Genetically Modified Rootworm-Resistant Corn, Risk, and Weather: Evidence from High Dimensional Methods

Serkan Aglasan[a] Barry K. Goodwin[b] Roderick Rejesus[c]

[a,b,c] Department of Agricultural and Resource Economics, North Carolina State University
[a] e-mail: saglasa@ncsu.edu

Selected Paper prepared for presentation at the 2020 Agricultural & Applied Economics Association Annual Meeting, Kansas City, MO
July 26-28, 2020

Genetically Modified Rootworm-Resistant Corn, Risk, and Weather: Evidence from High Dimensional Methods

Serkan Aglasan* Barry K. Goodwin† Roderick M. Rejesus‡

June 1, 2020

Abstract

This study analyzes whether genetically modified (GM) corn with rootworm resistant traits (GM-RW) decreases yield risk in the U.S. Central Corn Belt (CCB). A crop insurance actuarial performance measure, the loss cost ratio (LCR), is the primary variable used to represent yield risk. Since there are a large set of potential weather and other control variables that can influence yield risk, high dimensional methods are utilized in this study to maintain parsimony in the empirical specification, and facilitate estimation. Specifically, we employ the Cluster-Lasso (cluster-least absolute shrinkage and selection operator) procedure because it can produce uniformly valid inference on the main variable interest (i.e., the GM-RW variable) in a high dimensional panel data setting even in the presence of heteroskedastic, non-Gaussian, and clustered error structures. After controlling for a large set of potential weather confounders using Cluster-Lasso, we find consistent evidence that GM corn varieties with rootworm resistant traits reduce yield risk.

Keywords: Genetically modified corn, Yield risk, Loss-cost ratio, High-dimensional-sparse regression, Cluster-Lasso, Post-Cluster-Lasso, Post-double-selection, Inference under imperfect model selection

JEL Classification Numbers: C01, Q10, Q16

*Aglasan is PhD Candidate, Department of Agricultural and Resource Economics, NC State University, Raleigh, NC

†Goodwin is Distinguished Professor, Department of Agricultural and Resource Economics, NC State University, Raleigh, NC

‡Rejesus is Professor, Department of Agricultural and Resource Economics, NC State University, Raleigh, NC

1 Introduction

The United States (US) agricultural sector have benefited from technological advances over the years and these innovations have contributed to steady increases in corn yields over time. These technology-driven yield increases have been critical to feeding and clothing a growing population that is expected to reach 10 billion by 2050 (Wheeler and von Braun (2013), Godfray et al. (2010), McKenzie and Williams (2015), Armal et al. (2018)).

One of the key innovations that arguably have contributed to sustained corn yield increases over the last twenty years are genetically modified (GM) varieties with traits that confer resistance to various pests, and tolerance to herbicide applications (Moschini (2008), Fernandez Cornejo et al. (2014)). As such, there have been numerous studies that examined whether or not GM crop varieties have indeed increased mean yields over time (e.g., a non-exhaustive list includes Nolan and Santos (2012), Shi et al. (2013), Xu et al. (2013a), Chavas et al. (2014), Lusk et al. (2018)). The general consensus from this literature is that GM crop varieties have made major contributions to the increase in mean yields over the past few decades. However, it is important to note that some skepticism still remains, and there have been arguments that GM crops have not played a significant role in observed yield increases for major US commodity crops (Gurian-Sherman (2009), Foley (2014), Duke (2015), Hakim (2016)).

Aside from mean yield effects of GM crops, there is also a robust literature that has examined the effect of GM crops on yield risk (i.e., the impact of GM crops on yield variability and/or higher moments of the yield distribution) (Shi et al. (2013), Chavas and Shi (2015), Sanglestsawai et al. (2017), Goodwin and Piggott (2019)). In particular, Shi et al. (2013) and Chavas and Shi (2015) find that plot-level yield risk for GM corn with various traits tend to be smaller relative to non-GM corn, though climate variables are not controlled for in their specifications.[1] A number of studies have also examined how crop yield risk (specifically yield variability) is affected by climate change, and most of these studies find that crop yield risks are likely to increase under a scenario with increasing heat and moisture stresses due to climate change (Urban et al. (2012), Urban et al. (2015), Lobell et al. (2014), Tack et al. (2018)). Note that most of these "climate-change-focused"

[1] In addition, due to anecdotal evidence from Illinois farmers that GM crops suffered less yield loss from a localized 2005 drought, Monsanto soon after spearheaded a study of field trial data in the Midwest that indicates that planting GM crops result in lower yield risk relative to conventional varieties (Goodwin and Piggott (2019). Findings from this study became the basis for the the Biotech Endorsement (BE) introduced in 2008, which provides discounted crop insurance premiums to farmers who plant GM corn varieties (i.e., the lower yield risk for GM merits a reduction in premiums). The BE was eliminated in 2012 based on the argument that nearly all corn planted in the US is GM.

papers do not explicitly delineate between GM crops vis-à-vis non-GM crops, but since the data coverage of most of these studies coincide with high GM adoption rates, the implication has been that yield variability of GM crops has increased under climate change.

One notable paper in this literature is Lusk et al. (2018) where the authors explicitly evaluated the effect of GM corn adoption on mean yield and yield risk (e.g., yield variance and skewness), while at the same time accounting for climate variables. Lusk et al. (2018) argue that inferences on the yield effects of GM corn may be misleading if weather and climate variables are not accounted for in the empirical specification. A key remaining question is how one would select the climate variables to include in the specification. In their study, Lusk et al. (2018) do not find evidence that GM corn adoption leads to lower yield risk (e.g., insignificant GM effect on yield variance and skewness) when climate variables like degree day measures and precipitation terms are included in the specification. Overall, findings from the "GM-risk-climate-change" literature have largely been mixed and there is still considerable debate on this issue.

The objective of this study is to examine the yield risk effect of GM corn with rootworm resistant traits (GM-RW), while explicitly controlling for selected climate variables using the Cluster-Lasso (cluster least absolute shrinkage and selection operator) procedure developed by Belloni et al. (2016). To achieve this objective, we utilize a long-run, county-level panel data set from 1981 to 2015 with spatially consistent GM adoption, input expenditure, and weather information for three US Central Corn Belt states—Illinois, Indiana, and Iowa. In addition, county-level loss cost ratios (LCRs) are used as a measure of yield risk given that this variable provides straightforward information about the cost of providing a given level of risk protection when shortfalls in yields or revenues occurs (Goodwin and Piggott (2019). The LCR as a risk measure has the benefit of theoretically embodying the higher moment risk variables in one single risk indicator (i.e., rather than having separate measures for the higher moments, such as having one measure for variance and one for skewness, we only have one risk measure with LCR). We find that the Cluster-Lasso approach is a suitable procedure for selecting a parsimonious set of climate variables to serve as controls in empirical specifications that examine risk effects of GM crop varieties. Results from our study suggest that GM-RW tends to have lower yield risk relative to non-GM-RW after controlling for input expenditures and a large set of potential weather confounders through Cluster-Lasso.

The present study contributes to the literature in three ways. First, we introduce a

viable procedure that would allow for more systematic selection of climate variables to be included as control variables in studies that examine how crop varieties (like GM corn) influence mean yields and yield risk. Climate variables are typically "high dimensional" such that there are a large number that can be chosen as controls. Mainly for parsimony and ease of interpretation, Lusk et al. (2018) simply follows a common specification in the climate change literature where degree day measures and precipitation terms are used as the control climate variables (e.g., studies by Schlenker and Roberts (2006), Schlenker and Roberts (2009), and Annan and Schlenker (2015) also use this type of specification).[2] However, there are other studies that argue that vapor pressure deficit (VPD) measures, drought indices, and/or minimum/maximum temperatures may be more appropriate controls (see, for example, Yu and Babcock (2010), Xu et al. (2013a), and Lobell et al. (2014)). In addition, some studies use climate variables for the full growing season, while others use monthly measures of different climate variables. Given the interest in yield impacts of extreme weather events, there has also been significant variation as to the choice of climate variables (as well as the time dimension) used to represent "extreme weather". In general, the existing studies that have examined mean yield and yield risk effects of GM crops have largely used "ad hoc" methods for selecting climate variables to be included in the empirical specification (e.g., following what has been done in past studies, or based on some agronomic logic). Conceivably, the choice of weather controls may greatly influence the inferences drawn on the variable of interest (i.e., the GM corn in our case), and inadvertently excluding some important weather controls (for instance, excluding an important month or a key climate variable itself) might create biased results. The Cluster-Lasso procedure employed in this paper is a high dimensional method that "lets the data speak for itself" when choosing climate variables to be used as controls in the specification. A parsimonious specification can still be systematically established in a high dimensional setting that minimizes the risk of overfitting, multicollinearity issues, and noise from adding too many control variables.

Second, this study provides insights on the potential risk effects of a specific GM-trait—the rootworm resistant trait—rather than a bundle of different traits. As pointed out by Lusk et al. (2018), the GM nature of a crop is not a "single" thing. Over the period investigated in this study, some GM crops have one insect resistant trait (e.g., resistant

[2]Even though the studies cited here generally follow a specification that includes degree day and precipitation terms, it is important to note that the specific degree day variables and degree day thresholds utilized varies across studies.

to corn borer), some have herbicide tolerance only, and some GM crops are "stacked" varieties with multiple traits. However, the risk effects of GM rootworm resistant corn is of particular interest because the root architecture of these varieties might be more robust relative to corn varieties without this trait and, therefore, might be more resilient to extreme climate conditions such as droughts (Ma et al. (2009), Goodwin and Piggott (2019)).[3] Hence, the present study will provide new insights on whether the GM-RW in particular provides risk benefits when controlling for high dimensional climate variables.

Lastly, this article makes a contribution by utilizing spatially-consistent, county-level GM corn adoption data that allows for inferences about the relative contribution of the GM-RW to crop insurance LCR performance and yield risk. Note that the above mentioned paper by Lusk et al. (2018) uses a spatially-mismatched state-level GM corn adoption data merged with county-level yield measures (mainly due to data-availability constraints), which forced them to use estimation approaches that account for this issue. Other GM risk studies on the other hand use plot-level data sets without information on weather controls (Shi et al. (2013)). Hence, although the present study does not still use a panel data set at the finest level of aggregation (at the plot-level), it is still unique in the sense that it utilizes a spatially-consistent data set at a reasonable level aggregation and it has rich information on GM adoption and climate variables. In addition, the use of LCR as the main risk measure also has the benefit of providing additional insights about how the GM-RW per se has contributed to performance of the US crop insurance program over time (vis-à-vis the weather contribution). This type of information may be useful when there are concerns about adverse selection in the crop insurance program, and there is interest in fine-tuning the premium-rating procedure to find farm level indicators that better delineate lower-risk versus higher-risk producers.

