Section 4: Quantifying the Impact of Weather on Crop Yields

Shuo Yu

2024-02-12

Quantifying the yield-weather relationship is crucial for predicting crop responses, assessing climate change impacts, managing agricultural risks, informing policies, guiding technological innovations, and improving yield forecasts. This helps ensure food security and resilience in agriculture.

Data Sources

- Crop Yields: USDA NASS Quick Stats

  The USDA National Agricultural Statistics Service (NASS) Quick Stats is an online database providing extensive agricultural data for the United States. It allows users to query, download, and analyze annual, seasonal, and survey-based statistics on crop production, livestock, farm economics, and environmental factors.

- Weather Data: PRISM Climate Data

  The Parameter-elevation Regressions on Independent Slopes Model (PRISM) is a high-resolution climate dataset developed by the PRISM Climate Group at Oregon State University. It integrates point station data, elevation, and terrain influences to improve accuracy, particularly in complex geographic regions. PRISM provides daily, monthly, and long-term climate datasets for key weather variables such as precipitation, temperature (minimum, maximum, mean), and dew point. It is widely used in agricultural research, climate impact studies, hydrology, and environmental policy analysis due to its fine spatial resolution (e.g., 4 km or finer for some datasets).

```r
library(tidyverse)

setwd("C:/Users/shuoy/Dropbox/161/Sections/Section4")
```

```r
weather <- read.csv("Weather.csv", header = TRUE)
str(weather)
```

```
'data.frame':	132 obs. of  17 variables:
 $ system.index: chr  "1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_1_
 $ AFFGEOID    : chr  "0500000US19023" "0500000US17113" "0500000US18085" "0500000US19023" ..
 $ ALAND       : num  1.50e+09 3.06e+09 1.38e+09 1.50e+09 3.06e+09 ...
 $ AWATER      : num  4191852 7853695 59459185 4191852 7853695 ...
 $ COUNTYFP    : int  23 113 85 23 113 85 23 113 85 23 ...
 $ COUNTYNS    : int  465201 1784833 450367 465201 1784833 450367 465201 1784833 450367 4652
 $ GEOID       : int  19023 17113 18085 19023 17113 18085 19023 17113 18085 19023 ...
 $ LSAD        : int  6 6 6 6 6 6 6 6 6 6 ...
 $ NAME        : chr  "Butler" "McLean" "Kosciusko" "Butler" ...
 $ STATEFP     : int  19 17 18 19 17 18 19 17 18 19 ...
 $ Year        : num  1981 1981 1981 1982 1982 ...
 $ edd         : num  27.8 35.5 22.7 29.6 31.6 ...
 $ gdd         : num  1760 2001 1794 1714 1964 ...
 $ ppt         : num  632 915 842 618 524 ...
 $ tavg        : num  17.5 18.9 17.7 16.8 18.3 ...
 $ tmax        : num  23.8 24.5 23.2 22.6 24.3 ...
 $ .geo        : chr  "{\"type\":\"Polygon\",\"coordinates\":[[[-93.02700131102496,42.556808
```

```r
corn_yield <- read.csv("CornYields.csv", header = TRUE)
str(corn_yield)
```

```
'data.frame':	12304 obs. of  21 variables:
