Section 5: Quantifying the Impact of Weather on Crop Yields (Continued)

Shuo Yu

2024-02-19

```
rm(list = ls())

# Install package for fixed effect regressions
# install.packages("fixest")

library(tidyverse)
library(fixest)

setwd("C:/Users/shuoy/Dropbox/161/Sections/Section5")
```

Data

```
weather <- read.csv("Weather.csv", header = TRUE)
corn_yield <- read.csv("CornYields.csv", header = TRUE)
soy_yield <- read.csv("SoyYields.csv", header = TRUE)
```

Data Cleaning

```
# Weather Data
filtered_weather <- weather %>%
  select(GEOID, NAME, Year, edd, gdd, ppt, tavg, tmax, east_dummy)

# Corn Yield Data
filtered_corn_yield <- corn_yield %>%
  select(Year, County.ANSI, State.ANSI, Ag.District, Value) %>%  # Select relevant columns
```

```
  rename(CornYield = Value) %>%  # Rename 'Value' to 'CornYield'
  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, Ag.District, Value) %>%  # Select relevant columns
  rename(SoyYield = Value) %>%  # Rename 'Value' to 'SoyYield'
  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")) %>%
  mutate(Ag.District = coalesce(Ag.District.x, Ag.District.y)) %>%  # Fill missing with non-
  select(-Ag.District.x, -Ag.District.y)  %>%  # Drop original columns
  filter(!is.na(gdd) & !is.na(edd) & !is.na(ppt))

full_data$Year <- full_data$Year - 1980

full_data$GEOID <- as.factor(full_data$GEOID)
full_data$Year <- as.numeric(full_data$Year)
```

Let's use corn yields as an example. You can follow similar steps and adjust the code to obtain results for soybeans.

```
# Drops rows with missing values (na.omit()).
corn_data <- full_data %>%
  select(-SoyYield) %>%
  na.omit()

# Filters counties that have at least 21 observations to ensure major corn groing counties w
corn_data <- corn_data %>%
  group_by(GEOID) %>%
  filter(n() > 20) %>%
  ungroup()
```

```
summary(corn_data)
```

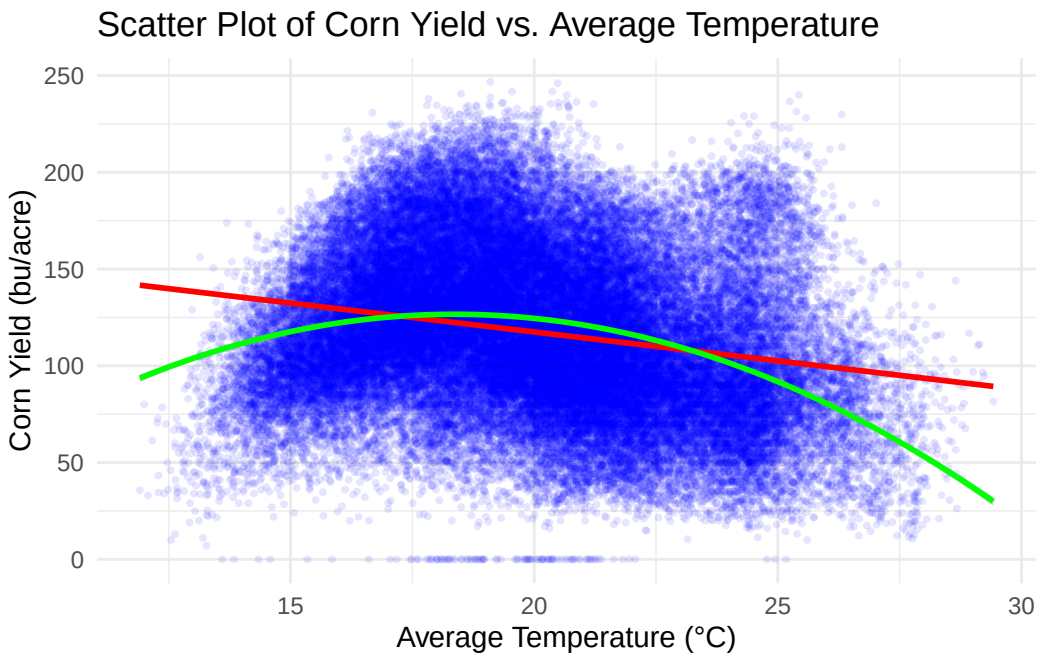
```
      Year         CornYield          GEOID           NAME
 Min.   : 1.00   Min.   :  0.0   1049   :   43   Length:76702
 1st Qu.:10.00   1st Qu.: 89.6   1077   :   43   Class :character
 Median :20.00   Median :117.0   1079   :   43   Mode  :character
 Mean   :20.56   Mean   :118.7   1083   :   43
 3rd Qu.:31.00   3rd Qu.:147.6   1089   :   43
 Max.   :43.00   Max.   :277.1   5067   :   43
                                 (Other):76444
      edd               gdd              ppt               tavg
 Min.   :   0.00   Min.   : 926.4   Min.   :   0.8864   Min.   :11.91
 1st Qu.:  48.78   1st Qu.:1798.4   1st Qu.: 455.8868   1st Qu.:17.54
 Median : 122.05   Median :2155.5   Median : 568.0678   Median :19.71
 Mean   : 166.62   Mean   :2189.6   Mean   : 571.3291   Mean   :19.83
 3rd Qu.: 242.75   3rd Qu.:2529.4   3rd Qu.: 685.9282   3rd Qu.:21.87
 Max.   :1558.21   Max.   :3781.0   Max.   :1705.4913   Max.   :30.17

      tmax         east_dummy     Ag.District
 Min.   :17.77   Min.   :0.0000   Length:76702
 1st Qu.:23.91   1st Qu.:1.0000   Class :character
 Median :26.07   Median :1.0000   Mode  :character
 Mean   :26.16   Mean   :0.9015
 3rd Qu.:28.26   3rd Qu.:1.0000
 Max.   :38.40   Max.   :1.0000
```