The remainder of this paper proceeds as follows. First, the data sources used and some descriptive statistics are presented in the next section. This is followed by a description of the empirical approach that allows us to examine the GM-RW effect on yield risk with high-dimensional weather variables. The main estimation results are discussed in the fourth section. Conclusions and implications are then presented in the last section.

[3]Though it should be noted here that, if there is no rootworm infestation, the root balls of crops with rootworm resistant traits should theoretically have the same robustness as other varieties without rootworm resistant traits. However, when there is rootworm infestation, it is reasonable to theoretically expect that crops with rootworm resistant traits should be healthier and more robust than those without the trait.

2 Data and Descriptive Statistics

The data used in this study comes from several sources and are discussed in turn below. As mentioned above, the main dependent variable of interest in this study is the loss cost ratio (LCR), which is widely considered a good measure of yield risk. In this study, LCR is defined as the ratio of indemnities to liabilities—particularly for the Yield Protection (YP) and Revenue Protection (RP) policies in US crop insurance.[4] YP and RP are the two biggest programs in US crop insurance and constitutes the majority of policies chosen by farmers. The present study utilizes the Summary of Business data from the Risk Management Agency (RMA), which has county-level information on indemnities and liabilities (among other crop insurance related variables). The LCR data we use spans the period from 1981 to 2015, and covers three US Central Corn Belt (CCB) states (Illinois, Indiana, Iowa).

Given our interest in the risk effects of GM-RW, the main independent variable of interest in our empirical analysis is the county-level GM-RW adoption rate. In contrast to Lusk et al. (2018), who used state-level GM adoption data from the National Agricultural Statistics Service (NASS), we utilize a commercial GM adoption data set collected by a private company called GfK Kynetec. Each year, GfK Kynetec conducts surveys throughout the US of randomly sampled farmers about decisions pertaining to seed and pesticide choices. This data set is considered one of the most comprehensive and detailed source of GM adoption information. From these surveys, county-level aggregates of the acres planted to particular GM crops is calculated, and this county-level GM data was the one available to us. Note that GM corn became commercially available in 1996 (James (2016)) and we have GM corn adoption data for the period 1997–2015. As mentioned in the introduction, GM corn varieties may have a single trait (e.g., corn that is only herbicide tolerant) or they may have multiple "stacked" traits (e.g., corn that has herbicide tolerance plus, say, resistance to corn rootwoorm and european corn borer, has three traits). Given our specific interest on the risk effects of rootworm resistant traits, we created a GM-RW variable that represents the proportion of acres in each county planted to GM corn varieties with at least one rootworm resistant trait. That is, we count the number of acres planted to any GM corn variety with at least one rootworm trait (regardless if the GM corn has a single-trait or multiple-traits), and divide it by the total number of corn

[4]The YP policy was formerly called the Actual Production History (APH) policy prior to 2010. Moreover, note that YP only protects against yield shortfalls, while RP covers both yield and price losses.

acres to get the GM-RW adoption rate variable of interest. Hence, we have a county-level GM-RW adoption rate variable that ranges from zero to one for each year in our data set.

A key component of the empirics in this paper is the weather-related variables to be considered as controls in the specification. As alluded to in the previous section, weather variables are considered high-dimensional such that there are a large number of these variables (at different time-scales) that can be included in the main estimation equation. However, adding all of these variables in the specification can cause a number of different econometric problems (i.e., more on this in the next section) and would not facilitate interpretation of the parameters of interest. Moreover, point estimates and inferences drawn may depend on what combination of weather variables the researcher chooses as controls; hence, the need for a more systematic weather variable selection procedure that can deal with high dimensionality.

In this study, we consider (and select among) the most commonly used weather variables that have been utilized in past literature. The first set of weather variables we consider are degree day measures that have been used in the literature to allow for non-linearities in the effect of warming on yield outcomes (see Ritchie and Nesmith (1991), Schlenker and Roberts (2006), Schlenker and Roberts (2009). A degree day measure is typically defined as the number of degrees that the daily mean temperature is above a threshold (or is within a range) per day, accumulated over a defined period of time (say, summed over a month or over a growing season). Note that the threshold temperatures used to define particular degree day measures can vary across studies.[5] For example Ritchie and Nesmith (1991) and Deschênes and Greenstone (2007) use 8°C as the lower threshold for delineating the range of potentially yield damaging cold temperatures, and 32°C for the upper threshold to delineate the range of potentially yield damaging high temperature. Moreover, Ritchie and Nesmith (1991) discuss the possibility of using 34°C as a reasonable upper threshold as well. On the other hand, Schlenker and Roberts (2009) set 8°C for the lower threshold and 29°C for the upper threshold for corn. In their specification, Annan and Schlenker (2015) include degree day measures for moderate temperatures defined as between 10–29°C, and a degree day measure for extreme heat being above 29°C. In contrast, Lusk et al. (2018) and Tack et al. (2018) incorporated three degree day measures in their empirical specification—one for yield damaging low temperatures (with threshold

[5]Most of the recent studies follow Schlenker and Roberts (2009)'s thresholds to create specific degree day measures (e.g., degree days for moderate temperatures and degree days for yield-damaging heat). However, one needs to keep in mind that changing environmental conditions and technological advances can alter the appropriate thresholds for specific crops and locations.

range between 0–10°C), one for moderate temperatures (with threshold range between 10-29°C), and one for yield damaging high temperatures (with a threshold above 29°C).

In the present study, we first consider two degree day measures along the lines of Schlenker and Roberts (2009): (i) a growing degree day (GDD) measure consistent with moderate temperatures, and (ii) a heating degree day (HDD) measure consistent with potentially yield damaging high temperatures. The GDD measure is defined based on the range of 8-29°C, and the HDD measure coincides with a high temperature threshold of above 29°C. To account for degree day measures with alternative thresholds (Lusk et al. (2018), Tack et al. (2018)), we also utilize a "low temperature" degree day measure (defined based on temperatures between 0–10°C) and an alternative "medium temperature" degree day indicator (defined based on the 10-29°C) akin to the aforementioned GDD variable. All degree day measures used are accumulated for each month over the typical May to September corn growing season (i.e., 4 degree day measures $\times$ 5 months in the growing season = 20 monthly degree day variables). In summary, the Cluster-Lasso procedure utilized in this study chooses among four monthly degree day measures—one "low temperature" degree day measure (0–10°C), two moderate temperature degree day measures (i.e., GDD between 8-29°C, and "medium temperature" degree days in 10-29°C range), and one "high temperature" degree day measure (i.e., HDD above 29°C).

Second, we also consider a weather variable that represents accumulated precipitation (in mm) summed over each month of the May to September corn growing season. A squared precipitation term is also included to account for potential non-linearity in the precipitation effect. We primarily utilize data from the "Parameter-Elevation Regression on Independent Slopes Model" (PRISM) for calculating the GDD, HDD, low temperature degree day, medium temperature degree day, and precipitation measures used in this study.[6] PRISM is gridded 4km resolution data, which can then be aggregated to obtain county averages. The PRISM data have been widely used in previous climate change studies (for example, Schlenker and Roberts (2009)). All processed PRISM data are freely available from the Ag-Analytics open-data platform (Woodard (2016))[7].

The third set of weather variables considered in this study are a variety of moisture measures and drought indices that have been used in past studies. We include a vapor

[6]PRISM was developed by the Spatial Climate Analysis Service at Oregon State University.

[7]The Ag-Analytics platform integrates, stores, and updates daily from a variety of sources, including PRISM weather data, SSURGO soil data, POLARIS soil data, USDA NASS agricultural statistics, USDA RMA crop insurance data, USDA ERS data, and CME futures price data, among many others. See: https://ag-analytics.org/

pressure deficit (VPD) measure (in kPA) that roughly represents the relative amount of moisture in the air and is a function of air temperature and humidity (as is used in Lobell et al. (2014)).[8] Specifically, we use an average VPD measure for each month of the May to September growing season. The drought indices considered in this study for each month of the May to September growing season are the following: the Palmer Drought Severity Index (PDSI), the Palmer Hydrological Drought Index (PHDI), the Modified Palmer Drought Severity Index (PMDI), the Palmer Z Index, and the Standardized Precipitation Index (SPI). Specifically for SPI, we consider six different SPI measures that represent an average SPI measure over the past month (SP01), the past two months (SP02), the past three months (SP03), the past six months (SP06), the past nine months (SP09), and the past twelve months (SP12).

All of the drought indices are typically standardized to the local climate and measures relative dryness or wetness relative to local norms. Positive index values typically represent wetness and negative values typically indicate dryness. As such, we categorize each drought index into groups to represent the following: severe wetness, mild to moderate wetness, near normal, mild to moderate drought, and severe drought. All indices are continuous variables and reflect different levels of drought and wetness (as described in the previous sentence). The thresholds and ranges used to create index categories for each drought index measure are presented in Appendix Figure A.1. Data on the drought indices are derived from the National Climatic Data Center (NCDC). Note that the drought indices from NCDC are aggregated at the climate division level. Hence, counties in our data set falling in one climate division have the same drought index values.

Lastly, we utilize several monthly temperature-related variables such as minimum temperature, maximum temperature, and mean temperatures as the final set of weather variables considered in the study. At this point, it is important to note that we consider monthly weather variables in our main estimation procedure in this study. Much of the previous "yield-climate-change" literature (for example, Lusk et al. (2018)), only utilize yearly weather variables that coincide with the growing season rather than having monthly weather variables for each month in the season. However, there have been studies that argue that accounting of the effects of weather month-to-month is important since the timing of the weather event likely affects yield outcomes differently. For instance, Thompson (1986) found that the highest corn yields were associated with above-normal July and

[8]The VPD variable is created by following the procedure in Roberts et al. (2013), and using the PRISM data.