 $ Program         : chr  "SURVEY" "SURVEY" "SURVEY" "SURVEY" ...
 $ Year            : int  2023 2023 2023 2023 2023 2023 2023 2023 2023 2023 ...
 $ Period          : chr  "YEAR" "YEAR" "YEAR" "YEAR" ...
 $ Week.Ending     : logi  NA NA NA NA NA NA ...
 $ Geo.Level       : chr  "COUNTY" "COUNTY" "COUNTY" "COUNTY" ...
 $ State           : chr  "ILLINOIS" "ILLINOIS" "ILLINOIS" "ILLINOIS" ...
 $ State.ANSI      : int  17 17 17 17 17 17 17 17 17 17 ...
 $ Ag.District     : chr  "" "CENTRAL" "CENTRAL" "CENTRAL" ...
 $ Ag.District.Code: int  99 40 40 40 40 40 40 40 40 40 ...
 $ County          : chr  "OTHER COUNTIES" "LOGAN" "MACON" "MASON" ...
 $ County.ANSI     : int  NA 107 115 125 113 129 143 175 179 203 ...
 $ Zip.Code        : logi  NA NA NA NA NA NA ...
 $ Region          : logi  NA NA NA NA NA NA ...
 $ watershed_code  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Watershed       : logi  NA NA NA NA NA NA ...
```

```
 $ Commodity       : chr  "CORN" "CORN" "CORN" "CORN" ...
 $ Data.Item       : chr  "CORN, GRAIN - YIELD, MEASURED IN BU / ACRE" "CORN, GRAIN - YIELD,
 $ Domain          : chr  "TOTAL" "TOTAL" "TOTAL" "TOTAL" ...
 $ Domain.Category : chr  "NOT SPECIFIED" "NOT SPECIFIED" "NOT SPECIFIED" "NOT SPECIFIED" ..
 $ Value           : num  208 211 225 208 223 ...
 $ CV....          : num  1 1.6 2.8 3.6 2.2 3.9 3 1.7 2.2 2.4 ...
```

```r
soy_yield <- read.csv("SoyYields.csv", header = TRUE)
str(soy_yield)
```

```
'data.frame':   12296 obs. of  21 variables:
 $ Program         : chr  "SURVEY" "SURVEY" "SURVEY" "SURVEY" ...
 $ Year            : int  2023 2023 2023 2023 2023 2023 2023 2023 2023 2023 ...
 $ Period          : chr  "YEAR" "YEAR" "YEAR" "YEAR" ...
 $ Week.Ending     : logi  NA NA NA NA NA NA ...
 $ Geo.Level       : chr  "COUNTY" "COUNTY" "COUNTY" "COUNTY" ...
 $ State           : chr  "ILLINOIS" "ILLINOIS" "ILLINOIS" "ILLINOIS" ...
 $ State.ANSI      : int  17 17 17 17 17 17 17 17 17 17 ...
 $ Ag.District     : chr  "" "CENTRAL" "CENTRAL" "CENTRAL" ...
 $ Ag.District.Code: int  99 40 40 40 40 40 40 40 40 40 ...
 $ County          : chr  "OTHER COUNTIES" "LOGAN" "MARSHALL" "MASON" ...
 $ County.ANSI     : int  NA 107 123 125 113 129 143 175 179 203 ...
 $ Zip.Code        : logi  NA NA NA NA NA NA ...
 $ Region          : logi  NA NA NA NA NA NA ...
 $ watershed_code  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Watershed       : logi  NA NA NA NA NA NA ...
 $ Commodity       : chr  "SOYBEANS" "SOYBEANS" "SOYBEANS" "SOYBEANS" ...
 $ Data.Item       : chr  "SOYBEANS - YIELD, MEASURED IN BU / ACRE" "SOYBEANS - YIELD, MEASU
 $ Domain          : chr  "TOTAL" "TOTAL" "TOTAL" "TOTAL" ...
 $ Domain.Category : chr  "NOT SPECIFIED" "NOT SPECIFIED" "NOT SPECIFIED" "NOT SPECIFIED" ..
 $ Value           : num  60.7 68.4 68.8 68.1 71.8 67.2 70.7 71.7 74.7 67.4 ...
 $ CV....          : num  1.2 2 3.3 3 1.5 1.6 2.6 1.2 1.4 3 ...
```

