Fluid Dynamics Laboratory. (n.d.). Global Climate Models, CM2.5 and FLOR. Retrieved from https://www.gfdl.noaa.gov/cm2-5-and-flor/.

16. United States Department of Agriculture Risk Management Agency. (2020). Cause of Loss Historical Data Files. Retrieved from https://www.rma.usda.gov/Information-Tools/Summary-of-Business/Cause-of-Loss.

17. United States Department of Agriculture. (2020). Summary of Business Report Generator. Retrieved from https://prodwebnlb.rma.usda.gov/apps/SummaryofBusiness/ReportGenerator.

18. University of California Fruit and Nut Research and Information Center. (2012). California Horticultural Crops Statistics. Retrieved from http://fruitsandnuts.ucdavis.edu/files/147809.pdf.

19. Los Angeles Times. (2019). How Do Wildfires Start and Spread? Retrieved from https://www.latimes.com/california/story/2019-10-29/how-do-wildfires-start.

20. Lee, H. (n.d.). The Role Of Crop Insurance In California. Retrieved from https://s.giannini.ucop.edu/uploads/giannini_public/97/14/971478e8-2956-405e-99ff-653da9065886/v2n2_2.pdf.

21. California Department of Food and Agriculture. (2017). California Agricultural Statistics Review. Retrieved from https://www.nass.usda.gov/Statistics_by_State/California/Publications/Annual_Statistical_Reviews/2017/2016cas-all.pdf.

22. United States Department of Agriculture Economic Research Service. (2019). Agricultural Act of 2014: Highlights and Implications. Retrieved from https://www.ers.usda.gov/agricultural-act-of-2014-highlights-and-implications/crop-insurance.aspx

23. Crop Insurance. (2020). Insurance Providers. Retrieved from https://cropinsurance inamerica.org/about-crop-insurance/insurance-providers-list/.

24. United States Department of Agriculture. (n.d.). USDA Definition of Specialty Crop. Retrieved from https://www.ams.usda.gov/sites/default/files/media/USDASpecialty CropDefinition.pdf.

25. University of California Agriculture and Natural Resources. (2019). California's Top Counties for Agriculture. Retrieved from https://ucanr.edu/blogs/blogcore/postdetail.cfm.

26. United States Department of Agriculture Natural Resource Conservation Service. (n.d.). Historically Underserved Farmers  Ranchers. Retrieved from https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/people/outreach/slbfr/?cid=nrcsdev11_001040.

27. Thapar, N. (2017). CA Farmer Equity Act Signed Into Law. Retrieved from https://www.theselc.org/farmer_equity_act_pr.

28. German, B. (2019, October 8). The Changing Demographics of Farming. Retrieved from http://agnetwest.com/changing-demographics-farming/.

29. United States Department of Agriculture Risk Management Agency. (2018). Farm Bill. Retrieved from https://www.rma.usda.gov/Topics/Farm-Bill.

30. Environmental Working Group. (2018). Top 5 Reasons to Reform Crop Insurance. Retrieved from https://www.ewg.org/agmag/2018/06/top-5-reasons-reform-crop-insurance.

31. Taxpayers for Common Sense. (n.d.). Durbin Grassley Crop Insurance Subsidy Reform Amendment. Retrieved from https://www.taxpayer.net/wp-content/uploads/2018/06/Durbin_Grassley_Crop_Insurance_prem_reduction.pdf.

32. United States Geological Survey. (2015). Total Water Use in the United States. Retrieved from https://www.usgs.gov/special-topic/water-science-school/science/total-water-use-united-states.

33. Natural Resources Conservation Service. (2020). Environmental Quality Incentives Program. Retrieved from https://www.nrcs.usda.gov/wps/portal/nrcs/main/national/programs/financial/eqip/.

34. Lee, H. Sumner, D. A. (2013). Risk Management and the Farm Bill: The Role of Crop Insurance. Retrieved from https://s.giannini.ucop.edu/uploads/giannini_public/cc/74/cc742f46-72ef-449b-b703-2f1fa7fdc2f7/v16n6_4.pdf.

35. California Department of Food and Agriculture. (n.d.). New Insurance Program for Avocado Growers. Retrieved from https://www.cdfa.ca.gov/egov/Press_Releases/Press_Release.asp.

# 6 Acknowledgements

We're deeply grateful to the following people who gave us the resources and opportunities to make this wonderful experience possible.

