Analysis and Prediction on Corn Yield in Illinois

Consultants: Shihao Duan, Smruthi Iyengar, Kagen Quiballo, Doris Wang, Lucy Zhao

Client: Mark Schleusener, National Agricultural Statistics Service &

Darren Glosemeyer, University of Illinois at Urbana-Champaign

University of Illinois at Urbana-Champaign

STAT 443

1. INTRODUCTION

1.1 Background & Motivation

Corn is a crop that is widely planted in the Midwest. In Illinois, there are nine agricultural districts, all of which plant a vast amount of corn. However, due to temperature and precipitation disparities, corn yields in different areas and different years may have huge differences. According to the requirements of the client, our goals are to tell if there is a difference between the corn yield in district 20 and district 60 over the past two decades, point out the reasons behind the difference, predict the corn yield by building statistical models, and figure out if innovative technology have an impact on the corn yield. We got the crop yield and progress data from Mark Schleusener from the National Agricultural Statistics Service (NASS). The crop yield data contains the information of the corn yields in different agricultural districts, so we believe this dataset can be used to see if there is a difference in corn yields between the districts 20 and 60. The weather and climate data were provided by the National Oceanic and Atmospheric Administration (NOAA), containing the detailed meteorological data in different weather stations in district 20 and 60 from 2000 to 2018, and we think these data are important because they provide important factors that may contribute to the difference of corn yields in different areas and years.

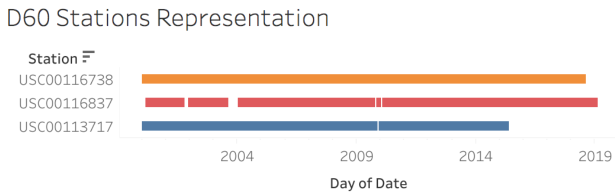
1.2 Data Preparation

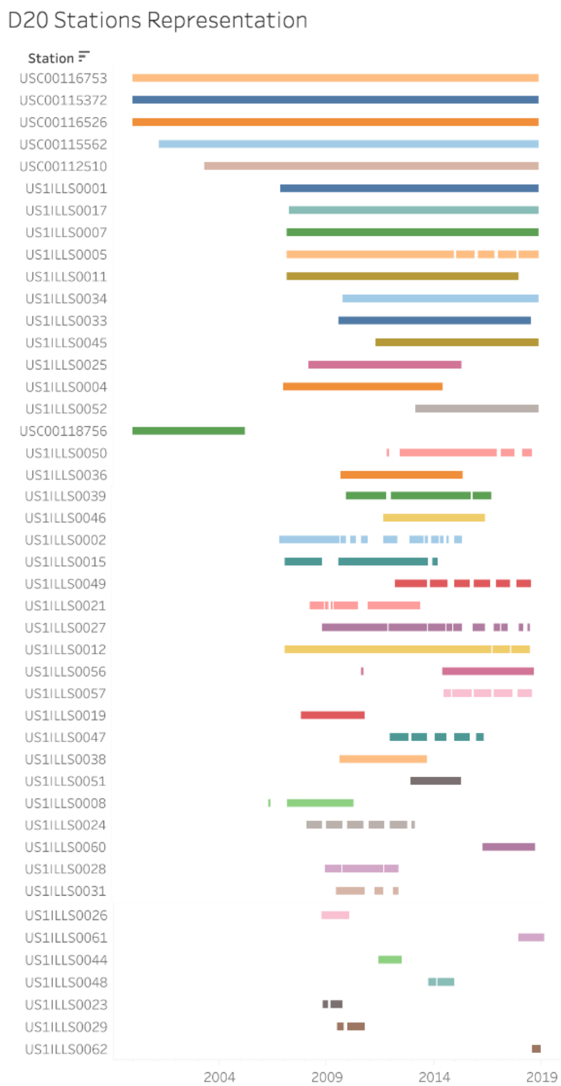
The four datasets given are *Illinois Corn Progress Data*, *Illinois Corn Yield Data*, *LaSalle County*, *District 20*, and *Pike County, District 60*. The *Illinois Corn Progress Data* includes 13 variables and 2868 observations, recording corn's growth status in 9 districts in Illinois weekly from April to November. The *Illinois Corn yield Data* has 11 variables and 190 observations. It records the corn grain yields per acre and acres harvested of 9 districts and the state yearly from 2000 to 2018. The *LaSalle County, District 20* data has weather information in district 20 including 7 variables and 95949 observations recorded by 45 weather stations. The *Pike County, District 60* data has weather information in district 60 including 7 variables and 18405 observations recorded by 3 weather stations. After reading these datasets into R, the primary analytical tool we are using for building predictive models, we performed data examination and cleaning. One of the issues that we noticed was that for the datasets *LaSalle County, District 20* and *Pike County, District 60*, there were lots of missing values that were useless for our analysis, so we subsetted these two datasets by removing the rows that contained missing values. For example, in *LaSalle County, District 20*, there were over 90k observations, after cleaning, only 19603 of them are found useful. The cleaned datasets are used for building regression models. The second problem existing in the datasets is that the variables in the datasets were recorded in different time scale. For example, the variables regarding precipitation and temperature were recorded daily in the datasets *LaSalle County, District 20* and *Pike County, District 60*, however, the data of the percentages of different corn growth stages were measured by week in *Illinois Corn Progress*, and the data of corn yield per acre and acres harvested in *Illinois Corn Yield Data* were measured by year in *Illinois Corn Yield Data*. How to utilize and reconcile, if necessary, these different measures when doing analysis is definitely what we should think about. We will address this issue later with the description of the specific analysis performed. Last but not least, to get better analysis, we added several variables that might be helpful in the variable analysis part.

2. VARIABLE ANALYSIS AND VISUALIZATION

2.1 Precipitation

When choosing to clean the precipitation data, we were faced with the issue of predicting annual yield with daily temperature data from multiple stations across districts. Converting the metrics of the two posed a problem as well as missing values on many dates, but the thought process of our methodology is the following. District 20 had 45 stations and District 60 had 3 stations that each recorded data from different days, some of them overlapping. The following is a graph representing which days each station recorded precipitation





The first method of creating a cumulative annual precipitation variable introduces the bias of over-representation from dates that have more data points. For example, if all the stations happen to record the day of a flood, the year would reflect precipitation that is too high, given that there are missing values from other dates that year from those given stations.

The next method of only choosing stations that have data points for all days from 2000-2018 would also be biased by not representing other areas in the district.

Therefore, the method we chose was to query the data by selecting daily averages across all stations in that district such that each day was equally represented despite missing values. From there, we take the sum of select days each year.

Looking at some simple statistics in relation to NASS standards, we see that dates with over 4 inches of precipitation may affect crop yields. This occurs once in 2013 for district 20 and twice in 2003 and 2014 from district 60. This may be taken note of when looking at the time vs. yield graph. Additionally according to NASS standards, approximately one inch of rain per week is good, and district 20 and 60 have an average weekly precipitation of 0.736 inches and 0.761 inches respectively. According to the Illinois State Water Survey, "Generally, annual rainfall exceeds the water requirement of Illinois crops. The average for southern Illinois is 45 inches, the rest of the state is 37 inches," and is reflected by the average annual rainfall of 38.45 inches and 39.77 inches of district 20 and district 60 respectively.

4 new precipitation variables were created as follows according to quotes from the Illinois State Water Survey. We used these quotes to create the variables because the time periods selected were indicated as critical time periods that may affect crop yield. Querying this data took a lot of time, and we were unable to integrate them into the predictive models. They have the potential to add more accuracy to the models and may be used in future directions. The 4 new queried variables are defined below and are available for viewing with this link: *tinyurl.com/STAT443grp6-prcp-data.*

"The critical time during the early growth lasts for approximately 30 days, from planting to tassel initiation."

Prec_Planted_30 is the sum of the daily averages across all stations in that district for **30 days after 50% is planted**

"Rainfall of 1 to 2 inches in the 2 weeks following corn pollination will generally result in the highest yield"

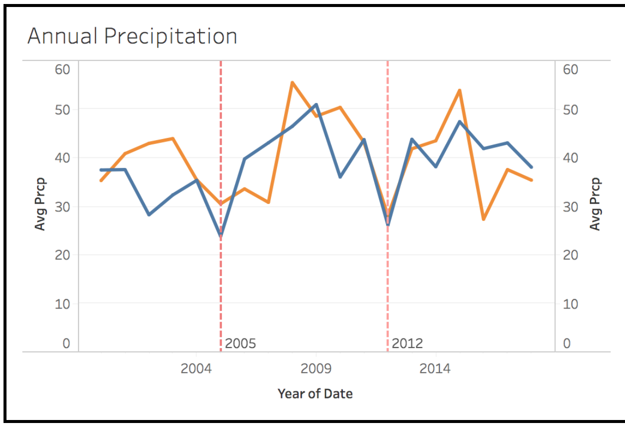
Prec_Pollinated_14 is the sum of the daily averages across all stations in that district for **2 weeks after 50% is silking**

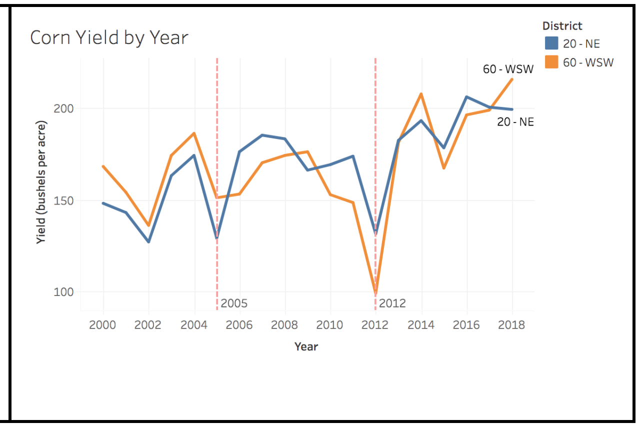
"During fallow season, there is usually enough precipitation to recharge the soil profile by January of the year. Otherwise February, March, and April are usually adequate to recharge the soil profile."