Data Cleaning

To prepare the data frame for visualization and regression analysis, we need to clean and structure the data into a tidy format that ensures consistency and usability. The key steps include:

- Understanding the Data: What is the unit of observation in the data frame? What key variables are needed for analysis? For this section, we will focus on a single county in

Iowa over time to establish a baseline understanding of the effect of weather on crop yields. As we progress, we will expand the analysis to include additional regions and examine other sources of variation in the dataset.

- Filtering the Data: Select relevant variables and filter out observations relevant to the specific study area.
- Merging Datasets: Combine multiple datasets based on year and region to ensure alignment.

This cleaned dataset will serve as the foundation for subsequent data visualization and regression modeling.

```r
# Weather Data
filtered_weather <- weather %>%
  select(GEOID, NAME, Year, edd, gdd, ppt, tavg, tmax) %>% # Select relevant columns
  filter(GEOID == 19023) # Filter for Iowa and Butler County (GEOID = 19023)

# Corn Yield Data
filtered_corn_yield <- corn_yield %>%
  select(Year, County.ANSI, State.ANSI, Value) %>%  # Select relevant columns
  rename(CornYield = Value) %>%  # Rename 'Value' to 'CornYield'
  filter(State.ANSI == 19, County.ANSI == 23) %>%  # Filter for Iowa and Butler County
  mutate(
      GEOID = State.ANSI*1000+County.ANSI # Create GEOID that is 5 digits
    ) %>%
  select(-State.ANSI, -County.ANSI)  # Drop State.ANSI and County.ANSI

# Soybean Yield Data
filtered_soy_yield <- soy_yield %>%
  select(Year, County.ANSI, State.ANSI, Value) %>%  # Select relevant columns
  rename(SoyYield = Value) %>%  # Rename 'Value' to 'SoyYield'
  filter(State.ANSI == 19, County.ANSI == 23) %>%  # Filter for Iowa (State.ANSI = 19) and Bu
  mutate(
      GEOID = State.ANSI*1000+County.ANSI # Create GEOID that is 5 digits
    ) %>%
  select(-State.ANSI, -County.ANSI)  # Drop State.ANSI and County.ANSI

# Merge Datasets
full_data <- filtered_corn_yield %>%
  left_join(filtered_soy_yield, by = c("GEOID", "Year")) %>%
  left_join(filtered_weather, by = c("GEOID", "Year")) %>%
  na.omit()
```

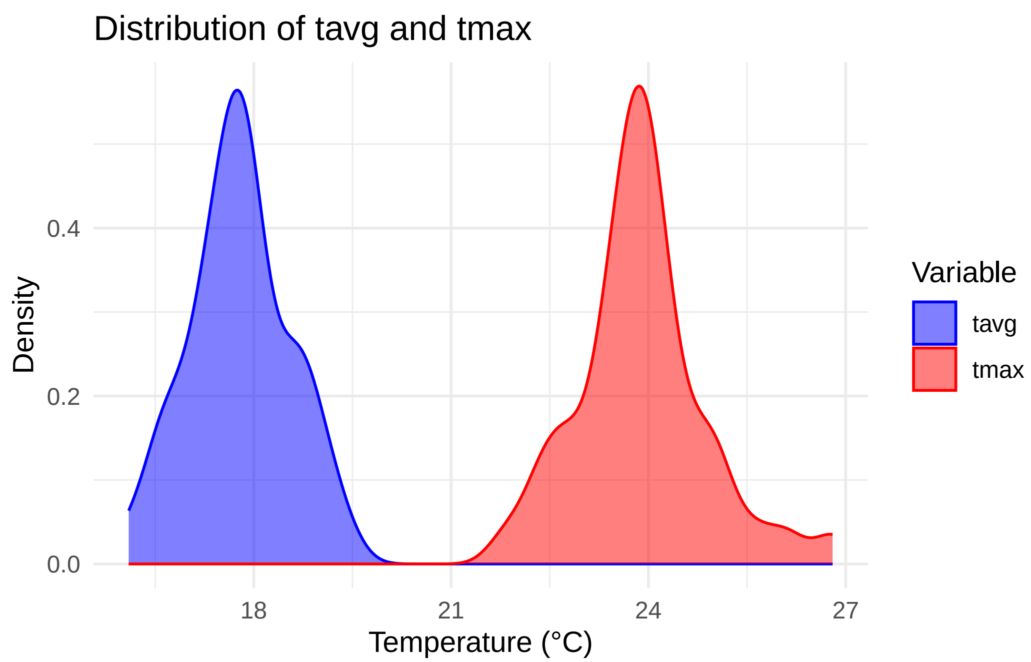
```r
summary(full_data)
```

```
      Year        CornYield         GEOID          SoyYield
 Min.   :1981   Min.   : 73.9   Min.   :19023   Min.   :24.40
 1st Qu.:1992   1st Qu.:130.6   1st Qu.:19023   1st Qu.:40.80
 Median :2002   Median :157.9   Median :19023   Median :47.50
 Mean   :2002   Mean   :157.2   Mean   :19023   Mean   :46.56
 3rd Qu.:2012   3rd Qu.:184.3   3rd Qu.:19023   3rd Qu.:53.05
 Max.   :2023   Max.   :212.4   Max.   :19023   Max.   :61.60
     NAME                edd                gdd            ppt
 Length:43          Min.   :  6.937   Min.   :1549   Min.   : 347.9
 Class :character   1st Qu.: 32.982   1st Qu.:1767   1st Qu.: 538.5
 Mode  :character   Median : 56.943   Median :1833   Median : 632.4
                    Mean   : 61.377   Mean   :1842   Mean   : 643.3
                    3rd Qu.: 76.101   3rd Qu.:1957   3rd Qu.: 737.7
                    Max.   :234.017   Max.   :2099   Max.   :1045.5
      tavg            tmax
 Min.   :16.10   Min.   :21.84
 1st Qu.:17.32   1st Qu.:23.39
 Median :17.78   Median :23.83
 Mean   :17.78   Mean   :23.90
 3rd Qu.:18.23   3rd Qu.:24.26
 Max.   :19.37   Max.   :26.80
```