Data Visualization

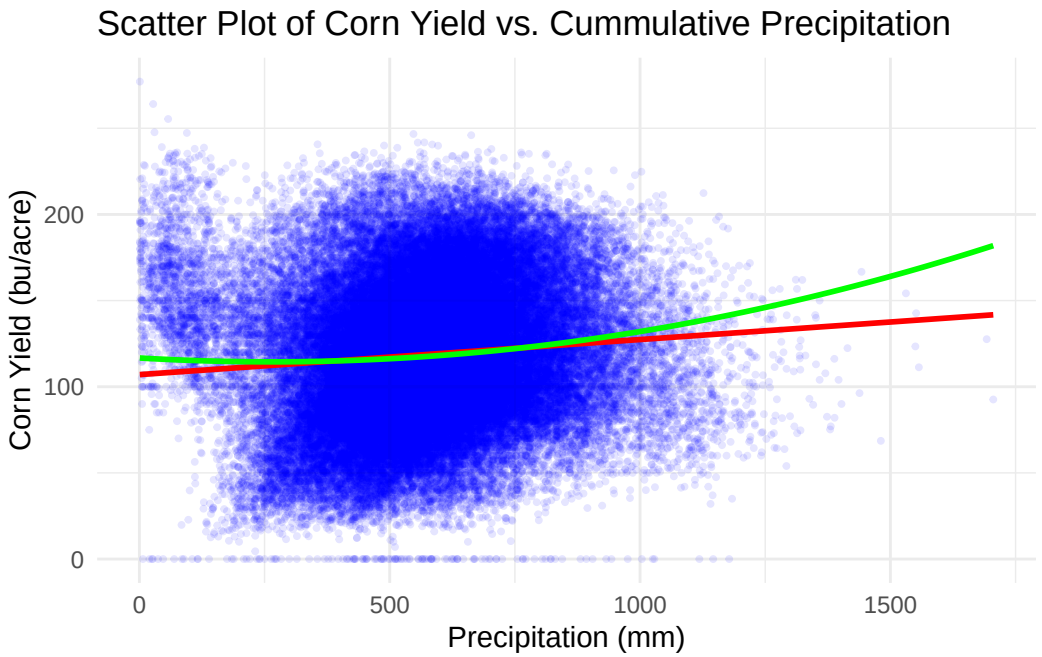
Scatter Plots

```
# Create scatter plot of corn yield vs. average temperature
scatterplot1 <- corn_data %>%
  filter(east_dummy==1) %>%
  ggplot(aes(x = tavg, y = CornYield)) +
  geom_point(color = "blue", alpha = 0.1, size = 0.7) +  # Scatter points
  geom_smooth(method = "lm", color = "red",
              se = TRUE, formula = y ~ x) +  # Linear fit
  geom_smooth(method = "lm", color = "green",
```

```
              se = TRUE, 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()

scatterplot1
```




```
# Create scatter plot of corn yield vs. precipitation
scatterplot2 <- corn_data %>%
  ggplot(aes(x = ppt, y = CornYield)) +
  geom_point(color = "blue", alpha = 0.1, size = 0.7) +  # 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()

scatterplot2
```



Multiple Regression with a Nonlinear Functional Form

This section introduces a quadratic term for precipitation to capture the inverted U-shaped relationship with crop yield. Additionally, we construct Growing Degree Days (GDD) and Extreme Degree Days (EDD) to account for the nonlinear effects of temperature on yield.