- Mrs. Ellen LeBlanc - Our team coach
- Mr. Jeff Johnson - Our mentor for the Modeling the Future Challenge
- All the members and supporters of the Actuarial Foundation

# 7 Appendix

## 7.1 Changelog: 17 April 2020

- [Updated] Section 2, Background Information, page 3
  - Included more information on agricultural industry in California
- [Added] Section 3.3.3, Exploratory Factor Analysis of Climate Model, page 12
  - Inserted subsection header to provide additional organization
  - Provided greater justification for conducting EFA
  - Included details on communality calculation
- [Updated] Section 3.3, Results and Model Analysis
  - Figures 3.3.2, 3.3.3, 3.3.4, and 3.3.6 (pages 9, 10, 12) - Made scatter dots larger to improve readability
  - [Added] Table 3.3.2 to highlight unusual values in crop loss results
- [Added] Section 3.3.4, Summary of Model and Analysis Results, page 14
  - Added summary of entire section 3.3 for increased structure and clarity
- [Updated] Section 4.1.1, Regional Risk Analysis, page 15
  - Included reference to new Appendix Table 6.2.1
- [Added] Section 7.1, Changelog: 17 April 2020 (Appendix)
- [Added] Section 7.2, Supplementary Information (Appendix)
  - New Table 7.2.1: Historical and Future Loss by County

## 7.2 Supplementary Information

**Table 7.2.1**: Historical and Future Loss by County

| | County | Average Loss 2000-2030 | Average Loss 2070-2100 | % Change |
|---|---|---|---|---|
| 0 | Alameda | 0.098877563 | 0.10720841 | 8.425417664 |
| 1 | Alpine | 0.167253181 | 0.169781535 | 1.511692761 |
| 2 | Amador | 0.150239815 | 0.157862905 | 5.073947918 |
| 3 | Butte | 0.153000488 | 0.164588267 | 7.573688026 |
| 4 | Calaveras | 0.147993304 | 0.154944966 | 4.697281397 |
| 5 | Colusa | 0.105449088 | 0.115098315 | 9.150602284 |
| 6 | Contra Costa | 0.098854814 | 0.106408244 | 7.640932574 |
| 7 | Del Norte | 0.127613287 | 0.128905391 | 1.012515003 |
| 8 | El Dorado | 0.156940792 | 0.165805702 | 5.648569926 |

| | | | | |
|---|---|---|---|---|
| 9 | Fresno | 0.103985681 | 0.112017228 | 7.723703817 |
| 10 | Glenn | 0.144995747 | 0.148278968 | 2.264357228 |
| 11 | Humboldt | 0.11544957 | 0.116669851 | 1.056981966 |
| 12 | Imperial | 0.19666483 | 0.196190139 | -0.241370872 |
| 13 | Inyo | 0.18065855 | 0.183389086 | 1.511434639 |
| 14 | Kern | 0.117391808 | 0.138684831 | 18.13842379 |
| 15 | Kings | 0.104469001 | 0.114730979 | 9.82298915 |
| 16 | Lake | 0.100666113 | 0.1103534 | 9.623186102 |
| 17 | Lassen | 0.184747711 | 0.183882813 | -0.468150618 |
| 18 | Los Angeles | 0.14809017 | 0.162591872 | 9.792481033 |
| 19 | Madera | 0.109441821 | 0.115288228 | 5.342022555 |
| 20 | Marin | 0.09892704 | 0.107307386 | 8.471239146 |
| 21 | Mariposa | 0.152453605 | 0.162435538 | 6.547521648 |
| 22 | Mendocino | 0.102505115 | 0.112041771 | 9.303590739 |
| 23 | Merced | 0.100523911 | 0.107085138 | 6.527031142 |
| 24 | Modoc | 0.17977572 | 0.181200922 | 0.792766581 |
| 25 | Mono | 0.175491006 | 0.17841301 | 1.665044979 |
| 26 | Monterey | 0.1221739 | 0.128445012 | 5.132938518 |
| 27 | Napa | 0.099371368 | 0.107881779 | 8.564248605 |
| 28 | Nevada | 0.162439103 | 0.166920184 | 2.758622166 |
| 29 | Orange | 0.158191832 | 0.171427453 | 8.366816863 |
| 30 | Placer | 0.160100401 | 0.166474318 | 3.981200095 |
| 31 | Plumas | 0.177774705 | 0.179054712 | 0.72001639 |
| 32 | Riverside | 0.206120826 | 0.193913258 | -5.922530473 |
| 33 | Sacramento | 0.101557517 | 0.109518027 | 7.838424608 |
| 34 | San Benito | 0.102196924 | 0.110859162 | 8.476026156 |
| 35 | San Bernardino | 0.153675203 | 0.165116045 | 7.444819842 |
| 36 | San Diego | 0.19188971 | 0.191639307 | -0.130493286 |
| 37 | San Francisco | 0.098882434 | 0.107282742 | 8.495247822 |
| 38 | San Joaquin | 0.098921489 | 0.106486943 | 7.647937193 |
| 39 | San Luis Obispo | 0.147806229 | 0.156779715 | 6.071114927 |
| 40 | San Mateo | 0.098790831 | 0.108007434 | 9.329411137 |
| 41 | Santa Barbara | 0.14440268 | 0.160723668 | 11.30241334 |
| 42 | Santa Clara | 0.108387461 | 0.114919731 | 6.026776497 |
| 43 | Santa Cruz | 0.109484875 | 0.119488836 | 9.137299405 |
| 44 | Shasta | 0.155205873 | 0.161306261 | 3.930513586 |
| 45 | Sierra | 0.178115081 | 0.179642741 | 0.857681115 |
| 46 | Siskiyou | 0.149748797 | 0.151777177 | 1.354522021 |
| 47 | Solano | 0.099571107 | 0.106934878 | 7.395489716 |
| 48 | Sonoma | 0.099133475 | 0.10768692 | 8.628210557 |
| 49 | Stanislaus | 0.101048663 | 0.111918014 | 10.75655099 |
| 50 | Sutter | 0.110089866 | 0.118797959 | 7.909986133 |