Prec_TOTAL is the sum of the daily averages across all stations in that district for **from the time 50% is planted to 50% is harvested**

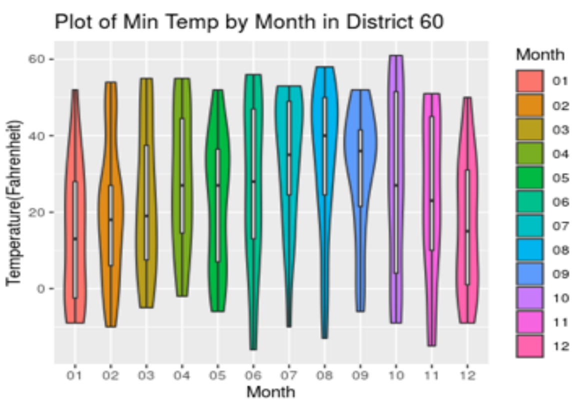
Prec_ANNUAL is the sum of the daily averages across all stations in that district for **from Jan 1 to Dec 31 of that year**

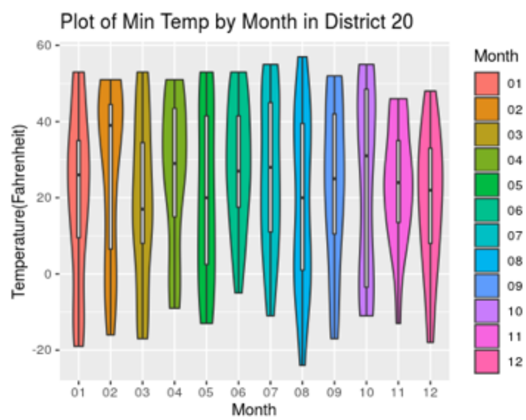
One last factor to consider is droughts. According to the Palmer Drought Severity Index on isws.illinois.edu, 2005 was considered drought and 2012 was considered extreme drought. The left graph indicates the years of these droughts have significantly lower annual precipitation than adjacent years. The right graph indicates that these droughts also impacted annual yield as shown by the significantly lower yields in 2005 and 2012.

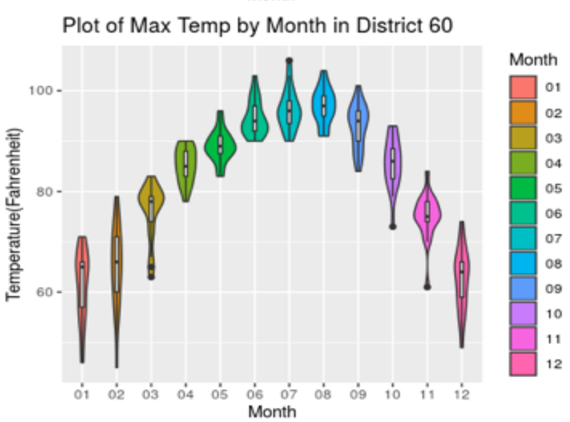


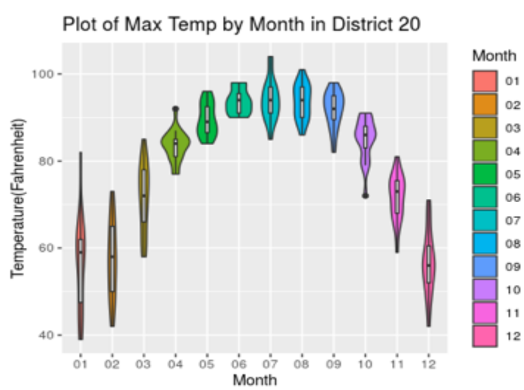


2.2 Temperature







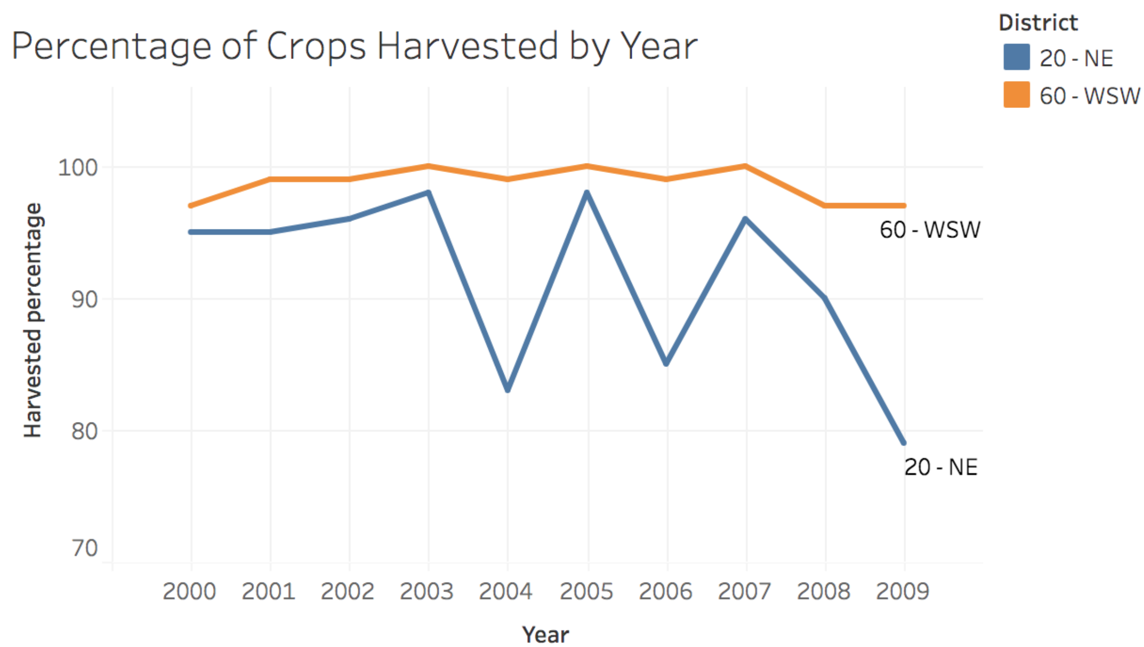


In order to examine the temperature variable, we want to see how temperature changes on average throughout the years. The dataset contains daily minimum and maximum temperatures of each district from the years 2000 to 2018. We created a monthly average of minimum temperature and maximum temperature throughout the years. This monthly average was made by picking the highest temperature value and the lowest temperature value of each month between January 2000 and December 2018 and creating a violin plot to determine if there are any major differences in temperature between the two districts. For both districts, the average temperatures were somewhat similar. But, District 20 had slightly lower averages than district 60 in both maximum temperature and minimum temperature. For both districts, the range in average the range of the minimum temperature is much more than the range of the maximum temperature. For average maximum temperature, the summer months have the smallest range in both districts.

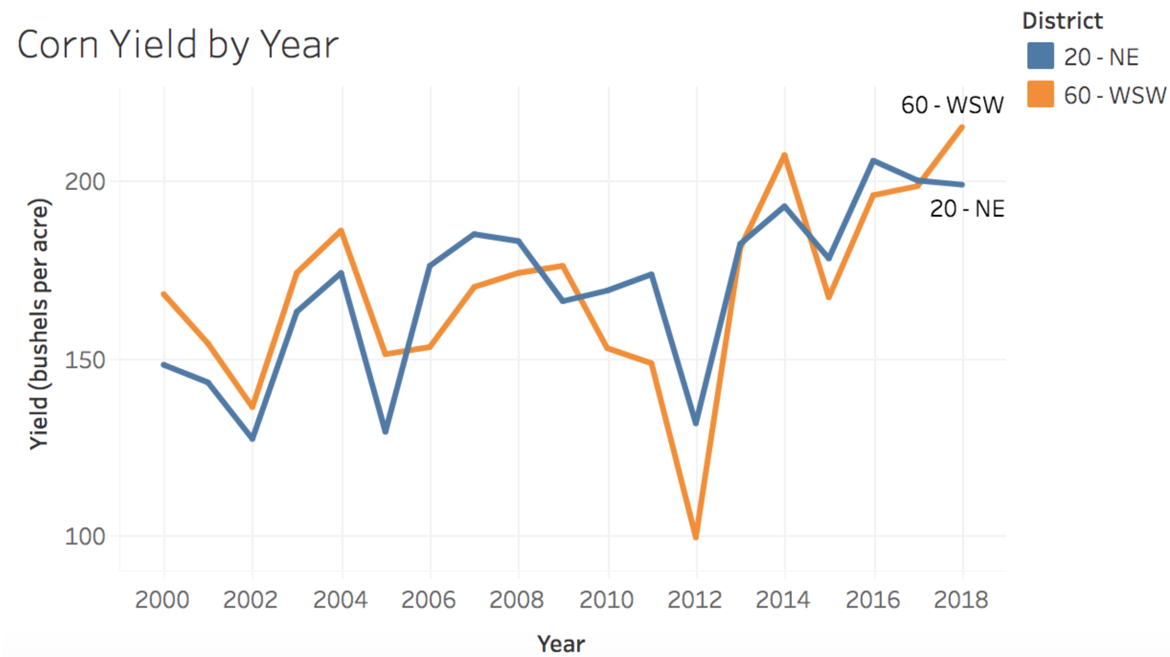
Although we can visualize that there are some differences in temperature between the two districts, we need to determine if the difference is statistically significant. The overall average maximum temperature in district 60 is 81.71°F and the average maximum temperature in district 20 is 78.72°F. After running a two-sample paired t-test for the maximum temperature variable we can determine that this difference in temperature is statistically different because the p-value is 0.03 and the difference has a 95 percent confidence interval of (-5.62, -0.37). The average minimum temperature in district 60 is 25.33°F and the average minimum temperature in district 20 is 23.24°F. The p-value from the paired two-sample t-test was .30 and the 95 percent confidence interval for the difference in minimum temperatures included 0 (-6.10, 1.91). So, the difference in minimum temperature between the two districts are not statistically significant.