August rainfall, normal June temperature, and below-normal temperature in July and August. McWilliams et al. (1999) has argued that the June to August months are the critical months for corn growth. During this period, crop growth is frequently affected by weather stresses. Xu et al. (2013b) use monthly climatic variables of GDD, HDD and Palmer Z index for the months May-September, and indicate that dry early growing season—May and June—is optimal. Tannura et al. (2008) indicate that wet weather in May would delay planting and consequently affect yields at harvest. During the early part of the growing season, dry weather in May and relatively wet weather in June was found to be the most ideal for corn yields (Tannura et al. (2008)).

To summarize, we first have four degree day measures (e.g., monthly GDD, HDD, low temperature degree days and medium temperature degree days), two precipitation terms (e.g., monthly precipitation and precipitation squared), and three temperature terms (e.g., monthly minimum, maximum and mean temperatures) for each growing season month (May-September). We then have one VPD measure for each month in the May to September period. We have a total of 250 drought index variables (e.g., ten drought indices multiplied by five categories, then multiplied over five months (for each month from May to September)). Therefore, a total of 300 weather variables are included in the initial set of variables that one can choose from.[9] Having 300 weather variables to choose from as potential controls is consistent with having a high dimensional problem.

Aside from weather variables, there may also be other confounding factors that may influence inference on the risk effects of GM-RW if not properly controlled for in the specification. The level of input used (e.g., fertilizer, chemicals, etc.) may be one such factor. Therefore, we examine specifications where we control for county-level per-acre input expenditure data on the following: (i) fertilizer and chemicals, and (ii) seeds to account for some portion of management practices. Input expenditure data was collected from the the Bureau of Economic Analysis (BEA).[10] Summary statistics for all the variables used in this research are presented in Table 1 (with weather variables pertaining to the full growing season, rather than monthly). Figures that show trends of some of these variables over time are in Appendix, Figures A.2–A.6.

[9]Note that we also created growing season averages for each climate variable described here for use in several other ad-hoc fixed effects estimation of different specifications. For conciseness, the summary statistics presented in Table 1 below only reflect the growing season averages for index variables, VPD, minimum, and maximum temperature variables. Full summary statistics of all the monthly weather variables are available from the authors upon request.

[10]The expenditure data used here represents expenditures not only for corn, but for all crops in the county. The data can be accessed from: https://www.bea.gov/regional/.

Table 1: Summary statistics of variables

Variable	Mean	SD	Min	Max
LCR (county avg. loss cost ratio)	0.066	0.116	-0.000	1.185
GM-RW (county avg. adoption of GM-RW)	0.139	0.266	0.000	1.000
HDD (harmful degree days in hundred, °C)	1.355	0.897	0.059	5.156
GDD (growing degree days in thousand,°C)	1.910	0.202	1.267	2.574
Precipitation ((mm) in thousand)	0.515	0.130	0.172	1.122
VPD (vapor pressure deficit)	68.937	8.797	46.728	107.346
Low temp. degree days (in hundred, °C)	1.522	0.009	1.469	1.530
Medium temp. degree days (in thousand, °C)	1.601	0.204	0.952	2.268
Minimum temp. (monthly min. temp, °C)	14.465	1.303	9.951	19.157
Maximum temp. (monthly max. temp, °C)	26.716	1.509	22.226	31.460
Palmer-Z (Palmer Z index)	0.354	1.328	-3.450	5.158
PDSI (palmer drought severity index)	0.790	2.126	-5.100	7.364
PHDI (palmer hydrological drought index)	0.977	2.205	-5.100	7.364
PMDI (modified palmer drought severity index)	0.789	2.077	-5.080	7.364
SPx01 (standardized precip. index, 1 months)	0.101	0.498	-1.504	1.642
SPx02 (standardized precip. index, 2 months)	0.177	0.664	-2.074	2.132
SPx03 (standardized precip. index, 3 months)	0.212	0.785	-2.522	2.340
SPx06 (standardized precip. index, 6 months)	0.263	0.961	-2.704	2.570
SPx09 (standardized precip. index, 9 months)	0.300	0.907	-2.472	2.714
SPx12 (standardized precip. index, 12 months)	0.333	0.870	-1.936	2.804
Fert.&Chemical exp. (p/a $ amount spent)	0.105	0.033	0.031	0.389
Seed exp. (p/a $ amount spent)	0.046	0.024	0.007	0.223
Observations	9484	9484	9484	9484

Notes: The weather variable values for degree day measures and precipitation variables presented in this table are those that were accumulated for the corn growing months of May-September. VPD, minimum and maximum temperatures, and index variables represent growing season average.

3 Empirical Framework

3.1 Estimation Strategy

The empirical context in this study allows for the possibility of using a numerous amount of independent variables, primarily due to the availability of a large number of weather variables that can serve as controls. However, using a large number of explanatory variables in a classical ordinary least squares (OLS) approach can cause several problematic issues. First, "overfitting" can be an issue where the use of many covariates can produce good in-sample fit (through high R-squareds), but results in inaccurate out-of-sample predictions (James et al. (2013). Second, the use of a large number of independent variables typically adds a lot of "noise" to the estimation, which consequently results in higher variance for the estimated parameters of interest (and inordinately high number of insignificant variables). Third, multicollinearity issues may arise when the additional number of variables included are highly correlated and there are limited number of observations. Fourth, adding more and more explanatory variables adds to model complexity such that ease of interpretation of the estimated parameters and model parsimony are compromised.

With all these issues associated with high dimensionality, researchers typically reduce the number of controls based primarily on economic intuition and/or other ad hoc methods (e.g., using controls that have been used by similar past studies). One problem with this traditional approach is that inadvertently dropping the "wrong" control variable can result in omitted variable bias. Another concern with unstructured selection of control variables is the issue of exploring results using different control variable combinations, and only presenting those that yield the expected (or best) results (i.e., the so-called "p-hacking" problem in science). In our empirical context, this is related to the problem of getting different results when using different combinations of weather controls.

To empirically deal with high dimensional control variables, an approach that is becoming increasingly common in economics is the use of least absolute shrinkage and selection operator (Lasso) procedures that has roots in the machine learning literature (Frank and Friedman (1993), Tibshirani (1996)). The main idea with the Lasso approach is that regression coefficients are still chosen to minimize the sum of squared residuals (as is done in OLS regression), but Lasso incorporates a penalization procedure to reduce the number of explanatory variables in the specification. The penalization can be based on a number of methods such as a data-driven approach (e.g., cross validation), a theory based approach (e.g., rigorous penalization), and an information criteria approach (AIC, BIC, EBIC) (Belloni et al. (2012). In practical terms, the penalization part of Lasso chooses the coefficient estimates that are set to zero, removing the corresponding "irrelevant" variables from the model. However, the classical Lasso procedure is prone to two types of model selection errors: (i) a variable may be considered relevant when in fact it has a zero coefficient (and thus has no true explanatory power), and (ii) a variable may be dropped from the model despite truly having a nonzero coefficient (see, for example, Belloni et al. (2014a) Belloni et al. (2014b)). Over the years, several improvements to the classical Lasso procedure have been developed to partly address these model selection errors. Zou (2006), Zou and Li (2008), Zou and Zhang (2009), Fan et al. (2014) employed adaptive weights to improve model selection properties of classical Lasso. Belloni et al. (2012) suggest that a theory-driven approach to setting the penalization parameter would also help alleviate the first type of model selection error above. Given these improvements, Lasso has been primarily viewed as a type of model selection or dimension reduction statistical technique (Tibshirani (1996), James et al. (2013), Belloni et al. (2014b)).

With its roots in the machine learning literature, Lasso methods have also been his-

torically used for the purpose of prediction rather than for causal inference. As such, one of the earliest limitation of classical Lasso procedures (used purely for model selection and prediction) is that it does not have the ability to produce reliable standard errors, confidence intervals, and p-values for causal inference, which is likely the main reason for its limited use in economics in the past. To overcome this shortcoming in terms of causal inference, one simple "fix" is to first use Lasso only for the purpose of selecting the control variables to be included in the specification. Then, in a second step, an OLS regression can be run where the outcome variable is made a function of the selected variables in the first step Lasso. Belloni et al. (2013) call this two-step approach for causal inference the Post-Lasso procedure (or the post-single-selection method).

Notwithstanding Post-Lasso's ability to facilitate causal inference by producing standard errors and p-values, there are still serious flaws in using this Post-Lasso approach for causal inference. The Post-Lasso procedure tends to not select covariates with small coefficients (Belloni et al. (2014a)). This issue comes about because Lasso minimizes prediction error subject to the constraint that the model is not too "complex", where "complexity" is defined as the sum of the absolute values of all the coefficients estimated. Parameter estimates with small magnitudes (even if the associated variable is theoretically important) might be picked out (and excluded) by this constraint. That is, important independent variables with small coefficients (that theoretically should be included in the model) might look similar to other variables with small coefficients that do not belong in the model. In addition, any variable that is highly correlated to the independent variable of interest (also called the target variable(s)) (i.e., the GM-RW variable in the present study) may be dropped. This is because including such a variable will not add much predictive power for the outcome given that the target variable is already in the model (i.e., the variable(s) correlated with the target variable will likely have a small coefficient). Therefore, mistakenly excluding these kinds of variables (even if they have small coefficients) can create bias in the parameter estimates associated with the variable(s) of interest (see Leeb and Pötscher (2008a,b), Belloni et al. (2014a)). Furthermore, Belloni et al. (2014b) show that Lasso or Post-Lasso procedures in general do not produce root-n consistent and asymptotically normal estimates of the parameter of interest (typically due to large omitted variable bias).