| | | | | |
|---|---|---|---|---|
| 51 | Tehama | 0.147307111 | 0.154266468 | 4.724386426 |
| 52 | Trinity | 0.142351201 | 0.145835655 | 2.447786619 |
| 53 | Tulare | 0.110295786 | 0.118323079 | 7.277968875 |
| 54 | Tuolumne | 0.154136929 | 0.163736402 | 6.227886604 |
| 55 | Ventura | 0.142462776 | 0.159335298 | 11.8434604 |
| 56 | Yolo | 0.103162173 | 0.111486334 | 8.069005174 |
| 57 | Yuba | 0.145635276 | 0.147742375 | 1.446832765 |

## 7.3 pca.py

```
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
import pickle
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from factor_analyzer import FactorAnalyzer

with open('yields2.pkl', 'rb') as f: # Our output file
    yields = pickle.load(f)

with open('hist_nc.pkl', 'rb') as f: # Historical climate data
    hist_nc = pickle.load(f)

with open('pred_nc.pkl', 'rb') as f: # Predicted climate data
    pred_nc = pickle.load(f)
#print(yields.loc[(yields['year'] == 1989) & (yields['county'] == 'Butte')])

nc = pd.concat([hist_nc, pred_nc], ignore_index=True) # Combine the two climate datas

# plt.scatter(np.array(nc['year']),
# np.array(nc['pr'])*8640*12)
# plt.show(block=True)

nc = nc.drop(['county', 'year'], axis=1) # Remove countyand year b/c they don't matter in PCA

scaler = StandardScaler()
scaler.fit(nc)
nc_scaled = scaler.transform(nc) # Scale everything with a mean of 0 and std dev of 1

print(nc_scaled)

fa = FactorAnalyzer(rotation='varimax', n_factors=3, # Analyze 3 factors and drop na values
                    use_smc=False, impute='drop', method='principal')
fa.fit(nc_scaled)

pca = PCA(3) # Do PCA with 3 components
pca.fit(nc_scaled)
print(pca.explained_variance_ratio_) # % variance for each PC
print(pca.explained_variance_)
print(list(zip(nc.columns, fa.get_communalities()))) # Get communalities correlated with names of metrics

print(list(zip(nc.columns, fa.loadings_))) # Get loadings correlated with names of metrics
```