3. DISTRICT VARIANCE ANALYSIS AND VISUALIZATION

As mentioned before, the datasets are given in different time scale, so it is difficult for us to compare the difference in yield using data with different time scales, as we notice that the dataset *Illinois Corn Yield Data* gives corn yield for all regions in Illinois, we made two subsets that contain the data of district 20 and district 60, respectively. We performed a paired t-test to compare the difference in means.



The result shows that when looking at the percentage of crops harvested after each season, there is a significant difference between the harvested percentage in district 20 and district 60. District 60 typically has a higher percentage harvested of about 98.7% compared to district 20's 91.5% average harvested. One may also note that data spans from 2000-2009. Our p-value for the test was 0.00573, so at a 0.05 significance level, we conclude that the average harvest percentage of district 20 is less than the average harvest percentage of district 60.

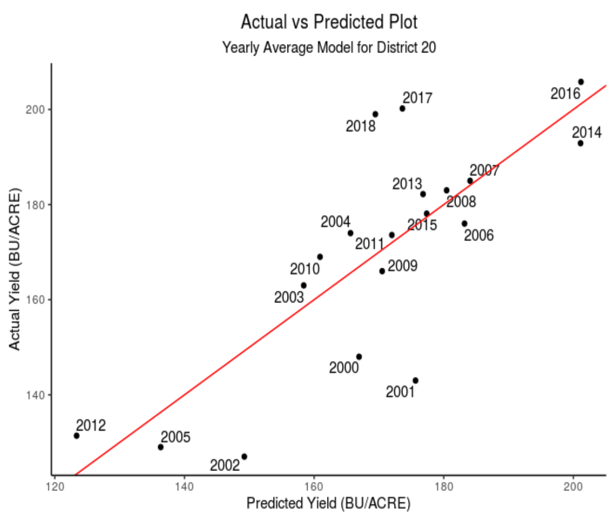


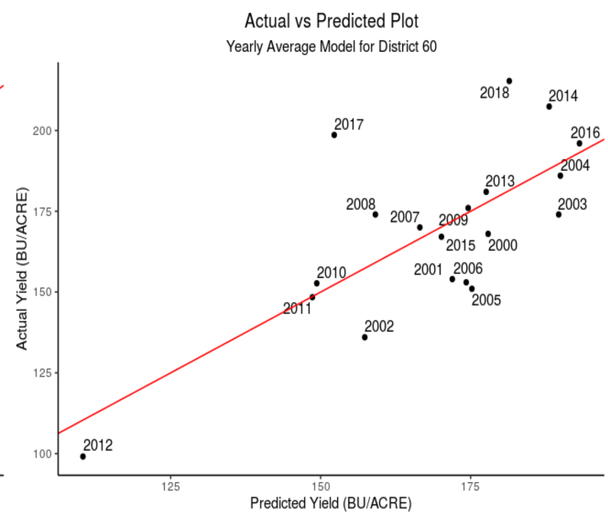
However, when running the same paired t-test to look at the <u>yield measured in BU/acres</u> for each year, there appears to be no significant difference between district 20 and district 60. They both have similar yields of about 169.8 BU/acre for district 20 and 168.82 BU/acre for district 60, which are relatively similar with large variations throughout each year ranging from 99.1 - 215.0. Our test resulted in a p-value of 0.799, so at a 0.05 significance level, we conclude that there is no significant difference between the average yield of district 20 and district 60 when paired by year.

4. PREDICTIVE MODEL ANALYSIS

4.1 Predictive Model - Yearly Averages

In order to predict future corn yield, we built linear regression models for each district using variables regarding temperature and precipitation from the datasets *LaSalle County, District 20* and *Pike County, District 60* as the predictors, and the corn yield variable measured in bu/acre from the dataset *Illinois Corn Yield Data* as the response variable. Since temperature and precipitation were measured daily while corn yield was measured yearly, we calculated the yearly averages of precipitation, minimum temperature, and maximum temperature variables to make the predictors in the same measurement as the response variable. The summary datasets we used for building the models are in Appendix 7.1.1, and they can also be viewed here. We started from fitting a full linear regression model with three yearly-averaged variables (precipitation, maximum temperature, and minimum temperature) and all their interaction terms, and then selected the variables in both directions based on the Akaike Information Criterion (AIC), which is an estimator of the relative quality of statistical models for a given set of data. The selection procedure showed that the best linear model for District 20 is *Yield (BU/ACRE)* = 11084.501 - 25371.63 * *Precipitation* - 182.966 * *Maximum Temperature* - 199.392 * *Minimum Temperature* + 427.621 * *Precipitation* * *Maximum Temperature* + 3.338 * *Maximum Temperature* * *Minimum Temperature*, and for District 60 is *Yield* = 933.316 - 820.936 * *Precipitation* - 24.251 * *Maximum Temperature* + 20.646 * *Minimum Temperature*. We proved that all the variables in both models are significant under 0.1 significant level (Appendix 7.1.2), so they have significant impacts on corn yield; the values of Multiple R-Squared, which measures how much variability of the respondent variable can be explained by the predictors, are 0.623 and 0.528, respectively, which means that more than half of the variability in corn yield can be predicted by precipitation, temperature, and their interaction effects. The values of Adjusted R-Squared, which adds a penalty for the number of coefficients to prevent using lots of predictors to boost the Multiple R-Squared values, for both models are 0.478 and 0.434, respectively. The actual vs. predicted corn yield graphs of the linear models of the two districts are presented as follows:





As we can see, the points are falling around the red line of the actual equal to the predicted, which means that most of our predictions are pretty close to the actual corn yields. However, there are also some year's corn yields that are obviously over- or under-estimated, such as the years 2001, 2002, 2017, and 2018 of District 20 as well as the years 2017 and 2018 of District 60. The Mean Absolute Errors (MAE), which are the means of the absolute differences between the actual and predicted values, of District 20 and 60 using our models are 10.639 and 13.565, respectively. These discrepancies in actual and predicted values may result from other unconsidered factors that may also affect corn yield, or from the fact that this model using yearly averages is too general to make accurate predictions. Therefore, we decided to move on to build a model based on averages by growth stages to see if the accuracy of predictions can be improved.

4.2 Predictive Model - Growth Stages

As discussed above, if a farmer possessing sufficiently detailed climate data wants to learn more about how each growing stage affects total corn yield of the year and makes more accurate predictions, he/she may want to use our predictive models for corn yield which was developed based on the averages by growth stage. Based on the beginning and ending dates of each growth stage of each year given in the dataset *Illinois Corn Progress Data* and the daily climate data from *LaSalle Country, District 20* and *Pike Country, District 60*, we calculated the average precipitation, maximum temperature, and minimum temperature by each growth stage. Since the climate daily data contains information from 2000 up to 2018 but *Illinois Corn Progress Data* only gives the dates of each stage up to 2009, so we assumed the time periods for the growth stages in the years 2010 to 2018:

Plant	Emerge	Silk	Dough	Dent	Mature	Harvest
04/20 - 06/01	05/01 - 06/15	07/01 - 08/10	07/15 - 09/01	08/10 - 09/20	09/01 - 10/20	09/10 - 11/15

So we got the summary datasets (Appendix 7.1.3) for District 20 and District 60 about the average precipitation and temperature by each growth stage for each year from 2000 to 2018, which have 19 observations and 21 climate properties, and built our predictive models based on these datasets. As the averages of maximum temperature and minimum temperature are significantly correlated, we build models using average precipitation and average maximum temperature or average minimum temperature separately for each of the districts. We still use the corn yield (BU/ACRE) for each year as our response variable and implement the technique of AIC selection for both directions, that is, the algorithm can either add or drop variables from the original full model to optimize AIC.

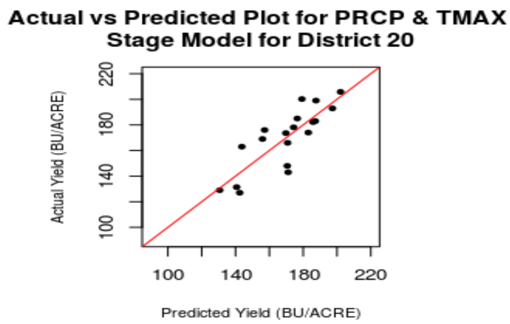
For District 20, the model we built based on precipitation and maximum temperature is: *Yield (BU/ACRE)* = 495. 053 + 119.327 * *Emerge PRCP* + 303.12 * *Dough PRCP* - 7.861 * *Dough TMAX* + 5.227 * *Dent TMAX* - 2.125 * *Harvest TMAX*, in which *PRCP* stands for the average precipitation, and *TMAX* stands for maximum temperature of that growth stage. All the coefficients except for that of *Harvest TMAX* are significant at p = 0.1 significance level (Appendix 7.1.4) with *Dough TMAX* being the most significant factor with a p-value of 0.001262, and the p-value of the F-test which compares this model with the model that only uses the intercept as its predictor is 0.005981, which means the model is also overall significant. The Multiple R-Squared is 0.681, meaning that 68.1% of the variability in the corn yield can be explained by the predictors, and the Adjusted R-Squared is 0.558, indicating the model is effective. The significant factors that have a positive effect on the corn yield of this district include the precipitation of emerging and dough and the maximum temperature of denting, and those have a negative effect is the maximum temperature of dough. The overall most influential factor that is significant is the precipitation of dough.