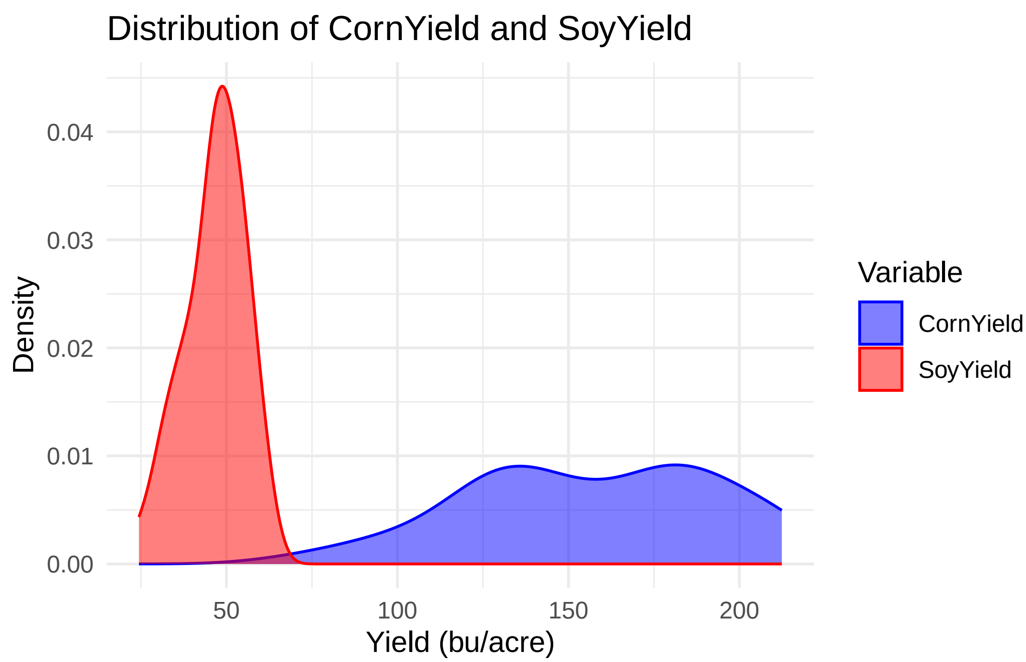
Data Visualization

Distribution of Key Variables

```r
# Create density plot for average and max temperature
full_data %>%
  ggplot() +
  geom_density(aes(x = tavg, fill = "tavg"), alpha = 0.5, color = "blue") +
  geom_density(aes(x = tmax, fill = "tmax"), alpha = 0.5, color = "red") +
  scale_fill_manual(values = c("tavg" = "blue", "tmax" = "red")) +  # Custom colors
  labs(title = "Distribution of tavg and tmax",
       x = "Temperature (°C)",
       y = "Density",
       fill = "Variable") +
  theme_minimal()
```