## 7.4 readFuture.py

```
import xarray as xr
from geopy.geocoders import Nominatim
import pandas as pd
from tqdm import tqdm
from datetime import datetime
from functools import reduce
from haversine import haversine

geolocator = Nominatim( # Server to get lat/long of county
    user_agent="mtf-stuff", domain='localhost/nominatim', scheme='http')

counties = [ # List of CA counties
    "Alameda", "Alpine", "Amador", "Butte", "Calaveras", "Colusa", "Contra_Costa", "Del_Norte", "El_Dorado", "Fresno",
    "Glenn", "Humboldt", "Imperial", "Inyo", "Kern", "Kings", "Lake", "Lassen", "Los_Angeles", "Madera", "Marin", "Mariposa",
    "Mendocino", "Merced", "Modoc", "Mono", "Monterey", "Napa", "Nevada", "Orange", "Placer", "Plumas", "Riverside", "Sacramento",
    "San_Benito", "San_Bernardino", "San_Diego", "San_Francisco", "San_Joaquin", "San_Luis_Obispo", "San_Mateo", "Santa_Barbara", "
        ↪ Santa_Clara",
    ”Santa_Cruz”, ”Shasta”, ”Sierra”, ”Siskiyou”, ”Solano”, ”Sonoma”, ”Stanislaus”, ”Sutter”, ”Tehama”, ”Trinity”, ”Tulare”,
    ”Tuolumne”, ”Ventura”, ”Yolo”, ”Yuba”]

# Historical data files for each metric
datas = [”cdd_A4.1861−2000.nc”, ”pr_A0.1861−2000.nc”,
         ”hwdi_A4.1861−2000.nc”, ”prsn_A0.1861−2000.nc”, ”r10_A4.1861−2000.nc”, ”tas_A0.1861−2000.nc”, ”hur_A0.1861−2000.nc”, ”prw_A0
               ↪ .1861−2000.nc”, ”r95t_A4.1861−2000.nc”]

# Prediction data files for each metric
# datas = [”cdd_A4.2001−2100.nc”, ”pr_A0.2001−2100.nc”,
# ”hwdi_A4.2001−2100.nc”, ”prsn_A0.2001−2100.nc”, ”r10_A4.2001−2100.nc”, ”tas_A0.2001−2100.nc”, ”hur_A0.2001−2100.nc”, ”prw_A0
      ↪ .2001−2100.nc”, ”r95t_A4.2001−2100.nc”]

# The end dataframe
out_data = pd.DataFrame(columns=(’county’, ’year’, ”cdd”, ”hwdi”, ”prsn”, ”r10”, ”tas”,
                                  ”hur”, ”pr”, ”prw”, ”r95t”))

# List of dataframes for each metric
raw_dfs = []

lats = [] # The possible latitude and longitudes in the file
longs = []

for data in datas: # Go through each file
    ds = xr.open_dataset(data)
    df = ds.to_dataframe()

    measurement = data.split(’_’)[0]

    # Some things are in days, so convert them to floats
    if isinstance(list(df.iloc[[0]].to_dict()[measurement].values())[0], pd._libs.tslibs.timedeltas.Timedelta):
        df[measurement] = df[measurement].dt.days.astype(’float’)

    curr_out = pd.DataFrame(columns=(’county’, ’year’, measurement)) # Our output dataframe, per metric

    if lats == []: # We haven’t populated the possible lats and longs (they’re the same in each metric, so we only do it once)
        lats = list(set(df.lat_bnds.index.get_level_values(1).values))
        longs = list(set(df.lon_bnds.index.get_level_values(2).values))

    for county in tqdm(counties):
        location = geolocator.geocode(county+”␣County,␣California”) # Get the lat/long of each location

        # print(county, location)

        county_loc = (location.latitude, location.longitude +
                      360 if location.longitude < 0 else location.longitude) # Convert to positive lat/longs

        lat_loc = 0
        for ilat in range(len(lats)−1): # Search for where our latitude would fit in
            if lats[ilat] <= county_loc[0] <= lats[ilat+1]:
                lat_loc = ilat
                break

        lon_loc = 0# Search for where our longitude would fit in
        for ilon in range(len(longs)−1):
            if longs[ilon] <= county_loc[1] <= longs[ilon+1]:
                lon_loc = ilon
                break

        x1 = (lats[lat_loc], longs[lon_loc]) # The four corners that surround the center of the county