To address the aforementioned inference issues with classical Lasso and Post-Lasso, Belloni et al. (2011), Belloni et al. (2014b), Chernozhukov et al. (2015) develop and demonstrate how a "post-double-selection" estimator can be used for valid inference. The post-

double-selection procedure has been shown to produce uniformly valid confidence intervals for the variable of interest over a wide range of potential underlying distributions for the data-generating-process. This approach is robust to the model selection errors and issues associated with classical Lasso and Post-Lasso discussed above. The post-double-selection method is conducted in three steps: (i) control variables are selected from a first-step Lasso procedure where the outcome is a function of all control variables, (ii) a second-step Lasso is conducted to select control variables that are strongly related with the variable of interest (i.e., the GM-RW variable in our context), and (iii) an OLS regression is run where the outcome variable is a function of the variable of interest and the union of the selected control variables from the first two steps. In particular, the second "selection" step is what makes the post-double-selection method robust to the model selection errors in classical Lasso and Post-Lasso. This second step also helps address the issue where Lasso tends to drop covariates that are strongly associated with the target variable, but are truly important factors that determine the outcome variable. Using the union of selected variables in the two selection steps greatly reduces the chance for omitted variables in the third step. Overall, the post-double-selection estimator produces parameter estimates associated with the target variable that are root-n consistent and asymptotically normal.[11]

Belloni et al. (2016) build on the post-double-selection approach developed in Belloni et al. (2014b), as well as the theory-driven penalization technique in Belloni et al. (2012),[12] by making these procedures applicable in panel data settings. This panel data Lasso approach is usually called the Cluster-Lasso procedure. Cluster-Lasso accounts for unobserved heterogeneity by "partialling out" the individual unit fixed effects, and also allows for clustering within individual units (i.e., allow for within-individual dependence across years). Belloni et al. (2016) show that the econometric procedures in their model result in uniformly valid inference for the parameter associated with the variable of interest, for a wide variety of distributions that potentially underlie the data generating processes (DGP).

[11] More formally, the post-double-selection estimator for the parameter of interest ($\check{\alpha}$) satisfies the following expression: $\sigma_n^{-1}\sqrt{n}\,(\check{\alpha}-\alpha_0) = i + o_P(1) \rightsquigarrow N(0,1)$.

[12] The theory-driven approach to choosing the penalization parameter (also called the tuning parameter) is a rigorous procedure that places high priority on controlling overfitting and allows for consistently deriving the theoretical properties of this particular Lasso variant.

More formally, consider the following compact panel data model:

$$y_{it} = x'_{it}\beta + \alpha_i + \varepsilon_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T \tag{1}$$

where y_{it} is the response variable for individual unit i in year t, x_{it} is the explanatory variables, β is the parameter to be estimated, α_i are the unobserved individual-unit fixed effects, and ε_{it} is the idiosyncratic error term (which is mean-zero conditional on the explanatory variables). In the first step of the Cluster-Lasso method, unobserved individual specific heterogeneity, which is treated as the fixed effects in equation (1), are "partialled out" of the model by time demeaning the dependent and independent variables, which results in a time-demeaned specification as follows:

$$\ddot{y}_{it} = \ddot{x}'_{it}\beta + \ddot{\varepsilon}_{it}, \tag{2}$$

where $\ddot{y}$ and $\ddot{x}$ denotes the time-demeaned dependent and independent variables, β is still the parameter to be estimated, and $\ddot{\varepsilon}$ is the time-demeaned error term. The time-demeaned specification in (2) is then used in a Lasso optimization to estimate the parameters of interest as follows:

$$\hat{\beta}_{CL} = \arg\min_{\beta} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \left(\ddot{y}_{it} - \ddot{x}'_{it}\beta\right)^2 + \frac{\lambda}{nT} \sum_{j=1}^{p} \hat{\phi}_j \left|\beta_j\right| \tag{3}$$

where λ is the penalization parameter and $\left\{\hat{\phi}_j\right\}_{j=1}^{p}$ are the clustered-penalty loadings set as:

$$\lambda = 2c\sqrt{n}\Phi^{-1}(1 - \gamma/(2p)) \tag{4}$$

where $\gamma = o(1)$, and

$$\hat{\phi}_j^2 = \frac{1}{nT} \sum_{i=1}^{n} \left(\sum_{t=1}^{T} \ddot{x}_{itj}\hat{\varepsilon}_{it} \right)^2 = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{t'=1}^{T} \ddot{x}_{itj}\ddot{x}_{it'j}\hat{\varepsilon}_{it}\hat{\varepsilon}_{it'} \tag{5}$$

In contrast to the traditional Lasso formulation, within-individual dependence (across years) is accounted for in the Cluster-Lasso method primarily through the cluster penalty loadings defined in equation (5) and used in the Lasso optimization along with (4) in equation (3).[13] Moreover, simulations performed by Belloni et al. (2016) show that the

[13] It is important to note here that the penalty loadings for Cluster-Lasso are constructed after the

Cluster-Lasso procedure perform better (i.e., lower bias and RMSE) than variable selection procedures that do not allow for clustering. The authors indicate that this difference in performance suggests that additional modifications of Lasso-type procedures to account for other dependence structures may be worthwhile.

The procedures embodied in equations (2) to (5) are utilized in the first two steps of the post-double-selection method used in this study. In the final step, estimation and inference proceed by OLS regression of $\ddot{y}_{it}$ as a function of $\ddot{x}_{it}$, where $\ddot{x}_{it}$ includes the variable of interest and the union of the set of control variables selected in the first two steps.[14]

3.2 Empirical Specification and Identification Issues

We use the Cluster-Lasso procedure described above to estimate the following empirical specification:

$$LCR_{it} = \beta_1 \mathrm{RW}_{it} + W'_{it}\beta_w + \alpha_i + \mu_t + \varepsilon_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T \tag{6}$$

where LCR_{it} is the lost cost ratio for county i in year t, RW_{it} is the county-level GM-RW adoption rate, W_{it} represents a vector of control variables that include climate variables (such as those discussed in the previous section) and input expenditure variables (e.g., seed, and fertilizer and chemicals), α_i is the unobserved county fixed effects that controls for time-invariant county-level factors that impact LCR (i.e., unobserved heterogeneity), μ_t denotes year fixed effects, and ε_{it} is the idiosyncratic error term for county i in year t (which is mean zero conditional on covariates, and within-individual dependence is allowed). The term β_1 is our parameter of interest given the study objective of determining the yield risk effects of GM corn with rootworm resistant traits.

To operationalize the Cluster-Lasso procedure for the empirical specification in equation (6), the time-demeaning step needs to be implemented first in order to "partial out"

time-demeaning step. See Appendix in Belloni et al. (2016) for details on the algorithm to calculate $\hat{\varepsilon}_{it}$ and $\left\{\hat{\phi}_j\right\}_{j=1}^{p}$. Also, one important condition in proving the favorable performance of the proposed model is: $\frac{\lambda\hat{\phi}_j}{nT} \geqslant 2c\left|\frac{1}{nT}\sum_{l=1}^{n}\sum_{t=1}^{T}\ddot{x}_{itj}\ddot{\epsilon}_{it}\right|$ for each $1 \leqslant j \leqslant p$ holds with probability one under the condition $l\phi_j \leqslant \hat{\phi}_j \leqslant u\phi_j$, with probability $1 - o(1)$ for some $\ell \to 1$ and $u \leqslant C < \infty$, and the λ set as in (4)

[14] Belloni et al. (2016) call this last step Post-Cluster-Lasso or post-double-selection using Cluster-Lasso for the first two steps.

the county-level fixed effects:

$$L\ddot{C}R_{it} = \beta_1 \ddot{\mathrm{RW}}_{it} + \ddot{W}'_{it}\beta_w + \mu_t + \ddot{\varepsilon}_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T. \tag{7}$$

In equation (7), note that the time fixed effects are still included in the specification to control for yearly time shocks that affects all the counties similarly. Hence, these time fixed effects are not penalized in the estimation procedure.

After time-demeaning, the Cluster-Lasso procedure is used to estimate a function with $L\ddot{C}R_{it}$ as the dependent variable, and $\ddot{\mathrm{RW}}_{it}$ and $\ddot{\mathrm{W}}_{it}$ as the explanatory variables. The next step is to implement Cluster-Lasso again, but now estimating a function where $\ddot{\mathrm{RW}}_{it}$ is the dependent variable and $\ddot{\mathrm{W}}_{it}$ as the independent variables. These two steps allow for dimension-reduction to select the climate variables (as well as other covariates) to leave in the specification. In particular, the union of the selected controls in the two Cluster-Lasso estimations will then be used in a traditional OLS regression of equation (7). This results in a more parsimonious model since it will only include selected control variables, and allow inference on the variable of interest $\ddot{\mathrm{RW}}_{it}$. Note that we use cluster-robust standard errors (Arellano (1987)) in the the final OLS step in addition to cluster-robust penalty loadings in the Cluster-Lasso.

Causal identification in our analysis relies primarily on the control of unobserved heterogeneity (e.g., unobserved time-invariant variables that affect both the adoption of GM-RW and the LCR outcome) through the time-demeaning transformation. We argue that unobserved management ability (in this case at the county-level) is one of the "main" unobservables that can be correlated with GM-RW usage and the LCR outcome, which may then cause identification issues. However, unobserved ability can be reasonably assumed as time-invariant and so the time-demeaning procedure that partials out the county-fixed effects take care of this issue. In addition, heterogeneous soil quality that varies across counties can also be reasonably assumed as time-invariant and absorbed by the county fixed effects since soil quality measure does not change a lot from year to year (i.e., it takes a long time to have substantial changes in soil quality). The geographical focus on the Central Corn Belt states also mitigates this issue since soils in this region should be fairly similar. Another potential source of endogeneity might be the year-to-year incidence and degree of corn rootworm infestation. Unfortunately, we do not have rootworm infestation data to use as controls (or valid instruments for it) but we believe that infestation levels across space (i.e., across the "I" states) in our data is fairly homogeneous. As such,

assuming corn rootworm infestation only varies considerably across time, then year fixed effects will soak up this potential confounding variation. In addition, year fixed effects in our model capture year-to-year changes in the federal insurance program that affects all counties in the sample similarly,[15] as well as overall price changes.

4 Results and Discussion

4.1 Estimation Results

Before we discuss estimation results from the main Cluster-Lasso procedure, we first present the estimation results using traditional panel data OLS fixed effects estimation techniques where the climate variables are chosen "ad hoc" based on what has been done in previous studies (see Table 2). As seen in Table 2, there are substantial differences in the magnitudes and the statistical significance of the GM-RW variable, depending on the set of climate variables included in the specification. For example, when only growing season GDD, HDD, precipitation, precipitation squared, and input expenditures (fertilizer and seed expenditures) are used as controls (Model 1), results from the traditional panel fixed effects model (with county and time fixed effects) suggests that GM-RW does not have a statistically significant effect on LCR (i.e., GM-RW has no risk effects). In contrast, when the specification includes growing season minimum temperatures, maximum temperatures, precipitation, and precipitation squared as control climate variables (together with input expenditures), the estimates indicate that GM-RW has a strong statistically significant negative effect on LCR, which indicates a risk reducing effect (Model 2). Several other specifications are used in Table 2 and inference on the risk effects of GM-RW varies depending on what weather variables are utilized in the specification. In Appendix, we also show that using different sets of degree day measures with varying thresholds also substantially change the inference drawn from the parameter estimate of GM-RW (Table A.1). Alternative combinations of drought indices also considerably alter inferences drawn from the GM-RW variable (Table A.2).