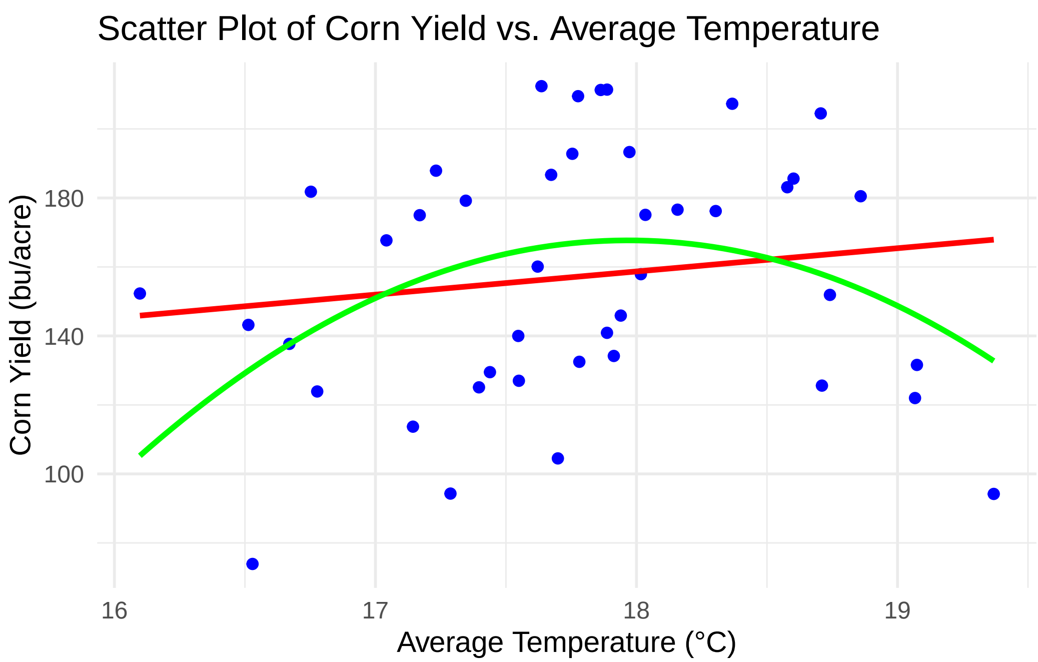



```
# Create density plot for corn yields and soybean yields
full_data %>%
  ggplot() +
  geom_density(aes(x = CornYield, fill = "CornYield"),
               alpha = 0.5, color = "blue") +
  geom_density(aes(x = SoyYield, fill = "SoyYield"),
               alpha = 0.5, color = "red") +
  scale_fill_manual(values = c("CornYield" = "blue", "SoyYield" = "red")) +  # Custom colors
  labs(title = "Distribution of CornYield and SoyYield",
       x = "Yield (bu/acre)", y = "Density", fill = "Variable") +
  theme_minimal()
```