        x2 = (lats[lat_loc+1], longs[lon_loc])
        y1 = (lats[lat_loc], longs[lon_loc+1])
        y2 = (lats[lat_loc+1], longs[lon_loc+1])

        # find distance from county centroid to each of 4 grid corners
        d1 = haversine(county_loc, x1)
        d2 = haversine(county_loc, x2)
        d3 = haversine(county_loc, y1)
        d4 = haversine(county_loc, y2)

        pd.options.mode.chained_assignment = None # Suppress error messages

        sub_df1 = df.query( # Get the rows for each latitude and longitude
            f”lat␣==␣{x1[0]}␣and␣lon␣==␣{x1[1]}”)
        sub_df1.columns = [ # Append _1 or _2 etc to each column name so they don’t conflict when we merge
            str(col) + ’_1’ if col == measurement else str(col) for col in sub_df1.columns]
        sub_df2 = df.query(f”lat␣==␣{x2[0]}␣and␣lon␣==␣{x2[1]}”)
        sub_df2.columns = [
            str(col) + ’_2’ if col == measurement else str(col) for col in sub_df2.columns]
        sub_df3 = df.query(f”lat␣==␣{y1[0]}␣and␣lon␣==␣{y1[1]}”)
        sub_df3.columns = [
            str(col) + ’_3’ if col == measurement else str(col) for col in sub_df3.columns]
        sub_df4 = df.query(f”lat␣==␣{y2[0]}␣and␣lon␣==␣{y2[1]}”)
        sub_df4.columns = [
            str(col) + ’_4’ if col == measurement else str(col) for col in sub_df4.columns]

        e = 1.0/d1 + 1.0/d2 + 1.0/d3 + 1.0/d4 # The e value on the bottom

        # print(sub_df1.index.get_level_values(−1))
        # print(sub_df1.index)

        # sub_df1.columns = sub_df1.columns.map(’|’.join).str.strip(’|’)
        sub_df1['time__'] = sub_df1.index.get_level_values(-1) # The year that it occured
        sub_df2['time__'] = sub_df2.index.get_level_values(-1)
        sub_df3['time__'] = sub_df3.index.get_level_values(-1)
        sub_df4['time__'] = sub_df4.index.get_level_values(-1)

        # sub_df1 = sub_df1.drop('lon_bnds', axis=1)
        # sub_df1 = sub_df1.drop('lat_bnds', axis=1)

        sub_df = reduce(lambda left, right: pd.concat( # Concatenate based on time into one whole dataframe
            [left, right], axis=1), [sub_df1.set_index('time__'), sub_df2.set_index('time__'), sub_df3.set_index('time__'), sub_df4.set_index('time__')])

        sub_df[measurement + "_total"] = (sub_df[measurement+"_1"]/d1 + # Create a final measurement according to the formula
                                          sub_df[measurement+"_2"]/d2 +
                                          sub_df[measurement+"_3"]/d3 + sub_df[measurement+"_4"]/d4) / e

        for _, line in sub_df.iterrows():
            # print(line)
            o = {'county': county, 'year': line.name.year} # Get the year for each line in the dataframe
            o[measurement] = line[measurement + '_total'] # And just put in our total
            curr_out = curr_out.append(o, ignore_index=True) # Append to our cleaned dataframe

    curr_out.groupby(['county', 'year'], as_index=False).mean() # In case there are any duplicates, just get the mean
    raw_dfs.append(curr_out)

out_data = reduce(lambda left, right: pd.merge(left, right, on=['county', 'year'],
                                               how='inner'), raw_dfs) # Combine our measurement dataframes into one giant dataframe

#out_data = out_data.drop_duplicates(ignore_index=True)

out_data = out_data.groupby(['county', 'year'], as_index=False).mean() # We want unique rows, so get the mean if there are any duplicates

out_data.to_pickle('hist_nc.pkl') # Pickle it so we don't have to do it again

print(out_data.head())
```