Given the inconsistent inferences on the GM-RW variable when one uses traditional panel data econometric techniques and different sets of climate variables, we use Cluster-

[15] For example, the 1994 Crop Insurance Reform Act (CIRA) increased premium subsidy rates to an average of about 40 percent and the 2000 Agricultural Risk Protection Act (ARPA) increased premium subsidy rates to an average of 62 percent. Since then, crop insurance participation expanded, and farmers enrolled more in the program. Since these types of program changes affects all counties, then the year fixed effects account for this.

Table 2: Estimates of the effect of GM-RW on LCR using traditional panel data methods

	Model 1	Model 2	Model 3	Model 4
GM-RW	-0.015	-0.030***	-0.017*	-0.014
	(-1.84)	(-3.51)	(-2.03)	(-1.73)
HDD	0.128***		0.120***	0.137***
	(20.86)		(19.20)	(19.70)
GDD	-0.272***		-0.256***	-0.253***
	(-9.01)		(-8.57)	(-8.23)
Precipitation	-0.556***	-0.762***		-0.587***
	(-9.19)	(-12.18)		(-9.27)
Precipitation sq.	0.685***	0.786***		0.699***
	(13.68)	(15.38)		(13.67)
VPD				-0.001**
				(-2.69)
Minimum temp		-0.000		
		(-0.05)		
Maximum temp		0.023***		
		(7.19)		
Palmer-Z			0.007***	
			(4.41)	
Palmer-Z sq.			0.009***	
			(13.67)	
Fertilizer exp	-0.177	-0.212*	-0.188	-0.184
	(-1.76)	(-2.15)	(-1.88)	(-1.85)
Seed exp	0.634**	0.669***	0.653***	0.623**
	(3.26)	(3.58)	(3.35)	(3.22)
County Fixed Effects	Y	Y	Y	Y
Year Fixed Effects	Y	Y	Y	Y
Observations	9484	9484	9484	9484
Adjusted R^2	0.502	0.445	0.494	0.503
AIC	-21718.571	-20678.507	-21564.467	-21725.128
BIC	-21425.120	-20385.056	-21271.015	-21424.519

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table presents estimates of the effect of genetically modified varieties with corn rootworm traits (GM-RW) on the loss cost ratio (LCR) for a panel of 293 U.S. Central Corn Belt (CCB) counties over the years 1981–2015. LCR is considered for the total of the two biggest insurance plans: Yield Protection and Revenue Protection. Each column employs the OLS fixed effects, regressing LCR on GM-RW with different covariates as shown in the table with *county and time fixed effects*.

Lasso to have a more structured approach to select among a large set of monthly weather variables (for parsimony) and arguably get more robust inference results. Results from the Cluster-Lasso estimation indicates that GM-RW statistically reduces LCR at the 5% level of significance, which is evidence of a risk reducing effect (see Table 3). The magnitude of the GM-RW parameter estimate suggests that one percentage point increase in GM-RW adoption decreases LCR by 0.00017, on average. For the period of time studied in this research, the average LCR in the CCB is 0.066. This means that a percentage point increase in the adoption of GM-RW decreases LCR by about 0.26%. Because the U.S. federal crop insurance program carries around $100 billion in liabilities annually, 0.26% decrease in LCR might still provide important insights for policy makers. The result from the Cluster-Lasso procedure provides evidence on the potential risk-reducing effect of GM corn varieties with the rootworm resistant trait after controlling for climate variables.

Table 3: Estimates of the effect of GM-RW on LCR using the cluster-lasso procedure

	Estimator
	Cluster-Lasso Post-Double Selection
GM-RW	-0.017*
	(-2.50)
Observations	9484

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table presents estimates of the effect of genetically modified varieties with corn rootworm trait (GM-RW) on the loss cost ratio (LCR) for a panel of 293 U.S. Central Corn Belt (CCB) counties over the years 1981–2015. LCR is considered for the total of the two biggest insurance plans: Yield Protection and Revenue Protection. LCR is regressed on GM-RW and additional covariates (that were selected through Cluster-Lasso as described in the text). *The specification includes a full set of year and county fixed effects and selected covariates.* In the specification, standard errors are clustered by county and they are robust to heteroskedastic and clustered error structure.

For the main Cluster-Lasso run in Table 3, the weather variables eventually selected by the procedure are presented in Appendix (Table A.3). In general, the selected weather controls include several degree day measures across different months (HDD, GDD, low and medium temperature degree days), minimum temperatures in the early months, and a good number of drought indices for different months. Note that precipitation terms are not among the selected variables. It is likely that the preponderance of the drought indices selected (for different dryness and wetness categories across months) already accounts for precipitation variables and some of the maximum temperature effects. Nonlinearity in the weather effects are also likely accounted for through the different degree days and wetness/dryness indices selected.

To augment the preferred model results in Table 3, we also explore whether inference on the GM-RW changes based on the number of weather variables to initially choose from. The Cluster-Lasso results in Table 3 is based on an initial set of 300 weather variables that the procedure can choose from. Hence, we first implement a Cluster-Lasso on a specification that does not include the 250 drought index variables. That is, the initial set of weather variables to choose from only include 50 monthly degree day measures, precipitation terms, VPD, and temperature variables (i.e., 10 weather terms for every month in the 5 month growing season). Results from this Cluster-Lasso run supports the risk reducing effect of GM-RW, with the magnitude of the effect similar to the one in Table 3 (see Table A.4 in Appendix).

A second Cluster-Lasso run is then implemented for the case where only yearly growing-season weather variables are the set to choose from. In this case, the weather variable

choice set include 60 variables in total (e.g., 50 variables for the 10 drought indexes having 5 categories, and 10 variables for degree days, precipitation terms, VPD, and temperature variables). Again, results (see Table A.5 in Appendix) from this Cluster-Lasso run still provides evidence of a risk-reducing effect and magnitude that is similar to the other cluster-runs above.

4.2 Robustness Checks

In the previous sub-section, we already provided some evidence on the robustness of our estimation approach when we examine whether inferences on the GM-RW variable changes when we change the weather variable choice set. Below we conduct further robustness checks to verify the strength and stability of our Cluster-Lasso results.

In the first robustness check, we consider a specification where we include crop insurance related variables in the control variable choice set. Since the outcome variable is a crop insurance measure, it can be argued that it is important to account for crop insurance related variables in the estimation.[16] Specifically, we consider county-level crop insurance participation rates and average county-level coverage level choice as potential control variables.[17] Results from this first robustness check is presented in the Model 2 column of Table A.6 (see Appendix).[18] Inference from the Cluster-Lasso procedure that considers crop insurance variables still supports findings from the earlier Cluster-Lasso runs where GM-RW statistically reduces LCR.

The second robustness check conducted is where we exclude the input expenditure variables from the control variable choice set that was used in Table 3 above. This is to explore whether only including the complete set of 300 weather related control variables (and not including input expenditure variables) will result in similar inferences. As seen in Table A.6 (Model 3), the estimated parameter associated with the GM-RW variable is consistent with the Cluster-Lasso runs that includes input expenditure variables as part of the control variables to choose from.

Lastly, we perform a robustness check where we only consider LCR for the YP policy

[16] The idea is to sharpen identification and avoid omitted variable bias by including these crop insurance variables. Though if county-level participation and coverage level choices are roughly homogeneous across the three "I" States from year-to-year, then one can argue that the time fixed effects can account for not including these variables in the specification. The disadvantage of including these variables in the specification is that some researchers may argue that adding these confounders may be endogeneous in and of itself.

[17] We calculate the insurance participation as the ratio of insured acres of corn to total planted acres of corn, which is also called an empirical insurance participation measure. Note that the data for coverage level is only available after 1988 (and so runs with this variable only includes data from 1989 onwards).

[18] Note that the Model 1 column is the same as that for our preferred model in Table 3.

(i.e., excluding the data from the RP policies) (see Table A.7 in Appendix). The idea is that only a YP-type insurance plan existed for the early part of our data (1980s) and so we would like to see if the inferences from the Cluster-Lasso procedure will change if only the losses from these YP plans are considered. However, note that RP-type plans came about in the mid-1990s and has become the predominantly chosen plan for corn since the mid-2000s. Nevertheless, using Cluster-Lasso only for the YP losses (i.e., and utilizing the same 300 weather variables plus other controls in Table 3) still produces the same inference where counties with higher levels of GM-RW adoption tend to have lower risk (as measured by LCR).

5 Conclusions

This study aims to explore how GM corn varieties with rootworm resistant traits (GM-RW) affect yield risk while accounting for a large number of potential climate variables that can serve as controls (e.g., the high dimensionality problem). Yield risk in this case is measured using a loss cost ratio (LCR) variable from crop insurance. A county-level panel data set was constructed for Illinois, Indiana, and Iowa covering the period 1981 to 2015. To accomplish the underlying aim of this study, the Cluster-Lasso procedure is used to handle the high dimensional nature of the climate variables, as well as the panel data structure of the county-level data utilized.

Our findings consistently indicate that counties with high levels of GM-RW adoption tend to have lower LCRs when Cluster-Lasso is used to deal with high dimensional weather control variables. This is evidence that GM-RW adoption in the Central Corn Belt has reduced yield risk associated with corn production. These results are consistent with some of the earlier studies that have shown that GM crops tend to reduce yield risk (Shi et al. (2013), Goodwin and Piggott (2019)). In addition, this risk reducing finding for GM-RW generally supports the idea behind the Biotech Yield Endorsement implemented in crop insurance from 2008 to 2012 where premium discounts are given to farmers adopting these type of GM varieties. Along these lines, the main empirical insight from this study may be useful when there are concerns about adverse selection in the crop insurance program, and there is a need to find farm level indicators that better delineate lower-risk versus higher-risk producers.