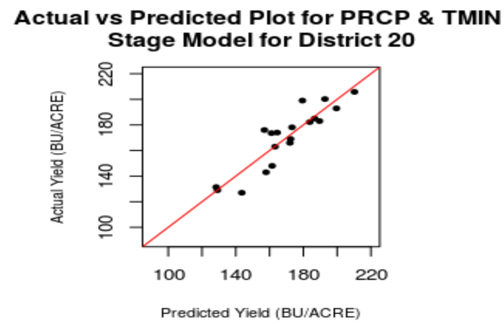
For the same district, the model we built based on precipitation and minimum temperature is: *Yield (BU/ACRE)* = 526.285 - 66.224 * *Plant PRCP* + 142.258 * *Emerge PRCP* + 353.579 * *Silk PRCP* + 504.496 * *Dough PRCP* - 452.319 * *Dent PRCP* + 201. 151 * *Mature PRCP* + 3.515 * *Planted TMIN* - 5.616 * *Emerge TMIN* - 6.956 * *Silk TMIN* + 8.063 * *Dented TMIN* - 7.359 * *Harvest TMIN*, in which *TMIN* stands for the average minimum temperature of that growth stage. The coefficients for *Silk PRCP*, *Dough PRCP*, *Silk TMIN*, *Dent TMIN*, and *Harvest TMIN* are significant under p = 0.1 significance level (Appendix 7.1.4) with *Silk TMIN* being the most significant factor with a p-value of 0.00585, and the p-value of the F-test is 0.08716 which shows that the model is overall significant. The Multiple R-Squared is 0.818, meaning that 81.76% of the variability in the corn yield can be explained by the predictors, and the Adjusted R-Squared is 0.531, from which we can see that although the Multiple R-Squared of this model is greater than that of the model with precipitation and maximum temperature, there is no obvious difference in the values of Adjusted R-Squared of these two models, so the greater Multiple R-Squared may mainly result from the increased predictors and thus we do not recommend one of the models over another. The significant factors that have a positive effect on the corn yield of this district include the precipitation of silking and dough and the minimum temperature of denting, those have a negative effect include the minimum temperature of silking and harvest. The most influential factor that is significant is the precipitation of dough.

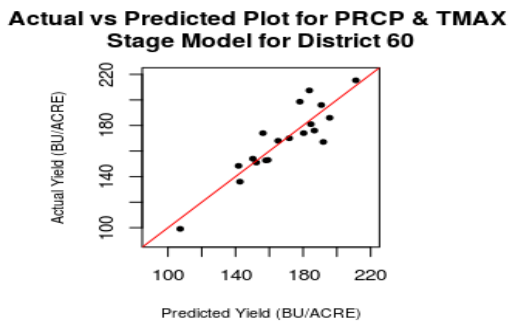
For District 60, the model we built based on precipitation and maximum temperature is: *Yield (BU/ACRE)* = 490.624 + 65.249 * *Emerge PRCP* + 252.249 * *Dent PRCP* - 198.729 * *Mature PRCP* - 142.653 * *Harvest PRCP* + 2.325 * *Emerge TMAX* - 2.487 * *Silk TMAX* - 7.414 * *Dough TMAX* + 4.165 * *Dent TMAX*. The coefficients for *Dent PRCP*, *Mature PRCP*, *Harvest PRCP*, *Dough TMAX*, and *Dent TMAX* are significant under p = 0.1 significance level (Appendix 7.1.5) with *Dent TMAX* being the most significant factor with a p-value of 0.023, and the p-value of the F-test is 0.007641, indicating that the model is overall significant. The Multiple R-Squared is 0.814, meaning that 81.4% of the variability of the response variable can be explained by the predictors, and the Adjusted R-Squared is 0.664. The significant factors that have a positive effect on the corn yield of this district include the precipitation of denting and the maximum temperature of denting, and those have a negative effect include precipitation of mature and harvest, and the maximum temperature of dough. The most influential factor that is significant is the precipitation of denting.

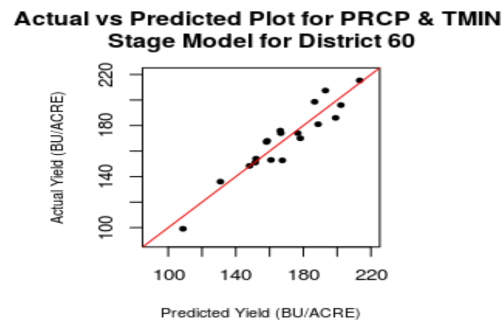
For the same district, the model we built based on precipitation and minimum temperature is: *Yield (BU/ACRE)* = 600.784 + 151.055 * *Emerge PRCP* + 383.585 * *Dough PRCP* - 90.994 * *Dent PRCP* + 4.38 * *Plant TMIN* - 2.222 * *Emerge TMIN* - 5.815 * *Silk TMIN* - 6.242 * *Dent TMIN* + 8.747 * *Mature TMIN* - 6.657 * *Harvest TMIN*. All coefficients except for that of *Emerge TMIN* are significant under p = 0.1 significance level (Appendix 7.1.5) with *Dough PRCP* being the most significant factor with a p-value of 0.000473, and the p-value of the F-test is 0.002069, indicating that the model is overall significant. The Multiple R-Squared is 0.893, meaning that 89.3% of the variability of the response variable can be explained by the predictors, and the Adjusted R-Squared is 0.786. Both Multiple R-Squared and Adjusted R-Squared of this model are better than those of the model with precipitation and maximum temperature, so the predictors of this model can better account for the changes in the corn yield. The significant factors that have a positive effect on the corn yield of this district include the precipitation of emerging and dough as well as the minimum temperature of planting and mature, and those have a negative effect include the precipitation of denting and the minimum temperature of planting, silking, and denting. The most influential factor that is significant is the precipitation of dough.

The following graphs show the Actual vs. Predicted corn yields for each of the four models discussed above:







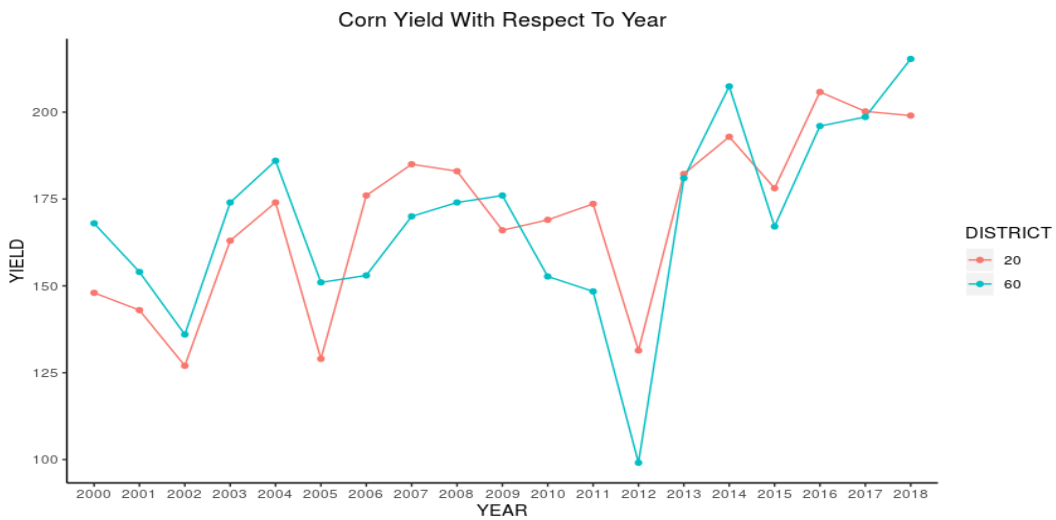


We can see that almost all the points are falling around the red line, which is the line of the actual equal to the predicted value, without any obvious outliers. The MAEs of the *PRCP & TMAX* and *PRCP & TMIN* models for District 20 are 10.863 and 7.982, respectively, and those of the *PRCP & TMAX* and *PRCP & TMIN* models for District 60 are 8.890 and 7.460, respectively, so according to the MAE, the models based on the average precipitation and minimum temperature perform better than those based on the average precipitation and maximum temperature. Looking at the values of Adjusted R-Squared and MAEs, the models incorporating growth stage information have higher values of Adjusted R-Squared and lower MAEs than the yearly-averages models, so we can conclude that the growth-stages models perform better.

5. INNOVATIVE TECHNOLOGY ANALYSIS

5.1 Overall Corn Yield Trend Analysis

In order to figure out whether technology affects the yield of corn in general, we assume that technology is growing year by year. From the yield vs. year graph below, we find that corn yield has an increasing trend with respect to time. Based on common sense, we made the assumption that technology has been growing year by year. By looking at the corn yield vs year graph, we find the graph has a general positive trend depending on the year, so that we think that technology has a general positive impact on corn yield. Yet we also find that there are some points on the graph showing a relatively low corn yield, as is analyzed previously, drought is a cause for low corn yield, but other reasons such as floods and frozen weather can also damage corn during different stages.