## 7.5 runForestRun.py

```
import pandas as pd
import pickle
import numpy as np
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score, GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.tree import export_graphviz
import pydot
import matplotlib.pyplot as plt
import plotly.express as px
from urllib.request import urlopen
import json
import sys

with urlopen('https://raw.githubusercontent.com/codeforamerica/click_that_hood/master/public/data/california-counties.geojson') as response:
    counties_json = json.load(response) # Get a GeoJSON of all the counties in California (for mapping with plotly)

with open('yields2.pkl', 'rb') as f:
    yields = pickle.load(f)

with open('hist_nc.pkl', 'rb') as f:
    hist_nc = pickle.load(f)

with open('pred_nc.pkl', 'rb') as f:
    pred_nc = pickle.load(f)

#print(yields.loc[(yields['year'] == 1989) & (yields['county'] == 'Butte')])

nc = pd.concat([hist_nc, pred_nc], ignore_index=True) # Load our climate data

#nc_current = nc.loc[nc['year'] <= 2019]

# print(nc_current)

# print(yields.describe())
# print(hist_nc.columns)

# cc = 'San Diego'
# cro = 'Avocado'

# plt.scatter(yields.loc[(yields['county'] == cc) & (yields['crop'] == cro)]['year'], yields.loc[(yields['county'] == cc) & (yields['crop'] == cro)]['yield'])
# plt.show(block=True)

merged = pd.merge(nc, yields, on=['county', 'year'],
                  how='outer') # Combine the climate data and yield data together

merged.loc[merged['yield'] == 0] = np.nan
merged = merged.dropna().reset_index() # Drop all rows with 0 yield
merged = merged.drop(['county', 'index', 'year'], axis=1) # Drop our county, index, and year (not inputs)
merged['yield'] = merged['yield'].astype('float') # Convert yield to float

print(merged.loc[merged['crop'] == 'Almond']['yield'].describe())

#merged = merged.loc[merged['crop'] == 'Almond']
#merged = merged.drop('crop', axis=1)

# print(merged['r95t'].describe())

features = pd.get_dummies(merged) # One hot encode the type of crop

labels = np.array(features['yield']) # Our output labels
features = features.drop(['yield'], axis=1) # Our input features

feature_list = list(features.columns) # The list of our inputs
features = np.array(features) # RF takes in a numpy array

print(feature_list)

train_features, test_features, train_labels, test_labels = train_test_split(
    features, labels, test_size=0.25, random_state=42) # Split into training testing data

print(train_features[0], train_labels[0])
print(test_features[0], test_labels[0])

print('Training Features Shape:', train_features.shape) # The shapes of our data, to make sure its consistent
print('Training Labels Shape:', train_labels.shape)
print('Testing Features Shape:', test_features.shape)
print('Testing Labels Shape:', test_labels.shape)

rf = RandomForestRegressor(n_estimators=100, random_state=42, # Make our RF regressor
                           max_features=0.33, max_depth=4, oob_score=True) # Some limits to make it faster
rf.fit(train_features, train_labels)

dot_data = export_graphviz(rf.estimators_[0], feature_names=feature_list, rounded=True, proportion=False,
                           precision=2, filled=True, rotate=True, out_file='tree.dot') # Get a tree in dot format
(graph,) = pydot.graph_from_dot_file('tree.dot') # Convert dot format to png
graph.write_png('tree.png')

# Use the forest's predict method on the test data
predictions = rf.predict(test_features)
# Print out the mean absolute error (mae)
errors = abs(predictions - test_labels)
print('Mean Absolute Error:', round(np.mean(errors), 2))

print(list(zip(feature_list, rf.feature_importances_))) # the importances of each of our features (sums to 1)

plt.rcParams.update({'font.size': 32})

future_ncs = nc.loc[(nc['year'] > 2019)] # Predict for years past 2019
future_ncs['crop_Almond'] = 0
future_ncs['crop_Avocado'] = 0
future_ncs['crop_Grape'] = 1 # We want to predict grapes
future_ncs = future_ncs.groupby(['year'], as_index=False).mean() # Combine all counties
future_ncs2 = future_ncs.drop(['year'], axis=1) # Drop our years (not an input)
print(future_ncs2.head())