Moreover, the implementation of the Cluster-Lasso procedure in this study demonstrates its potential as a viable econometric approach when researchers are faced with

the challenge of choosing among a large number of control variables in estimation. High-dimensional regressors could arise for different reasons in agricultural economics and other disciplines (e.g., a large set of potential weather confounders in our case). Hence, the Cluster-Lasso procedure utilized in this study is a potential new tool for researchers to use when they are unsure about the exact set of control variables to include in a model and there are a large number of controls to choose from. In existing studies that have examined yield effects of various agricultural technologies and practices (e.g., precision technologies, fertilization practices, soil health practices, etc.), especially where weather variables need to serve as controls to help assure identification, researchers normally just use "ad hoc" methods to maintain parsimony in the specification (e.g., following what has been done in past studies, or based on some agronomic logic). We show that "ad hoc" choice of weather controls can substantially influence the eventual inferences drawn on the variable of interest (i.e., GM-RW in our case). In this case, the Cluster-Lasso procedure can serve as a more structured alternative for researchers and economists seeking to estimate panel data production functions with high dimensional weather variables as controls. It can complement intuition based choices and can provide uniformly valid inference in the presence of unobserved individual specific heterogeneity under high-dimensional panel data settings.

Notwithstanding the contributions of this article to empirical understanding of the risk effects of GM crops and in highlighting the potential of Cluster-Lasso as a means to deal with high dimensionality, it is important to recognize the limitations of the study and mention promising opportunities for future research. First, although we strived to avoid potential omitted variable bias in the estimation and assure proper identification of the GM-RW effect, it is possible that there are still time-varying and county-varying unobservables that cause endogeneity in the GM-RW variable. For instance, if rootworm infestation levels vary substantially across time and space, then not having this variable in the specification may affect inferences from our estimation. Even if there are Lasso procedures that allow for instrumental variables to address the endogeneity described above (see Chernozhukov et al. (2015)), we do not have valid instruments to implement this procedure. We leave this potential extension for future research. Second, we only utilize county-level data for three corn belt states in the analysis. Future research may explore the use of individual farm-level data and evaluate the risk effects of GM-RW for a wider geographical area. Lastly, the focus in this paper is specifically on corn production

and the rootworm trait. A such, future studies may consider the yield risk effects of various GM traits for other crops (like soybeans).

References

Annan, F. and Schlenker, W. (2015). Federal crop insurance and the disincentive to adapt to extreme heat. *American Economic Review*, 105(5):262–66.

Arellano, M. (1987). Practitioners'corner: Computing robust standard errors for within-groups estimators. *Oxford bulletin of Economics and Statistics*, 49(4):431–434.

Armal, S., Devineni, N., and Khanbilvardi, R. (2018). Trends in extreme rainfall frequency in the contiguous united states: Attribution to climate change and climate variability modes. *Journal of Climate*, 31(1):369–385.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Belloni, A., Chernozhukov, V., et al. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Belloni, A., Chernozhukov, V., and Hansen, C. (2011). Inference for high-dimensional sparse econometric models. *arXiv preprint arXiv:1201.0220*.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014b). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Belloni, A., Chernozhukov, V., Hansen, C., and Kozbur, D. (2016). Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics*, 34(4):590–605.

Chavas, J.-P. and Shi, G. (2015). An economic analysis of risk, management, and agricultural technology. *Journal of agricultural and resource economics*, pages 63–79.

Chavas, J.-P., Shi, G., and Lauer, J. (2014). The effects of gm technology on maize yield. *Crop Science*, 54(4):1331–1335.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review*, 105(5):486–90.

Deschênes, O. and Greenstone, M. (2007). The economic impacts of climate change: evidence from agricultural output and random fluctuations in weather. *American Economic Review*, 97(1):354–385.

Duke, S. O. (2015). Perspectives on transgenic, herbicide-resistant crops in the united states almost 20 years after introduction. *Pest management science*, 71(5):652–657.

Fan, J., Xue, L., and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of statistics*, 42(3):819.

Fernandez Cornejo, J., Wechsler, S., Livingston, M. J., and Mitchell, L. (2014). Genetically engineered crops in the united states. *USDA-ERS Economic Research Report*, (162).

Foley, J. (2014). Gmos, silver bullets and the trap of reductionist thinking. *Ensia, February*, 25.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M., and Toulmin, C. (2010). Food security: The challenge of feeding 9 billion people. *Science*, 327(5967):812–818.

Goodwin, B. K. and Piggott, N. E. (2019). Has technology increased yield risk? evidence from the crop insurance biotech endorsement. *Forthcoming in American Journal of Agricultural Economics.*

Gurian-Sherman, D. (2009). *Failure to yield: Evaluating the performance of genetically engineered crops.* Union of Concerned Scientists Cambridge, MA.

Hakim, D. (2016). Doubts about the promised bounty of genetically modified crops. *New York Times*, 29.

James, C. (2016). Global status of commercialized biotech/gm crops:2016. *ISSA Briefs*, (52).

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Leeb, H. and Pötscher, B. M. (2008a). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376.

Leeb, H. and Pötscher, B. M. (2008b). Guest editors'editorial: Recent developments in model selection and related areas. *Econometric Theory*, 24(2):319–322.

Lobell, D. B., Roberts, M. J., Schlenker, W., Braun, N., Little, B. B., Rejesus, R. M., and Hammer, G. L. (2014). Greater sensitivity to drought accompanies maize yield increase in the us midwest. *Science*, 344(6183):516–519.

Lusk, J. L., Tack, J., and Hendricks, N. P. (2018). Heterogeneous yield impacts from adoption of genetically engineered corn and the importance of controlling for weather. In *Agricultural Productivity and Producer Behavior*. University of Chicago Press.

Ma, B., Meloche, F., and Wei, L. (2009). Agronomic assessment of bt trait and seed or soil-applied insecticides on the control of corn rootworm and yield. *Field Crops Research*, 111(3):189–196.

McKenzie, F. and Williams, J. S. (2015). Sustainable food production: constraints, challenges and choices by 2050. *Food Security*, 7:221–233.

McWilliams, D. A., Berglund, D. R., and Endres, G. (1999). Corn growth and management quick guide.

Moschini, G. (2008). Biotechnology and the development of food markets: retrospect and prospects. *European Review of Agricultural Economics*, 35(3):331–355.

Nolan, E. and Santos, P. (2012). The Contribution of Genetic Modification to Changes in Corn Yield in the United States. *American Journal of Agricultural Economics*, 94(5):1171–1188.

Ritchie, J. T. and Nesmith, D. S. (1991). Temperature and crop development. *Modeling plant and soil systems*, (modelingplantan):5–29.

Roberts, M. J., Schlenker, W., and Eyer, J. (2013). Agronomic weather measures in econometric models of crop yield with implications for climate change. *American Journal of Agricultural Economics*, 95(2):236–243.

Sanglestsawai, S., Rodriguez, D. G. P., Rejesus, R. M., and Yorobe, J. M. (2017). Production risk, farmer welfare, and bt corn in the philippines. *Agricultural and Resource Economics Review*, 46(3):507–528.

Schlenker, W. and Roberts, M. J. (2006). Nonlinear effects of weather on corn yields. *Review of agricultural economics*, 28(3):391–398.

Schlenker, W. and Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to us crop yields under climate change. *Proceedings of the National Academy of sciences*, 106(37):15594–15598.

Shi, G., Chavas, J.-P., and Lauer, J. (2013). Commercialized transgenic traits, maize productivity and yield risk. *Nature biotechnology*, 31(2):111.

Tack, J., Coble, K., and Barnett, B. (2018). Warming temperatures will likely induce higher premium rates and government outlays for the us crop insurance program. *Agricultural economics*, 49(5):635–647.

Tannura, M. A., Irwin, S. H., and Good, D. L. (2008). Weather, technology, and corn and soybean yields in the u.s. corn belt. *https://ssrn.com/abstract=1147803*.

Thompson, L. M. (1986). Climatic change, weather variability, and corn production. *Agronomy Journal*, 78(4):649–653.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Urban, D., Roberts, M. J., Schlenker, W., and Lobell, D. B. (2012). Projected temperature changes indicate significant increase in interannual variability of us maize yields. *Climatic change*, 112(2):525–533.

Urban, D. W., Sheffield, J., and Lobell, D. B. (2015). The impacts of future climate and carbon dioxide changes on the average and variability of us maize yields under two emission scenarios. *Environmental Research Letters*, 10(4):045003.

Wheeler, T. and von Braun, J. (2013). Climate change impacts on global food security. *Science*, 341(6145):508–513.

Woodard, J. (2016). Big data and ag-analytics: An open source, open data platform for agricultural & environmental finance, insurance, and risk. *Agricultural Finance Review*, 76(1):15–26.

Xu, Z., Hennessy, D. A., Sardana, K., and Moschini, G. (2013a). The realized yield effect of genetically engineered crops: U.s. maize and soybean.

Xu, Z., Hennessy, D. A., Sardana, K., and Moschini, G. (2013b). The realized yield effect of genetically engineered crops: Us maize and soybean. *Crop Science*, 53(3):735–745.

Yu, T. and Babcock, B. A. (2010). Are us corn and soybeans becoming more drought tolerant? *American Journal of Agricultural Economics*, 92(5):1310–1323.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37(4):1733.

Appendix

Degree days and Palmer indices

Our results also indicate that using different thresholds for the different degree day measures also change the statistical significance of the GM-RW variable. If cooling degree days and heating degree days are used instead of HDD and GDD, then the inference on GM-RW may change. In its simplest form, growing degree days are calculated in the same way as cooling degree days, but the threshold temperatures are different. Heating degree days (HeDD) is calculated as the average temperature on a day subtracted from 65°F and cooling degree days (CDD) is calculated as the average temperature minus 65°F. Both degree days come from from NCDC-NOAA. Additionally, we also show that results vary if low-temperature degree days (0–10°C), medium temperature degree days (10–29°C), and high temperature degree days (above 29°C) are employed in the specification instead of HDD, GDD, HeDD and CDD. Table A.1 below shows the results for these specifications. Although it is trivial to select the most suitable set of degree day measures via our model selection approach (cluster-lasso), our point here is to show that inferences from traditional panel OLS methods are sensitive to the selection of degree day measures used.[19]

Further, note that some researchers prefer include Palmer Z drought index as water stress variable without separately breaking out the wetness and dryness categories (e.g., nonlinearity in the simplest definition).[20] In Table A.2, we also show that there are differences in the magnitudes and the statistical significance of the GM-RW variable based on drought indices (say Palmer Z and PDSI) in the econometric specification. Significance of GM-RW changes based on what index variable is used (PDSI vs Palmer Z) and how it is incorporated.