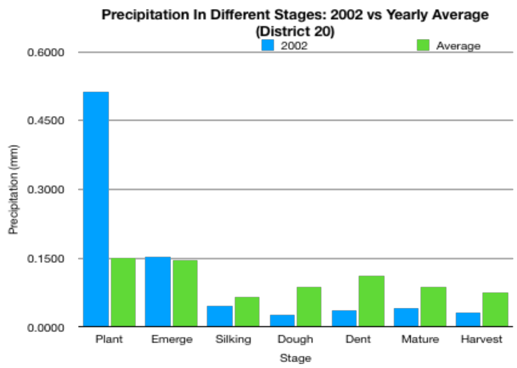


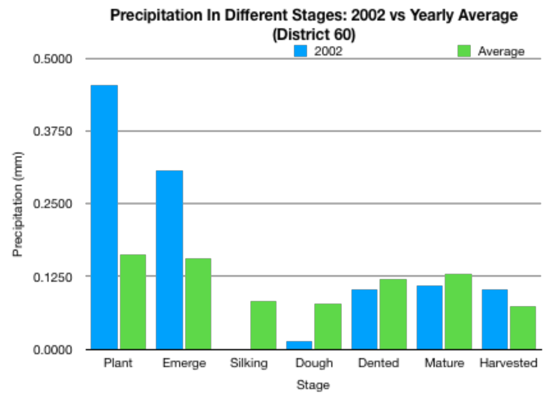
In order to show the pure effect of technology on corn yield, we want to pull out other factors that we already know to have an effect on corn yield. The yield vs. year graph shows that yield valleys occur in the years 2002, 2005, 2012 and 2015. So we decide to take a closer look at the temperature and precipitation for different stages in these four years.

We use the same dataset as we used to fit the growth stage predictive model to plot the precipitation vs stage histogram graph for district 20 and district 60. In order to decide if the temperature or precipitation at a particular stage can be considered as abnormal, we made comparisons between the actual temperature/precipitation variable in each stage to its 19-year average. The following shows the unusual precipitation we have detected:

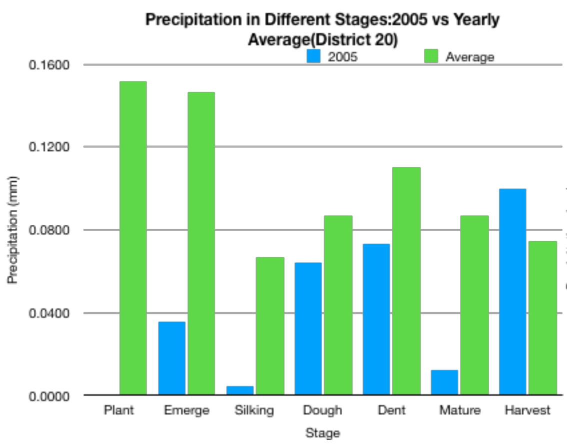
5.2 Pointing Out Climate Factors for Abnormal Years

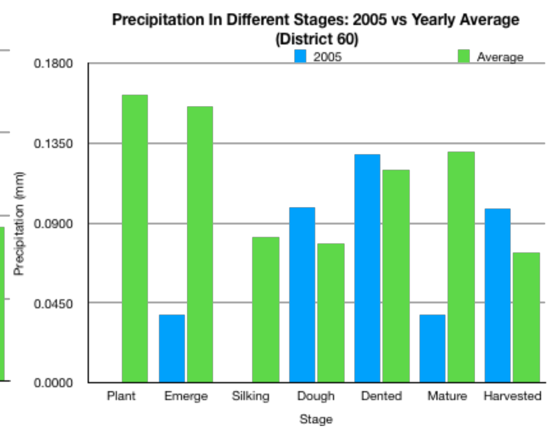
Looking at the two histograms for 2002's precipitation, we believe that unusual high precipitation in the plant stage is the main attribution for low corn yield in 2002. For District 20, the average precipitation in the plating stage is 0.1516mm, however, in 2002, the precipitation in the plating stage is 0.5122 mm, which is more than 3 times higher than usual, from the stage predictive model, the coefficient for *Plant PRCP* is -66.224, which means precipitation at plant stage is negatively correlated with yield. For District 60, we noticed that floods occurred during both the plant stage and the emerging stage, for the plating stage the average daily precipitation is 0.4550 mm, and 0.3075 mm for the emerging stage, which is much higher than the average precipitation in this region, thou the predictive model does not include the planting stage precipitation, we still think that a three times higher precipitation would be harmful to corn yield according to NASS's recommendation.





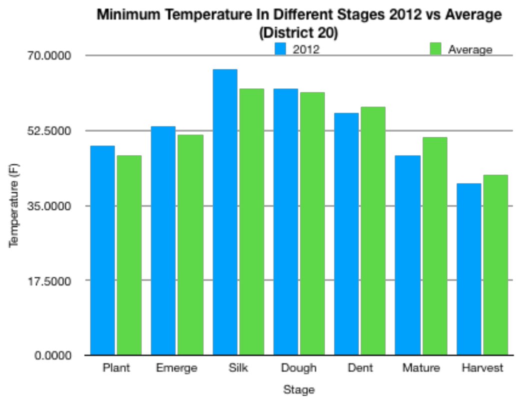
Just as we mentioned previously, droughts occurred during the year 2005, we find droughts happened in multiple stages during the year 2005. By looking at the two histograms below, we find that low precipitation during multiple stages is the main cause of droughts in 2005.

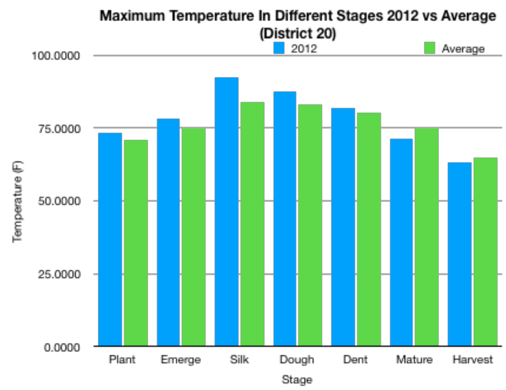


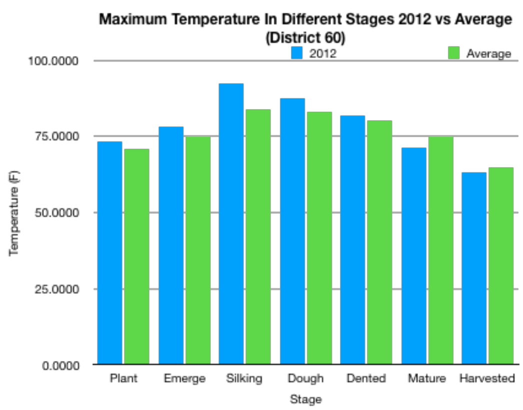


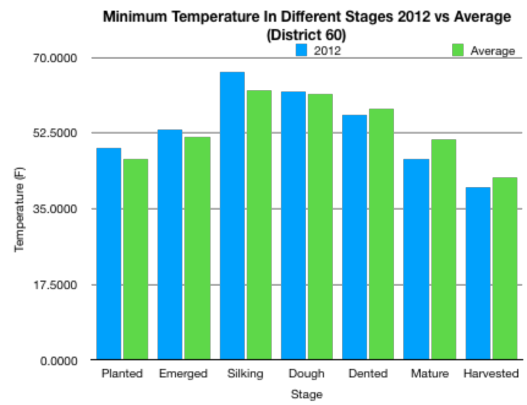
Another severe drought happened in 2012, however, from the analysis, we find that precipitation is actually normal if we compare the rainfall in 2012 to the average precipitation. However, comparing to the 19-year average, temperatures were higher overall during the first half of 2012 and colder during the second half of 2012. During the plant, emerge, silk and dough stage, maximum and minimum temperature is 4~8 degrees higher, especially in the silk stage, the maximum temperature is 8.36 degrees Fahrenheit higher than the average and min temperature is 4.34 degrees Fahrenheit higher than the average.

District 60 has just the same situation as district 20, by looking at the two histograms for district 60, for the silking stage, district 60's maximum temperature is 8.3615 degrees higher than usual, and the minimum temperature is 4.3635 degrees higher than the average.

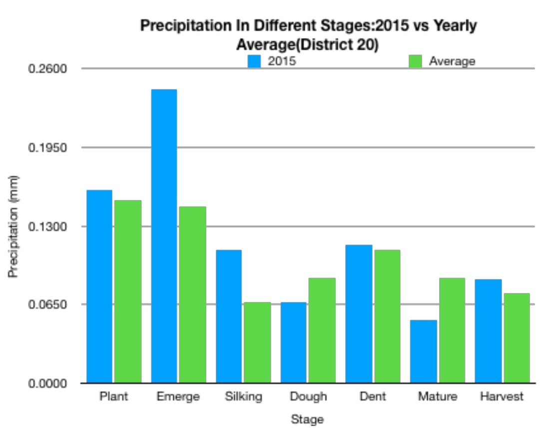


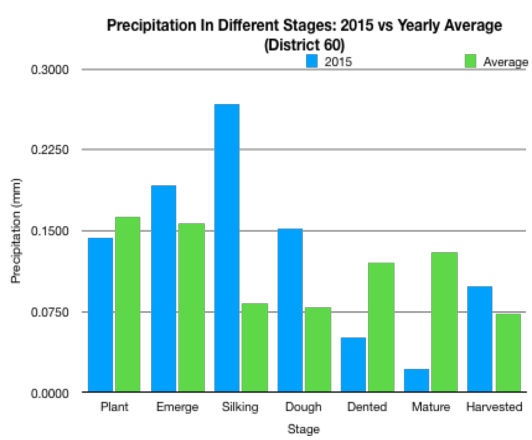






The two histograms below show the precipitation in different stages during the year 2015. 2015's corn yield is a low for the period 2013-2018. We find that unusual high precipitation is the main reason for relatively low corn yield. For District 20, daily precipitation is 0.2432mm, which is 1.7 times more than the average precipitation. Floods seem to be more severe during the silking stage, daily precipitation for the silking stage is 0.2674mm, which is 3 times more than the average precipitation.