pred_features = np.array(future_ncs2)

out = rf.predict(pred_features) # Predict our crop loss

for yr in range(2020, 2120, 20): # Get key data for years 2020, 2040, 2060, 2080, 2100
    print("PREDICTED Year ", yr, out[yr-2020])
sys.exit(0)

plt.scatter(np.array(future_ncs['year']), out) # Plot it
plt.ylabel("Crop Loss (proportion)")
plt.xlabel("Year")
# plt.show(block=True)

f, (ax1, ax2, ax3) = plt.subplots(1, 3)

ax1.scatter(np.array(future_ncs['year']), np.array(future_ncs['tas'])) # Plot our climate data in the future
ax2.scatter(np.array(future_ncs['year']), np.array(future_ncs['hwdi']))
ax3.scatter(np.array(future_ncs['year']), np.array(future_ncs['cdd']))

ax1.set_xlabel("Year")
ax2.set_xlabel("Year")
ax3.set_xlabel("Year")

ax1.set_ylabel("Temperature (K)")
ax2.set_ylabel("Heat Wave Duration Index")
ax3.set_ylabel("Consecutive Dry Days (days)")

# plt.show(block=True)


def get_avg_map(change_df): # Get the average mapping

    counties = change_df['county'].unique() # List of counties
    change_df['crop_Almond'] = 1 # Do it for almonds
    change_df['crop_Avocado'] = 0
    change_df['crop_Grape'] = 0
    #one_year_predict = one_year_predict.groupby(['year'], as_index=False).mean()
    change_predict2 = change_df.drop(['year', 'county'], axis=1)

    # DO FOR EACH CROP
    plot_map_almond = rf.predict(np.array(change_predict2))
    change_predict2['crop_Almond'] = 0
    change_predict2['crop_Avocado'] = 1
    change_predict2['crop_Grape'] = 0
    plot_map_avocado = rf.predict(np.array(change_predict2))
    change_predict2['crop_Almond'] = 0
    change_predict2['crop_Avocado'] = 0
    change_predict2['crop_Grape'] = 1
    plot_map_grape = rf.predict(np.array(change_predict2))

    plot_map_all = (plot_map_almond + plot_map_avocado + plot_map_grape) / 3 # Average our crops

    change_df['yield'] = plot_map_all
    # print(change_predict.head())

    dropped_yield = change_df.loc[:, change_df.columns.intersection(
        ['yield', 'county'])] # Remove everything except yield and county name

    # grouped = dropped_yield.groupby(
    # ['county'])

    # grouped_outliers = grouped.transform(
    # lambda group: (group - group.mean()).div(group.std()))
    # outliers = grouped_outliers.abs() > 2 # Calculate outliers (2 std devs)

    # # print(outliers)

    # dropped_yield = dropped_yield[outliers.any(axis=1)].groupby([
    # 'county']).size()

    # o = dropped_yield.to_frame().reset_index().rename(columns={0: 'yield'})

    # print(counties)

    # print(o['county'])

    # for county in counties:
    # if county not in o['county'].unique():
    # print("Adding", county)
    # o = o.append({'county': county, 'yield': 0}, ignore_index=True)
    # return o

    dropped_yield = dropped_yield.groupby( # Average all data by county
        ['county'], as_index=False).mean()

    # dropped_yield = dropped_yield.groupby(
    # ['county']).std().reset_index() # Plot the standard deviation

    return dropped_yield


change_predict = nc.loc[(nc['year'] >= 2000) & (
    nc['year'] <= 2030)] # Our base years is 2000 to 2030

change_pred_future = nc.loc[(nc['year'] >= 2070) & (nc['year'] <= 2100)] # Our future years are 2070 to 2100

plot_hist = get_avg_map(change_predict)
plot_future = get_avg_map(change_pred_future)

plot_change = pd.DataFrame( # Calculate percent change, and remove County if its in the name
    {'yield': 100 * (np.array(plot_future['yield']) - np.array(plot_hist['yield'])) / np.array(plot_hist['yield']), 'county': [county.replace('_County', '') for county in plot_future['county']]})

# print(plot_map_df)

plot_change.columns = ['loss', 'county'] # Rename to loss

# print(counties_json["features"][10])

fig = px.choropleth(plot_change, geojson=counties_json, color="loss", # Plot our map
                    locations="county", featureidkey="properties.name",
                    color_continuous_scale='Plasma_r'
                    )
fig.update_geos(fitbounds="locations", visible=False) # Make it fit to california and ignore everything else
fig.update_layout(margin={"r": 0, "t": 0, "l": 0, "b": 0})
fig.show()
```