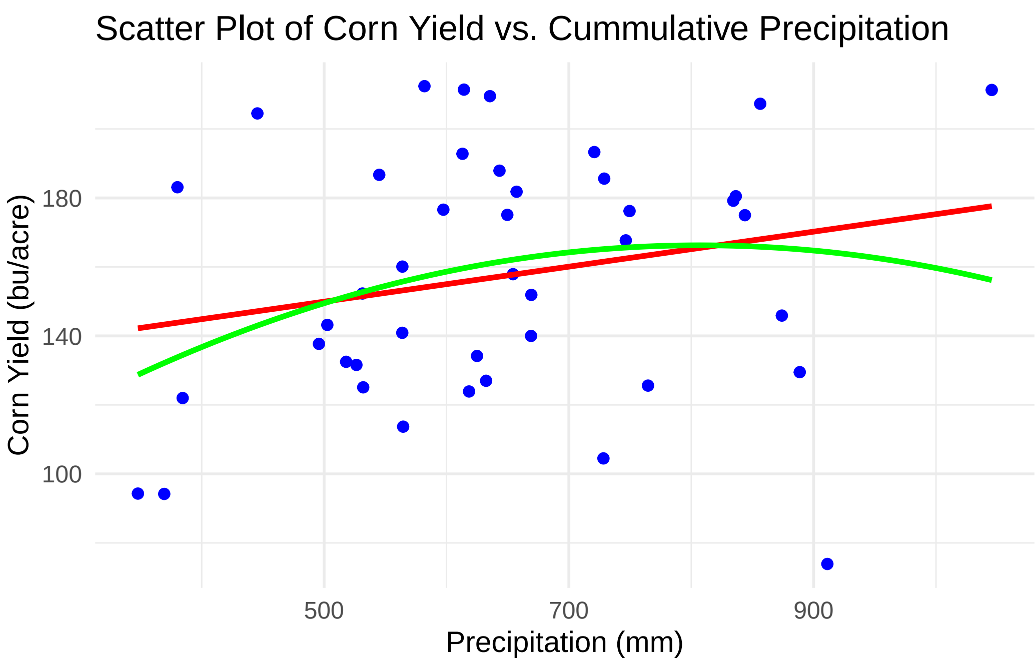


Scatter Plots

```
# Create scatter plot of corn yield vs. average temperature
full_data %>%
  ggplot(aes(x = tavg, y = CornYield)) +
  geom_point(color = "blue") +  # Scatter points
  geom_smooth(method = "lm", color = "red",
              se = FALSE, formula = y ~ x) +  # Linear fit
  geom_smooth(method = "lm", color = "green",
              se = FALSE, formula = y ~ poly(x, 2)) +  # Quadratic fit
  labs(title = "Scatter Plot of Corn Yield vs. Average Temperature",
       x = "Average Temperature (°C)",
       y = "Corn Yield (bu/acre)") +
  theme_minimal()
```




```r
# Create scatter plot of corn yield vs. precipitation
full_data %>%
  ggplot(aes(x = ppt, y = CornYield)) +
  geom_point(color = "blue") +  # Scatter points
  geom_smooth(method = "lm", color = "red",
              se = FALSE, formula = y ~ x) +  # Linear fit
  geom_smooth(method = "lm", color = "green",
              se = FALSE, formula = y ~ poly(x, 2)) +  # Quadratic fit
  labs(title = "Scatter Plot of Corn Yield vs. Cummulative Precipitation",
       x = "Precipitation (mm)",
       y = "Corn Yield (bu/acre)") +
  theme_minimal()
```



Multiple Variable Linear Regression Model

We model the relationship between corn yield and key weather variables as follows:

$$CornYield_t = \beta_0 + \beta_1 AvgTemperature_t + \beta_2 Precipitation_t + \beta_3 Year_t + \varepsilon_t$$

where:

- Outcome variable: $CornYield_t$ – Corn yield in year t
- Independent variables:
  - $AvgTemperature_t$ Average temperature during the growing season (April to September
  - $Precipitation_t$ – Cummulative precipitation during the growing season (April to September)
  - $Year_t$ - Linear time trend
- ε_t represents the error term, capturing unobserved factors affecting corn yield.

This model allows us to estimate the impact of temperature and precipitation on corn yield, providing insights into how climate conditions influence agricultural productivity.

```r
# Run multiple linear regression
model <- lm(CornYield ~ tavg + ppt + Year, data = full_data)

# Display summary of the model
summary(model)
```

```

Call:
lm(formula = CornYield ~ tavg + ppt + Year, data = full_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-69.876  -8.499   3.002  11.626  25.439 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.639e+03  4.760e+02  -9.745 5.29e-12 ***
tavg        -8.534e-01  4.066e+00  -0.210    0.835    
ppt          2.634e-02  1.921e-02   1.371    0.178    
Year         2.395e+00  2.439e-01   9.818 4.30e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.2 on 39 degrees of freedom
Multiple R-squared:  0.7342,	Adjusted R-squared:  0.7137 
F-statistic:  35.9 on 3 and 39 DF,  p-value: 2.642e-11
```

Question: What are the main takeaways?

- Corn yield has a significant upward trend over time (Year coefficient = 2.395, $p < 0.001$). Each additional year is associated with a 2.395 bushel/acre increase in corn yield. This could reflect technological advancements or improved farming practices.
- Temperature (`tavg`) and precipitation (`ppt`) are not statistically significant predictors of yield (p_tavg = 0.835, p_ppt = 0.178).
- Model explains 71.4% of yield variation (`R² = 0.7137`), but residual standard error (19.2 bu/acre) suggests other unobserved factors.
- F-statistic = 35.9, p-value = 2.642e-11 → The model is highly significant, meaning at least one predictor is statistically relevant in explaining corn yield.

References

Schlenker, Wolfram, and Michael J. Roberts. "Nonlinear effects of weather on corn yields." *Review of agricultural economics* 28, no. 3 (2006): 391-398. link

Schlenker, Wolfram, and Michael J. Roberts. "Nonlinear temperature effects indicate severe damages to US crop yields under climate change." *Proceedings of the National Academy of sciences* 106, no. 37 (2009): 15594-15598. link