6. CONCLUSION

According to the requirements of the client, our goals are to tell if there is a difference between the corn yield in district 20 and district 60 over the past two decades, point out the reasons behind the difference, predict the corn yield by building statistical models, and figure out if innovative technology have an impact on the corn yield.

From the 'District Variance Analysis' section, although there is no significant difference between the yield between the districts, we see a statistically and significantly higher harvested percentage in district 60 than district 20.

From the analysis of the precipitation variable, we see that district 60 also has a higher average weekly precipitation and higher annual precipitation which may be the reason for the higher harvested percentage.

From the analysis of the temperature variable, we can determine that the difference in temperatures between the two districts is only significant for the average maximum temperature. The average maximum temperature is higher in district 60 which can explain why district 60 usually has a higher yield. Higher temperatures lead to higher yields.

From the analysis of the predictive models, we define the most significant factor of a model as the one with the smallest p-value and the most influential factor as the one with the largest absolute coefficient. We summarized the most significant and influential factors given by each of the growth-stage models into the following table:

		Most Significant Factor	Most Influential Factor being Significant
District 20	**Model with PRCP & TMAX**	The average maximum temperature during dough	The average precipitation during dough
	Model with PRCP & TMIN	The average minimum temperature during silking	The average precipitation during dough
District 60	**Model with PRCP & TMAX**	The average maximum temperature during denting	The average precipitation during denting
	Model with PRCP & TMIN	The average precipitation during dough	The average precipitation during dough

We can see that the precipitation is highly influential in the corn growth as it is the most influential factor of all the models. For the growth stages, we believe that the stages of dough, silking, and denting play a relatively significant role in the growing processes which the farmers should pay special attention to.

From the analysis of the innovative technology, we found that the corn yield has a general increasing trend over the years except for several valleys at the years 2002, 2005, 2012, and 2015. We found that all of these yield valleys were caused by the abnormal climate of those years, so it is reasonable to believe that increasingly innovative technology is the push behind the ascending yield.

Because higher corn yield is linked with a lower selling price according to the economic supply and demand curve as the demand for corn yield is relatively stable across the years. We recommend farmers to pay attention to weather conditions, especially in the dough stage. We can also advise farmers to be sensitive to any abnormal rainfall because we find droughts or floods can be destructive. If we notice that a flood or drought happens in an early growth stage, we would recommend farmers to plan ahead for expected low corn yield.

APPENDIX

7.1 Relevant Tables

7.1.1 Datasets for Building the Yearly Averages Models
District 20

YEAR	PRCP	TMAX	TMIN	TOBS	ACRES	DISTRICT	YIELD
2000	0.10392	61.28721	40.14396	44.83019	1062000	20	148
2001	0.104786	62.30705	41.56224	46.31466	1048000	20	143
2002	0.079157	61.22169	40.54765	45.28039	1064000	20	127
2003	0.09	60.61448	38.8131	43.34276	1034000	20	163
2004	0.096374	60.58516	39.98901	44.60714	1100000	20	174
2005	0.067619	61.2328	39.87654	44.31217	1112000	20	129
2006	0.111304	61.72544	41.74747	46.08907	1028000	20	176
2007	0.127906	60.74832	40.19404	44.71758	1229000	20	185
2008	0.125291	58.61425	37.98982	42.83349	1167000	20	183
2009	0.139088	58.27095	38.92831	43.37337	1114000	20	166
2010	0.097173	61.3011	40.09116	44.59392	1137000	20	169
2011	0.136569	58.70793	39.51792	43.99891	1185000	20	173.6
2012	0.073672	65.04096	42.80508	47.5226	1153000	20	131.4
2013	0.111001	59.49499	38.47783	43.97282	1135000	20	182.2
2014	0.111975	57.37941	37.1763	42.50917	1038000	20	192.9
2015	0.129711	59.45593	39.25836	44.11702	1034000	20	178.1
2016	0.113657	62.57143	42.94	47.61143	1020000	20	205.8
2017	0.105849	62.15965	41.16546	46.13788	974000	20	200.2
2018	0.091452	59.96593	40.05926	44.90519	996000	20	199

District 60

YEAR	PRCP	TMAX	TMIN	TOBS	ACRES	DISTRICT	YIELD
2000	0.098405	64.94018	43.99233	49.98926	1486000	60	168
2001	0.116181	62.94137	42.04523	47.64992	1525000	60	154
2002	0.127988	64.51085	43.6657	49.233	1524000	60	136
2003	0.110388	59.66262	38.79369	45.05825	1573000	60	174
2004	0.104993	62.69896	42.18927	48.65723	1602000	60	186
2005	0.082523	64.99719	43.29081	49.89493	1681000	60	151
2006	0.091429	64.99804	43.60274	49.65851	1597000	60	153
2007	0.08493	65.16465	43.16186	48.70419	1835000	60	170
2008	0.150201	60.96463	40.44073	46.03537	1679000	60	174
2009	0.128167	62.61444	42.26778	47.94889	1717000	60	176
2010	0.135775	64.51176	43.58824	48.91765	1758000	60	152.7
2011	0.18321	62.35152	42.89567	47.70787	1754000	60	148.4
2012	0.093333	69.21348	45.55306	51.45443	1727000	60	99.1
2013	0.112337	61.56903	40.54043	46.62032	1600000	60	181
2014	0.10775	61.37255	40.63674	46.72136	1654000	60	207.4
2015	0.134626	62.13273	41.7384	48.16366	1627000	60	167.1
2016	0.070717	66.43759	45.40029	52.52224	1638000	60	196
2017	0.09867	66.66524	44.79399	51.91845	1508000	60	198.6
2018	0.088129	63.88602	42.50323	49.28602	1457000	60	215.3

7.1.2 R Outputs of the Yearly Averages Models
District 20

```
lm(formula = YIELD ~ PRCP + TMAX + TMIN + PRCP:TMAX + TMAX:TMIN,
    data = factor_year20)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.648  -7.271   1.606   6.716  29.548 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)  11084.501   4370.104   2.536  0.02482 * 
PRCP        -25371.630   8351.492  -3.038  0.00952 **
TMAX          -182.966     72.098  -2.538  0.02476 * 
TMIN          -199.392     94.513  -2.110  0.05484 . 
PRCP:TMAX      427.621    138.251   3.093  0.00856 **
TMAX:TMIN        3.338      1.550   2.154  0.05060 . 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.48 on 13 degrees of freedom
Multiple R-squared:  0.6226,	Adjusted R-squared:  0.4775 
F-statistic:  4.29 on 5 and 13 DF,  p-value: 0.0159
```

District 60

```
lm(formula = YIELD ~ PRCP + TMAX + TMIN, data = factor_year
60)

Residuals:
    Min      1Q  Median      3Q     Max 
-24.212 -13.479  -0.228   3.437  46.335 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  933.316    191.340   4.878 0.000201 ***
PRCP        -820.936    237.999  -3.449 0.003577 ** 
TMAX         -24.251      8.055  -3.011 0.008782 ** 
TMIN          20.464      9.656   2.119 0.051171 .  
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.4 on 15 degrees of freedom
Multiple R-squared:  0.528,	Adjusted R-squared:  0.4336 
F-statistic: 5.592 on 3 and 15 DF,  p-value: 0.008876
```

7.1.3 Datasets for Building the Growth Stage Models

District 20

YEAR	PRCP_Planted	PRCP_Emerged	PRCP_Silking	PRCP_Dough	PRCP_Dented	PRCP_Mature	PRCP_Harvested
2000	0.0085	0.162670455	0.0625	0.054285714	0.0475	0.056785714	0.05
2001	0.0175	0.140909091	0.021	0.148928571	0.1465625	0.1515625	0.218181818
2002	0.5121875	0.151931818	0.04625	0.026785714	0.035357143	0.042142857	0.032
2003	0.060833333	0.061875	0.027083333	0.015	0.032916667	0.038214286	0.108863636
2004	0.149166667	0.082916667	0.0525	0.0325	0.0321875	0	0.007777778
2005	0	0.035714286	0.004166667	0.063888889	0.073333333	0.012222222	0.1
2006	0.22	0.160555556	0.005833333	0.06125	0.128666667	0.106666667	0.100740741
2007	0.205625	0.22	0	0.070555556	0.062857143	0.035	0.005357143
2008	0.179	0.225	0.182	0.009047619	0.267619048	0.364	0.008148148
2009	0.030416667	0.002222222	0.019333333	0.1075	0.098148148	0.020416667	0.013428571
2010	0.153902439	0.226136364	0.097948718	0.078071429	0.07225	0.046666667	0.043538462
2011	0.235660377	0.21046729	0.113409091	0.160654206	0.145505618	0.133055556	0.152222222
2012	0.169125	0.153604651	0.039342105	0.074565217	0.114875	0.108444444	0.081967213
2013	0.157866667	0.164942529	0.033157895	0.1	0.100632911	0.053789474	0.101574803
2014	0.173670886	0.165764706	0.065974026	0.139680851	0.223625	0.116382979	0.056269841
2015	0.15975	0.243176471	0.10987013	0.067222222	0.114444444	0.052428571	0.085943396
2016	0.150769231	0.150470588	0.183733333	0.225054945	0.201558442	0.063510638	0.056434109
2017	0.200253165	0.073837209	0.1348	0.132340426	0.0575	0.121477273	0.113495935
2018	0.095479452	0.146	0.062077922	0.081290323	0.142467532	0.124431818	0.084237288