## 7.6 totalLand.py

```
import os
import pandas as pd
import numpy as np

fileLand = open(”totalLand.txt”, ”r”) # The text file with total land for each county
rawdata = fileLand.readlines()
headers = rawdata[0].split(”|”)
data = [x.split(”|”) for x in rawdata[1:]] # Its separated by pipes
land = [[[0 for i in range(3)] for j in range(58)] for k in range(32)] # Initialize a 3d array for each crop, county, year

# comodity name”: 1
# area: 13
# county: 7
counties = [
    ”Alameda”, ”Alpine”, ”Amador”, ”Butte”, ”Calaveras”, ”Colusa”, ”Contra␣Costa”, ”Del␣Norte”, ”El␣Dorado”, ”Fresno”,
    ”Glenn”, ”Humboldt”, ”Imperial”, ”Inyo”, ”Kern”, ”Kings”, ”Lake”, ”Lassen”, ”Los␣Angeles”, ”Madera”, ”Marin”, ”Mariposa”,
    ”Mendocino”, ”Merced”, ”Modoc”, ”Mono”, ”Monterey”, ”Napa”, ”Nevada”, ”Orange”, ”Placer”, ”Plumas”, ”Riverside”, ”Sacramento”,
    ”San␣Benito”, ”San␣Bernardino”, ”San␣Diego”, ”San␣Francisco”, ”San␣Joaquin”, ”San␣Luis␣Obispo”, ”San␣Mateo”, ”Santa␣Barbara”, ”
        ↪ Santa␣Clara”,
    ”Santa␣Cruz”, ”Shasta”, ”Sierra”, ”Siskiyou”, ”Solano”, ”Sonoma”, ”Stanislaus”, ”Sutter”, ”Tehama”, ”Trinity”, ”Tulare”,
    ”Tuolumne”, ”Ventura”, ”Yolo”, ”Yuba”]

crop = −1
count = 0
for line in data:
    count+=1
    if ”ALMONDS” in line[1].upper(): # Find the crops we’re interested in
        crop = 0
    elif ”GRAPES” in line[1].upper():
        crop = 1
    elif ”AVOCADOS” in line[1].upper():
        crop = 2
    if crop != −1:
        try:
            land[int(line[0])−1989][counties.index(line[7]) # Conevrt to float the amount in acres of land
                                        ][crop] += float(line[13])
        except:
            pass
    crop = −1

counter = 1989
for year in land:
    print(”\nyear:␣” + str(counter))
    counter += 1
    for i in range(3):
        sum = 0
        for line in year:
            sum += line[i] # Get the sum for each county+crop
        print(sum)
# year, county, crop
lostLand = [[[0 for i in range(3)] for j in range(58)] for l in range(31)] # Now get land lost
for root, dirs, files in os.walk(”./colsom”, topdown=False): # Go through each file
    for name in files:
        if name[−4:] == ”.txt”:
            fileIn = open(os.path.join(root, name)) # Open up the file for the year
            rawdata = fileIn.readlines()
            data = [x.split(”|”) for x in rawdata]
            for line in data:
                if line[1] == ”06”:
                    crop = −1
                    if ”ALMONDS” in line[6].upper():
                        crop = 0
                    elif ”GRAPES” in line[6].upper():
                        crop = 1
                    elif ”AVOCADOS” in line[6].upper():
                        crop = 2
                    if crop != −1:
                        lostLand[int(
                            line[0])−1989][counties.index(line[4].strip())][crop] += float(line[27]) # Add together the lost land
yields = [[[0 for i in range(3)] for j in range(58)] for l in range(31)] # Calculate the yields
for i in range(3):
    for j in range(58):
        for k in range(31):
            if land[k][j][i] != 0: # If we have no land
                yields[k][j][i] = 1.0∗lostLand[k][j][i] / land[k][j][i]
                if yields[k][j][i] > 1.0: # If we have a loss > 1(for some reason), set it to 1
                    yields[k][j][i] = 1.0

np.save(”./yeilds”, yields)


#yields = np.load(’yeilds.npy’)

# year = index/58
# county = index%58

crops = [’Almond’,’Grape’,’Avocado’]

df = pd.DataFrame({},columns=[’year’, ’county’, ’crop’, ’yield’]) # Make a dataframe

yield2D = [[0 for i in range(3)] for j in range(1798)]
for i in range(3):
    for j in range(1798): # Convert everything into the dataframe
```

```
        year = j//58
        county = j%58
        df = df.append({'year': year+1989, 'county': counties[county], 'crop': crops[i], 'yield': yields[year][county][i]}, ignore_index=True)
        #yield2D[j][i] = yields[year][county][i]

# df = pd.DataFrame(yield2D,columns=['Almond','Grape','Avocado'])



print(df)
df.to_pickle("./yields2.pkl") # Pickle the dataframe
```