[19] One needs to keep in mind that the changing environmental conditions and technological advances might change the different thresholds used to determine the degree day measures. Hence, we wanted to show robustness of the cluster-lasso approach when choosing among degree day measures with different thresholds.

[20] Note that in our main model specification, we further consider more flexibility of index variables by incorporating different specifications for wetness and dryness (i.e., severe, moderate)

Table A.1: Estimates of the effect of GM-RW on LCR: Alternative degree day thresholds

	Model 1	Model 2	Model 3
GM-RW	-0.015	-0.021*	-0.003
	(-1.84)	(-2.48)	(-0.39)
HDD	0.128***		
	(20.86)		
GDD	-0.272***		
	(-9.01)		
Precipitation	-0.556***	-0.751***	-0.793***
	(-9.19)	(-13.05)	(-13.94)
Precipitation sq	0.685***	0.771***	0.800***
	(13.68)	(15.71)	(16.44)
Low temp DD			-0.673***
			(-4.82)
Medium temp DD			0.001
			(0.05)
High temp			1.286***
			(13.91)
HeDD		0.033***	
		(7.75)	
CDD		0.344***	
		(14.90)	
Fertilizer exp	-0.177	-0.243*	-0.194*
	(-1.76)	(-2.51)	(-2.05)
Seed exp	0.634**	0.657***	0.602***
	(3.26)	(3.68)	(3.40)
County Fixed Effects	Y	Y	Y
Year Fixed Effects	Y	Y	Y
Observations	9484	9484	9484
Adjusted R^2	0.502	0.455	0.529
AIC	-21718.571	-20855.315	-22248.948
BIC	-21425.120	-20561.863	-21948.339

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table presents estimates of the effect of genetically modified varieties with corn rootworm traits (GM-RW) on the loss cost ratio (LCR) for a panel of 293 U.S. Central Corn Belt (CCB) counties over the years 1981–2015. LCR is considered for the total of the two biggest insurance plans: Yield Protection and Revenue Protection. Each column employs the OLS fixed effects, regressing LCR on GM-RW with different covariates as shown in the table with *county and time fixed effects.*

Table A.2: Estimates of the effect of GM-RW on LCR: Alternative drought indices

	Model 1	Model 2	Model 3	Model 4	Model 5
GM-RW	-0.015	-0.017*	-0.019*	-0.014	-0.013
	(-1.84)	(-2.03)	(-2.45)	(-1.69)	(-1.57)
Palmer-Z	0.019***	0.007***			
	(13.57)	(4.41)			
Palmer-Z sq		0.009***			
		(13.67)			
Wet-Palmer-Z			0.043***		
			(22.93)		
Dry-Palmer-Z			0.040***		
			(14.56)		
PDSI				0.009***	0.003***
				(12.90)	(5.55)
PDSI sq					0.003***
					(16.33)
Fertilizer exp.	-0.175	-0.188	-0.092	-0.167	-0.175
	(-1.77)	(-1.88)	(-0.98)	(-1.73)	(-1.79)
Seed exp.	0.565**	0.653***	0.574**	0.552**	0.578**
	(3.09)	(3.35)	(3.15)	(3.10)	(3.10)
County Fixed E.	Y	Y	Y	Y	
Year Fixed E.	Y	Y	Y	Y	
Observations	9484	9484	9484	9484	9484
Adjusted R^2	0.470	0.494	0.503	0.465	0.484
AIC	-21127.338	-21564.467	-21729.355	-21036.245	-21368.202
BIC	-20841.044	-21271.015	-21435.903	-20749.950	-21074.750

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table presents estimates of the effect of genetically modified varieties with corn rootworm traits (GM-RW) on the loss cost ratio (LCR) for a panel of 293 U.S. Central Corn Belt (CCB) counties over the years 1981–2015. LCR is considered for the total of the two biggest insurance plans: Yield Protection and Revenue Protection. Each column employs the OLS fixed effects, regressing LCR on GM-RW with different covariates as shown in the table with *county and time fixed effects.*

Table A.3: Selected Variables through Cluster-Lasso

tmin6 tmin9 HDD5 VPD5 low_temp5 HDD6 VPD6 med_temp6 GDD7 HDD7 med_temp7 HDD8 low_temp9 may_zndx_wet_s may_zndx_n jul_zndx_wet_s jul_zndx_dry_s sep_zndx_n may_phdi_wet_s aug_phdi_dry_s sep_phdi_n sep_phdi_dry_m may_pmdi_n jun_pmdi_wet_m sep_pmdi_wet_m may_sp01_wet_s may_sp01_wet_m jun_sp01_wet_m jun_sp01_dry_m jun_sp01_dry_s sep_sp01_dry_m may_sp02_n jun_sp02_wet_m jun_sp02_n jun_sp02_dry_s jul_sp02_wet_s jul_sp02_dry_m aug_sp02_wet_s aug_sp02_n jun_sp03_wet_s jun_sp03_wet_m jun_sp03_dry_s aug_sp03_wet_s aug_sp03_n aug_sp03_dry_s jun_sp06_wet_m jul_sp06_dry_s aug_sp06_wet_s may_sp09_n aug_sp09_wet_m aug_sp09_dry_m jul_sp12_wet_s aug_sp12_wet_m sep_sp12_wet_m sep_sp12_dry_m seed_exp

Table A.4: Estimates of the effect of GM-RW on LCR: 50 monthly weather variables

	Estimator
	Cluster-Lasso Post-Double Selection
GM-RW	-0.020**
	(-2.64)
Observations	9484

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table presents estimates of the effect of genetically modified varieties with corn rootworm trait (GM-RW) on the loss cost ratio (LCR) and employs the Cluster-Lasso method as used in the main result, Table 3. Note that instead of full 300 weather variables, *only monthly weather variables without drought indices*, which are 50 in total are used to select. LCR is regressed on GM-RW and additional covariates selected among 50 weather and two expenditure variables. *The specification includes a full set of year and county fixed effects and selected covariates.* In the specification, standard errors clustered by county and they are robust to heteroskedastic and clustered error structure.

Table A.5: Estimates of the effect of GM-RW on LCR: 60 yearly weather variables

	Estimator
	Cluster-Lasso Post-Double Selection
GM-RW	-0.016*
	(-2.12)
Observations	9484

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table presents estimates of the effect of genetically modified varieties with corn rootworm trait (GM-RW) on the loss cost ratio (LCR) and employs the Cluster-Lasso method as used in the main result, Table 3. Note that instead of full 300 weather variables, *only growing season weather with growing season index weather variables*, which are 60 in total are used to select. LCR is regressed on GM-RW and additional covariates selected among 60 weather and two expenditure variables. *The specification includes a full set of year and county fixed effects and selected covariates.* In the specification, standard errors clustered by county and they are robust to heteroskedastic and clustered error structure.

Table A.6: Estimates of the effect of GM-RW on Loss Cost Ratio: Adding crop insurance related variables and excluding input expenditure variables

	Model 1	Model 2	Model 3
GM-RW	-0.017*	-0.014*	-0.015*
	(-2.50)	(-2.08)	(-2.32)
Observations	9484	9484	9484

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table presents estimates of the effect of genetically modified varieties with corn rootworm trait (GM-RW) on the loss cost ratio (LCR) for a panel of 293 U.S. Central Corn Belt (CCB) counties over the years 1981–2015. LCR is considered for the total of the two biggest insurance plans: Yield Protection and Revenue Protection. *Each column employs the same Cluster-Lasso methods with slightly different set of variables.* In each specification, LCR is regressed on GM-RW and additional covariates (selected variables through the Cluster-Lasso as described in the text). The first column provides our main results with the same set of variables used in Table 3. The second column adds insurance variables to the specification in the first column. The third column excludes the input expenditure variables from the first column. *Each specification includes a full set of year and county fixed effects.* In each specification, standard errors clustered by county and they are robust to heteroskedastic and clustered error structure.

Table A.7: Estimates of the effect of GM-RW on Loss Cost Ratio only for YP

	Cluster-Lasso Post-Double-Selection
	Yield Protection
GM-RW	-0.017*
	(-2.38)
Observations	9484
9484	

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table presents estimates of the effect of genetically modified varieties with corn rootworm trait (GM-RW) on the loss cost ratio (LCR) for a panel of 293 U.S. Central Corn Belt (CCB) counties over the years 1981–2015. *LCR is considered for only Yield Protection.* LCR is regressed on GM-RW and additional covariates same as in the Table 3 as described in the text. *The specification includes a full set of year and county fixed effects and selected covariates.* In the specification, standard errors clustered by county and they are robust to heteroskedastic and clustered error structure.