TMAX_Planted	TMAX_Emerged	TMAX_Silking	TMAX_Dough	TMAX_Dented	TMAX_Mature	TMAX_Harvested
71.5	76.35227273	80.45833333	81.03571429	80.96875	79.10714286	68.91666667
75.6875	74.625	87.25	83.71428571	79.46875	73	67.36363636
67.125	71.50568182	89.4375	88.14285714	85.82142857	79.85714286	63.275
71.22222222	72.75	83.83333333	83.5625	81.625	74.03571429	62.56818182
67.33333333	73.125	77.8	77	78.9375	76.55	65.41666667
66.94444444	69.28571429	88	85.77777778	83.53333333	84.05555556	70.73333333
64.2	70.05555556	86.91666667	84.20833333	79.6	75.83333333	61
75.5625	76.13333333	81.91666667	79.72222222	78.23809524	76.5	69.82142857
70.25	76.8125	81.4	80.47619048	80.19047619	77.4	58.33333333
73.95833333	77.44444444	78.46666667	78.20833333	69.92592593	58.54166667	50.71428571
72.09756098	75.46969697	85.90598291	86.13571429	83.325	75.86111111	69.7025641
67.58490566	74.95327103	87.875	85.60747664	78.50561798	70.94444444	64.20261438
73.2125	78.1744186	92.25	87.60869565	81.7625	71.31111111	63.03278689
72.68	74.86206897	81.80263158	82.07608696	81.5443038	76	65.96062992
70.89873418	74.49411765	79.77922078	81.27659574	78.8375	69.96808511	62.57936508
69.85	75.48235294	81.85714286	82.68055556	80.75555556	73.41428571	66.90566038
70.57692308	75.24705882	83.37333333	83.69230769	81.92207792	76.59574468	70.28682171
68.43037975	74.58139535	82.49333333	80.67021277	78.60526316	77.28409091	66.53658537
76.09589041	79.7125	83.06493506	82.52688172	83.1038961	74.18181818	63.59322034

TMIN_Planted	TMIN_Emerged	TMIN_Silking	TMIN_Dough	TMIN_Dented	TMIN_Mature	TMIN_Harvested	YIELD
44.2	54.63636364	61.625	61.89285714	61	57.42857143	46	148
49.5625	53.17613636	67.15	63.78571429	59.9375	52.625	44.27272727	143
43.75	49.11931818	64.0625	62.60714286	57.78571429	50.39285714	39.55	127
43.80555556	45.96875	61.45833333	59.625	57.58333333	50.32142857	42.18181818	163
42.20833333	49.66666667	57.8	55.6875	55	50.8	40.08333333	174
41.33333333	45.38095238	62.33333333	63.83333333	59.33333333	57.44444444	46.26666667	129
43.93333333	48.22222222	64.33333333	63.91666667	59.6	51.66666667	39.48148148	176
49.3125	52.86666667	57.5	61.27777778	58.71428571	52.4375	43.67857143	185
46.35	51.3125	60.06666667	56.14285714	55.52380952	52.13333333	38.40740741	183
47.33333333	52.94444444	59.66666667	58.54166667	50.85185185	39.95833333	32.25714286	166
48.69105691	53.78030303	65.77777778	65.05	59.05	48.09027778	42.70769231	169
46.69811321	52.68224299	67.97727273	65.12149533	56.50561798	48.4537037	42.49019608	173.6
49.05	53.38372093	66.73684211	62.15217391	56.5625	46.6	40.00819672	131.4
49.6	51.5862069	60.31578947	58.98913043	57.34177215	51.02105263	42.18897638	182.2
47.92405063	51.42352941	59.46753247	61.72340426	59.2625	48.42553191	40.5952381	192.9
47.4875	53.84705882	61.66233766	62.66666667	60.26666667	50.81428571	44.85849057	178.1
47.64102564	51.44705882	64.98666667	65.62637363	62.67532468	56.18085106	48.84496124	205.8
45.44303797	50.69767442	60.77333333	58.78723404	54.59210526	52.72727273	45.67479675	200.2
50.89041096	56.45	61.77922078	61.92473118	61.83116883	51.93181818	42.88135593	199

District 60

YEAR	PRCP_Planted	PRCP_Emerged	PRCP_Silking	PRCP_Dough	PRCP_Dented	PRCP_Mature	PRCP_Harvested
2000	0.069	0.145609756	0.000833333	0.022857143	0.02	0.137142857	0.139444444
2001	0.07375	0.204651163	0	0.006363636	0.020833333	0.03	0.069090909
2002	0.455	0.3075	0	0.012857143	0.102142857	0.108571429	0.101764706
2003	0.107777778	0.0925	0.125	0.105	0.133333333	0.135714286	0.082727273
2004	0.318571429	0.06	0.09	0.030833333	0.004	0	0.173
2005	0	0.038	0	0.098888889	0.128888889	0.038333333	0.098
2006	0.146666667	0.135384615	0	0.065	0.058666667	0	0.018181818
2007	0.145	0.176666667	0.001666667	0.004444444	0.0025	0.002857143	0.01
2008	0.2	0.228888889	0.019285714	0.07	0.762941176	1.06	0.00047619
2009	0.01	0.02125	0.010714286	0.056666667	0.0905	0.055	0.006315789
2010	0.20159292	0.224344262	0.280769231	0.240526316	0.131320755	0.076911765	0.139444444
2011	0.249875	0.39244186	0.116595745	0.043421053	0.042368421	0.095735294	0.069090909
2012	0.113977273	0.061684211	0.043243243	0.036616541	0.134150943	0.156511628	0.101764706
2013	0.232136752	0.228671875	0.092242991	0.074016393	0.069439252	0.106615385	0.082727273
2014	0.093055556	0.151858407	0.121111111	0.125887097	0.205740741	0.206376812	0.173
2015	0.143	0.192413793	0.26739726	0.151555556	0.051486486	0.022209302	0.098
2016	0.137974684	0.08	0.208730159	0.122597403	0.116538462	0.0575	0.018181818
2017	0.307866667	0.147804878	0.103421053	0.124782609	0.075769231	0.073666667	0.01
2018	0.077162162	0.07691358	0.076	0.09725	0.130333333	0.107272727	0.00047619

TMAX_Planted	TMAX_Emerged	TMAX_Silking	TMAX_Dough	TMAX_Dented	TMAX_Mature	TMAX_Harvested
75.6	78.24390244	82.33333333	83.07142857	83.0625	80.07142857	69.22222222
76.375	75.76744186	90.8	86.72727273	83.91666667	78.1	68.72727273
69.8125	73.60227273	89.75	88.85714286	84.78571429	79.85714286	65.88235294
72.11111111	73.875	85.66666667	85.25	81.16666667	74	62.81818182
66	75.33333333	78.6	80.58333333	82.26666667	80.1	68.2
70.41176471	73.45	90.16666667	89.72222222	88.11111111	85.5	72.83333333
69.33333333	71.76923077	88	87.45833333	83.53333333	79.25	64.31818182
78.11111111	77.6	83	88.77777778	85.29166667	81.57142857	73.96666667
73.71428571	79.33333333	85.42857143	82.57894737	80	77.09090909	59
75.18181818	79.3125	82.21428571	79.76190476	70.85	63.57142857	54.52631579
72.86725664	77.51639344	88.35897436	87.92481203	83.11320755	76.16911765	69.22222222
67.375	74.81395349	88.80851064	87.46052632	81.42105263	72.36764706	68.72727273
77	80.03157895	94.95495495	90.80451128	83.48113208	71.43410853	65.88235294
70.05128205	74.0390625	83.8411215	83.90163934	84.74766355	77.7	62.81818182
73.77777778	76.32743363	81.28703704	82.65322581	80.2962963	71.84782609	68.2
72.55	77.1954023	84.23287671	84.41111111	84.04054054	77.62790698	72.83333333
73.69620253	78.54117647	85.93650794	86.77922078	85.01282051	80.69791667	64.31818182
72.21333333	78.29268293	87.86842105	85.23913043	82.41025641	79.67777778	73.96666667
81.22972973	86.60493827	88.05714286	85.825	85.96666667	80.84848485	59

TMIN_Planted	TMIN_Emerged	TMIN_Silking	TMIN_Dough	TMIN_Dented	TMIN_Mature	TMIN_Harvested	YIELD
47.8	57.48780488	64.08333333	64.92857143	63.5625	61	47.66666667	168
53	55.94186047	68	62.45454545	61	54.6	43.54545455	154
49.375	53.79545455	69.875	68.85714286	61.64285714	53.35714286	42.76470588	136
48.55555556	50.625	62.83333333	61.125	56	49	43.63636364	174
50	57.83333333	58.8	55.75	53.86666667	50	40.9	186
43	48.4	64.5	65.77777778	61.38888889	58.94444444	48.03333333	151
46.33333333	51.84615385	64.5	65.91666667	63.26666667	53.25	39.63636364	153
54.05555556	56.86666667	59.66666667	63.11111111	59.79166667	54.33333333	47.56666667	170
51.47619048	57.66666667	64.64285714	59.68421053	57	54.27272727	38.57142857	174
53.72727273	56.75	62.5	60.47619048	51.95	40.64285714	34.31578947	176
52.11504425	57.98360656	67.88888889	66.69924812	60.21698113	50.38235294	47.66666667	152.7
49.2	54.87209302	68.27659574	63.80263158	57.19736842	48.61764706	43.54545455	148.4
52.89772727	55.12631579	66.90990991	61.63909774	55.90566038	46.86821705	42.76470588	99.1
49.99145299	53.984375	62.04672897	61.85245902	59.43925234	51.93076923	43.63636364	181
51.97222222	56	59.73148148	62.60483871	60.33333333	49.43478261	40.9	207.4
52.0125	58.62068966	65.80821918	64.07777778	61.2027027	52.74418605	48.03333333	167.1
53.11392405	56.98823529	66.6984127	67.49350649	63.92307692	56.85416667	39.63636364	196
51.34666667	56.18292683	65.17105263	61.92391304	56.42307692	53.72222222	47.56666667	198.6
57	63.56790123	66.17142857	64.425	65.66666667	59.24242424	38.57142857	215.3

7.1.4 R Outputs of the Growth Stage Models - District 20

The Model with PRCP & TMAX

```
lm(formula = YIELD ~ PRCP_Emerged + PRCP_Dough + TMAX_Dough +
    TMAX_Dented + TMAX_Harvested, data = phase_20)

Residuals:
    Min      1Q  Median      3Q     Max 
-28.112  -6.920  -1.645   9.956  20.944 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     495.053    113.199   4.373 0.000754 ***
PRCP_Emerged    119.327     60.076   1.986 0.068504 .  
PRCP_Dough      303.120     87.865   3.450 0.004309 ** 
TMAX_Dough       -7.861      1.919  -4.096 0.001262 ** 
TMAX_Dented       5.227      2.425   2.156 0.050422 .  
TMAX_Harvested   -2.125      1.205  -1.764 0.101147    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.08 on 13 degrees of freedom
Multiple R-squared:  0.6807,	Adjusted R-squared:  0.5578 
F-statistic: 5.542 on 5 and 13 DF,  p-value: 0.005981
```

The Model with PRCP & TMIN

```
lm(formula = YIELD ~ PRCP_Planted + PRCP_Emerged + PRCP_Silking +
    PRCP_Dough + PRCP_Dented + PRCP_Mature + TMIN_Planted + TMIN_Emerged +
    TMIN_Silking + TMIN_Dented + TMIN_Harvested, data = phase_20)

Residuals:
    Min      1Q  Median      3Q     Max
-16.566  -6.303  -1.576   6.122  19.565

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     526.285    199.239   2.641  0.03335 *
PRCP_Planted    -66.224     57.256  -1.157  0.28537
PRCP_Emerged    142.258    122.252   1.164  0.28269
PRCP_Silking    353.579    137.181   2.577  0.03661 *
PRCP_Dough      504.496    204.909   2.462  0.04334 *
PRCP_Dented    -452.319    247.677  -1.826  0.11055
PRCP_Mature     201.151    134.518   1.495  0.17848
TMIN_Planted      3.515      2.811   1.250  0.25136
TMIN_Emerged     -5.616      3.175  -1.769  0.12025
TMIN_Silking     -6.956      1.780  -3.907  0.00585 **
TMIN_Dented       8.063      4.106   1.964  0.09031 .
TMIN_Harvested   -7.359      3.838  -1.917  0.09670 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.56 on 7 degrees of freedom
Multiple R-squared:  0.8176,    Adjusted R-squared:  0.5311
F-statistic: 2.853 on 11 and 7 DF,  p-value: 0.08716
```

7.1.5 R Outputs of the Growth Stage Models - District 60

The Model with PRCP & TMAX

```
lm(formula = YIELD ~ PRCP_Emerged + PRCP_Dented + PRCP_Mature +
    PRCP_Harvested + TMAX_Emerged + TMAX_Silking + TMAX_Dough +
    TMAX_Dented, data = phase_60)

Residuals:
    Min      1Q  Median      3Q     Max
-24.848  -6.468  -1.798   4.644  23.716

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     490.624    160.296   3.061   0.0120 *
PRCP_Emerged     65.249     50.377   1.295   0.2243
PRCP_Dented     252.402    107.972   2.338   0.0415 *
PRCP_Mature    -198.729     80.107  -2.481   0.0325 *
PRCP_Harvested -142.653     72.223  -1.975   0.0765 .
TMAX_Emerged      2.325      1.324   1.755   0.1097
TMAX_Silking     -2.487      1.846  -1.347   0.2076
TMAX_Dough       -7.414      3.081  -2.406   0.0369 *
TMAX_Dented       4.165      1.555   2.679   0.0231 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.71 on 10 degrees of freedom
Multiple R-squared:  0.8135,    Adjusted R-squared:  0.6642
F-statistic: 5.451 on 8 and 10 DF,  p-value: 0.007641
```

The Model with PRCP & TMIN

```
lm(formula = YIELD ~ PRCP_Emerged + PRCP_Dough + PRCP_Dented +
    TMIN_Planted + TMIN_Emerged + TMIN_Silking + TMIN_Dented +
    TMIN_Mature + TMIN_Harvested, data = phase_60)

Residuals:
     Min       1Q   Median       3Q      Max
-14.8660  -7.7912   0.2557   8.1261  14.3492

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     600.784    118.988   5.049 0.000691 ***
PRCP_Emerged    151.055     44.697   3.380 0.008133 **
PRCP_Dough      383.585     71.941   5.332 0.000473 ***
PRCP_Dented     -90.994     26.840  -3.390 0.007996 **
TMIN_Planted      4.380      2.001   2.189 0.056318 .
TMIN_Emerged     -2.222      1.950  -1.140 0.283816
TMIN_Silking     -5.815      1.143  -5.088 0.000656 ***
TMIN_Dented      -6.242      2.017  -3.095 0.012829 *
TMIN_Mature       8.747      1.834   4.769 0.001016 **
TMIN_Harvested   -6.647      1.356  -4.903 0.000844 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.55 on 9 degrees of freedom
Multiple R-squared:  0.8929,	Adjusted R-squared:  0.7859
F-statistic: 8.339 on 9 and 9 DF,  p-value: 0.002069
```

7.2 Code

The code for this report will be submitted in the zip file. Each file has commented explanation in the code, a description in the title of the file, and/or an indication of what it is used for in the report. Important data files will also be included in the zip file containing this report.

7.3 Future Directions

From the analysis of the precipitation variable section, incorporating some of the 4 created variables may help improve the accuracy of predictive models because it is less biased and focuses on critical times indicated by the Illinois State Water Survey

We realized that many variables are highly correlated in this study, which means that the multicollinearity problem occurs in our predictive model. In a regression model, we hope that we can learn how a one-unit change of independent variables can cause variability in the dependent variable. However, in our case, due to multicollinearity, the coefficient estimates can swing wildly based on which variables to include in the model. This is why we find that the value of coefficient for a variable varies greatly in two different models we estimated to predict the same corn yield (For example, in the model that uses stage precipitation and maximum temperatures to estimate district 20’s corn yield, the coefficient for emerge PRCP is 119.327 and the coefficient for Dough PRCP is 303.12, but in the model that uses stage precipitation and minimum temperature to estimate district 20's corn yield, the coefficient for *Emerge PRCP* is 142.258 and the coefficient for Dough PRCP is 504.496), so that it is difficult for us to know the real effect of each attribute on corn yield. Some of the estimated coefficients we got from the variable are contradicted with the point-out analysis for abnormal corn yields. Even if a temperature variable has a positive coefficient, this does not mean that higher temperature is always the better. In the pull-out factor analysis, we realize that high temperature shall cause drought, which is actually detrimental towards corn yield.

Another limitation we had with the model is that we only had 19 years' observations, which is too few for us to predict a good model, if we want to better understand how temperature and precipitation are impact corn yield, more observations are needed in order to predict a more accurate model.

Moreover, for the innovative technology, though we already find that innovative technology plays a role in the increasing corn yields throughout the years, figuring out how to quantify the technology factor and putting it into the predictive models to further explore its impact on corn yield still require future relevant analyses.

7.4 References

Hollinger and Angel - 'Weather and Crops'
http://extension.cropsciences.illinois.edu/handbook/pdfs/chapter01.pdf
This source was used to indicate critical stages of corn growth to focus on for the precipitation variable.

State Climatologist Office for Illinois - 'Drought Trends in Illinois'
https://www.isws.illinois.edu/statecli/climate-change/ildrought.htm
This source was used to indicate that 2005 and 2012 were droughts using the Palmer Severity Index.

7.5 Group Member Contributions

Shihao Duan worked on the Introduction part, which includes Background & Motivation and Data Preparation.

Smruthi Iyengar worked comparing the temperatures in both district 60 and district 20. She created four violin plots in R to visualize the differences in temperature. She also conducted the T-test to determine if there were any statistical differences in temperature. She worked on section 2.2 and the code in the temperature data file.

Kagen Quiballo worked on the analysis of the precipitation variable under the 'Variable Analysis and Visualization' section which discusses methods for creating 4 different precipitation variables at the following website (*tinyurl.com/STAT443grp6-prcp-data*) in parallel with Illinois State Water Survey article, as well as visualizing 2005 and 2012 drought effects on annual corn yield. He also worked on the 'District Variance Analysis and Visualization' section which uses 2 sample t-tests to compare the average yield and average harvested percentage differences across districts 20 and 60. Lastly, he wrote the majority of all sections in the appendix.

Doris Wang worked on processing and cleaning datasets, calculating averages, and creating summary tables to prepare the variables used in the predictive models. Building, analyzing, and evaluating all the predictive models using R (the code file about predictive models), and making the slides and writing the report for the predictive model part. In addition, she wrote the majority of the Data Preparation part, formatted, and created a title page for the report. She created the visualization of Corn Yield with respect to Year used in the section of Innovative Technology Analysis. She also prepared and presented all the tables in the Appendix.

Lucy Zhao worked on analyzing "How technology has affected corn yield" and abnormal precipitation and temperatures each year to figure out the cause of low yield in specific years. She created 10 histograms to examine the effects of the abnormal temperature and precipitation on corn yield. She is also responsible for the future direction part.