Figure A.1: Thresholds used for the drought index variables used in the study

Weather Indices Range

PHDI, PDSI, PMDI		Palmer-Z
Range	Category	Range
> 3.00	Severe wetness	> 2.50
1.50, 2.99	Mild to moderate wetness	1.00, 2.49
-1.49, 1.49	Near normal	-1.24, 0.99
-1.50, -2.99	Mild to moderate drought	-1.25, -1.99
< -3.00	Severe drought	<-2.75

SPxx

Category	Range
Severe wetness	> 1.50
Mild to moderate wetness	1.00, 1.49
Near normal	-0.99, 0.99
Mild to moderate drought	-1.49, -1
Severe drought	<-1.50

Figure A.2: Loss Cost Ratios

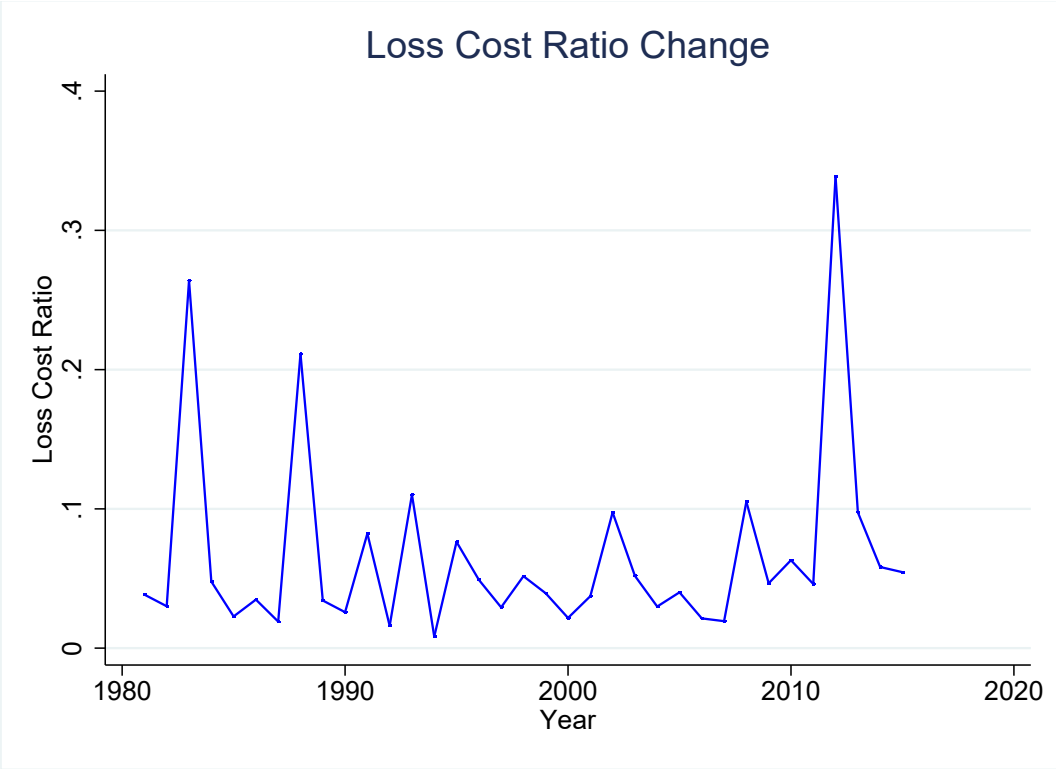

Figure A.3: GM Adoption

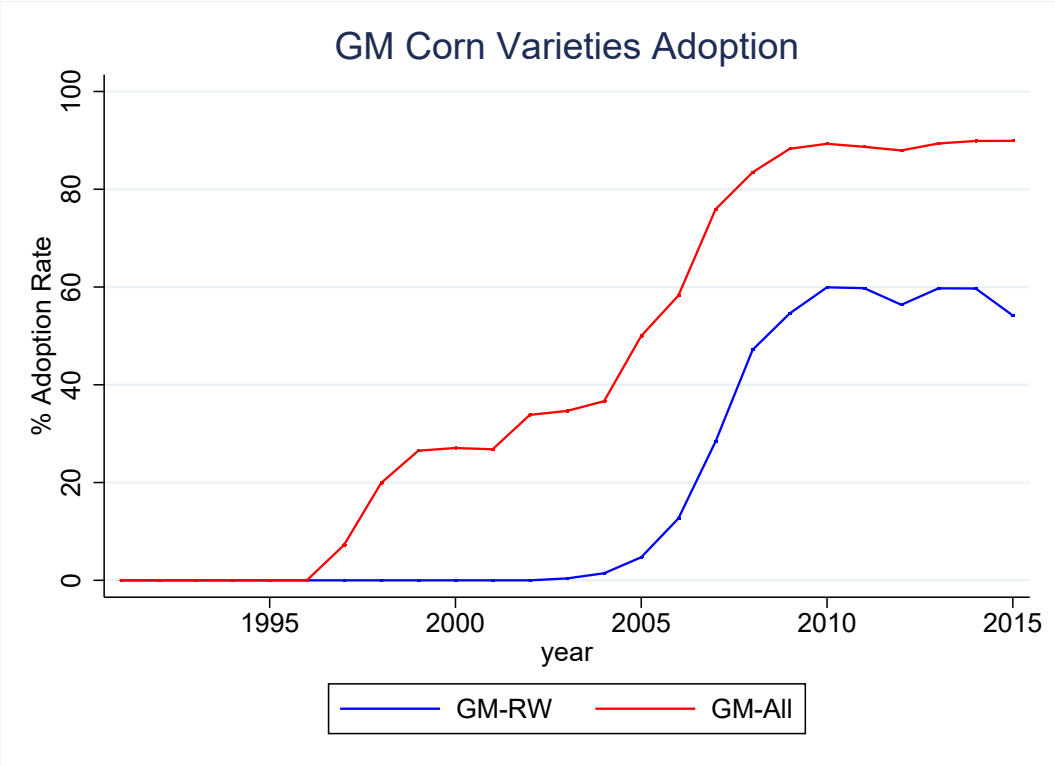

Figure A.4: Degree Days Measures

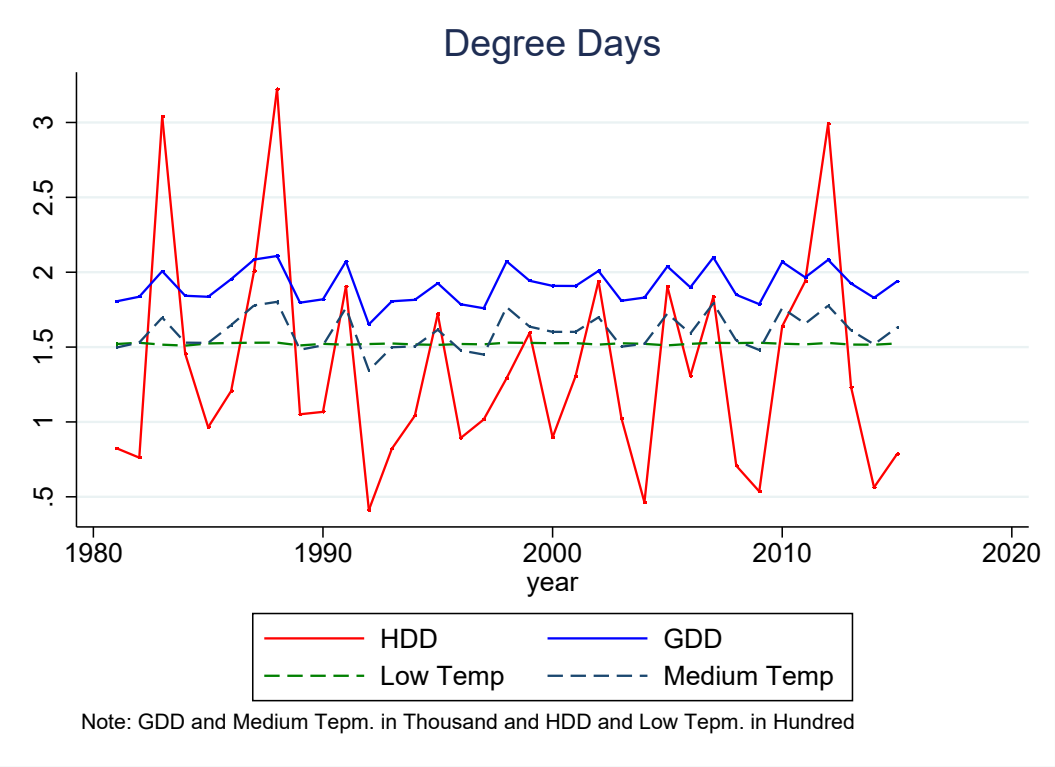

Figure A.5: Weather Variables

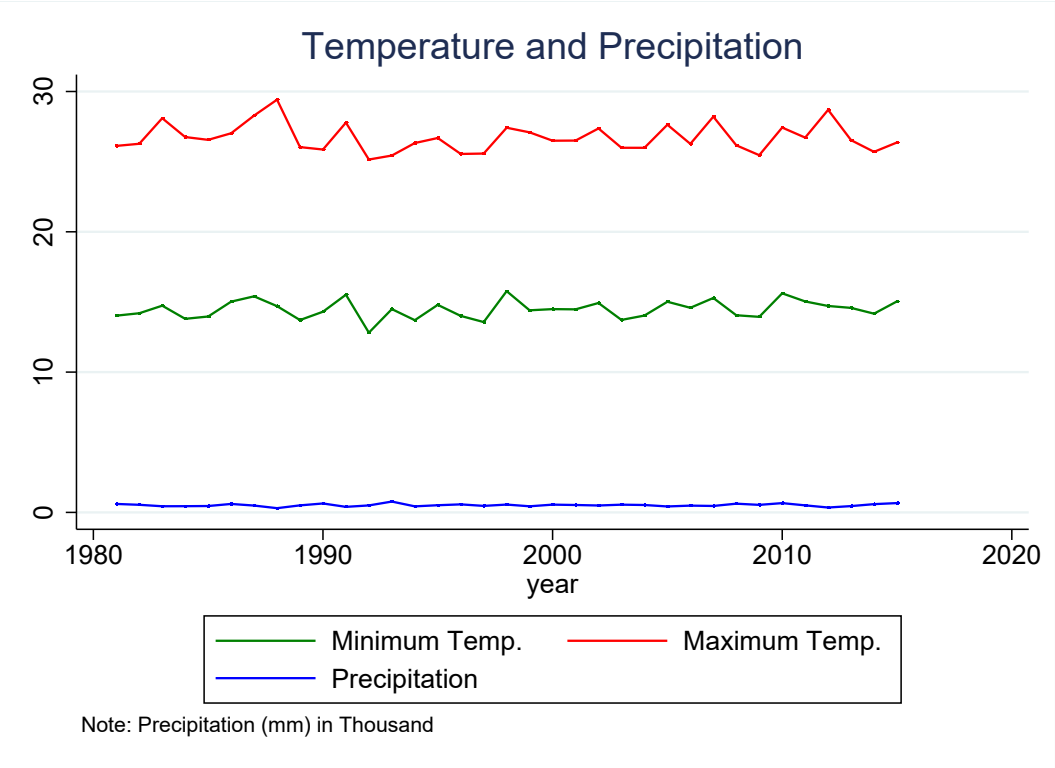

Figure A.6: Index Variables