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

$$CornYield_{it} = \beta_0 + \beta_1 GDD_{it} + \beta_2 EDD_{it} + \beta_3 Precipitation_{it} + \beta_4 Precipitation_{it}^2 + \beta_5 Year_{it} + \varepsilon_{it}$$

where:

- Outcome variable: $CornYield_t$ – Corn yield in year t
- Independent variables:
  - GDD_t Growing degree days during the growing season (April to September)
  - EDD_t Extreme degree days during the growing season
  - $Precipitation_t$ – Cumulative precipitation during the growing season
  - $Year_t$ - Linear time trend
- ε_t represents the error term, capturing unobserved factors affecting corn yield.

Adding a Squared Term of Precipitation

Diminishing returns and the inverted U-shape relationship:

- Moderate precipitation is essential for crop growth, leading to higher yields.
- However, excessive rainfall can lead to water logging, nutrient leaching, or disease outbreaks, which negatively impact yields.
- By including both precipitation (linear term) and precipitation squared (quadratic term), the model can account for an inverted U-shaped relationship, where yield increases with precipitation up to a certain point and then starts to decline.

Using Growing Degree Days (GDD) and Extreme Degree Days (EDD)

- Crops have optimal temperature ranges for photosynthesis and growth. Extreme heat (>30°C) damages plant cells, reduces enzyme activity, and disrupts photosynthesis (e.g. Schlenker & Roberts, 2009).
- When high temperatures are combined with low precipitation, the negative impact on crop yields is significantly amplified. This interaction leads to severe water stress, heat stress, and nutrient deficiencies, which reduce plant productivity. High temperatures accelerate evapotranspiration, meaning crops lose water faster than they can absorb it from the soil.
- **GDD** measure heat accumulation used to predict plant development stages. GDD between 8 and 30°C is defined as

$$GDD = \sum_{t=1}^{T} \max(\min(T_{avg}, 30) - 8, 0)$$

- **EDD** measure heat stress on crops when temperatures exceed a critical threshold. EDD above 30°C is defined as

$$EDD = \sum_{t=1}^{T} \max(T_{max} - 30, 0)$$

```r
# Run multiple linear regression
model <- lm(CornYield ~ gdd + edd + ppt + I(ppt^2) + Year, data = corn_data)

# Display summary of the model
options(scipen = 999)
summary(model)
```

```
Call:
lm(formula = CornYield ~ gdd + edd + ppt + I(ppt^2) + Year, data = corn_data)

Residuals:
     Min       1Q   Median       3Q      Max
-145.390  -20.841    0.728   22.194  171.951

Coefficients:
                 Estimate    Std. Error t value            Pr(>|t|)
(Intercept) 112.641807933   0.892391077  126.22 <0.0000000000000002 ***
gdd           0.007547841   0.000492079   15.34 <0.0000000000000002 ***
edd          -0.094064069   0.001661103  -56.63 <0.0000000000000002 ***
ppt          -0.097580399   0.002760052  -35.35 <0.0000000000000002 ***
I(ppt^2)      0.000062391   0.000002139   29.17 <0.0000000000000002 ***
Year          1.865457710   0.009889831  188.62 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.56 on 76696 degrees of freedom
Multiple R-squared:  0.3704,    Adjusted R-squared:  0.3704
F-statistic:  9025 on 5 and 76696 DF,  p-value: < 0.00000000000000022
```

Question:

- What are the key takeaways? Were any estimates unexpected?
- Is the model correctly specified? What potential issues might arise with this estimation equation?

Some counties have naturally better soil or farming techniques, those county-specific factors might bias our results. Solution? Use county fixed effects, which account for each county's unique characteristics that don't change over time (like soil quality or historical farming practices). This way, our model focuses only on how changes in weather impact yield.

Fixed Effects Model

Fixed effects in regression help control for time-invariant differences between groups, allowing us to isolate the impact of changes in other factors on the outcome. Conceptually, this approach is similar to subtracting each county's average yield from its yearly yields and adjusting weather variables—such as EDD, GDD, and precipitation—by their county-level means. By doing so, we eliminate the influence of permanent characteristics (e.g., soil quality, historical farming

practices) and focus on how within-county variations in weather conditions affect crop yields over time.

$$CornYield_{it} = \beta_0 + \beta_1 GDD_{it} + \beta_2 EDD_{it} + \beta_3 PPT_{it} + \beta_4 PPT_{it}^2 + \alpha_i + \delta_t + \gamma_i Year_{it} + \varepsilon_{it}$$

- **County Fixed Effects (α_i):** Removes the impact of factors that differ across counties but stay the same over time (e.g., soil quality, local policies).
- **Year Fixed Effects (δ_t):** Controls for yearly shocks affecting all counties (e.g., nationwide economic conditions, federal policies).
- **County-Specific Time Trends (γ_i):** Allows each county to have its own trend over time (e.g., gradual technology adoption in some regions).

```
fe_model <- feols(CornYield ~ gdd + edd + ppt + I(ppt^2) |
                  Year + GEOID[Year],
                  data = corn_data)

summary(fe_model)
```

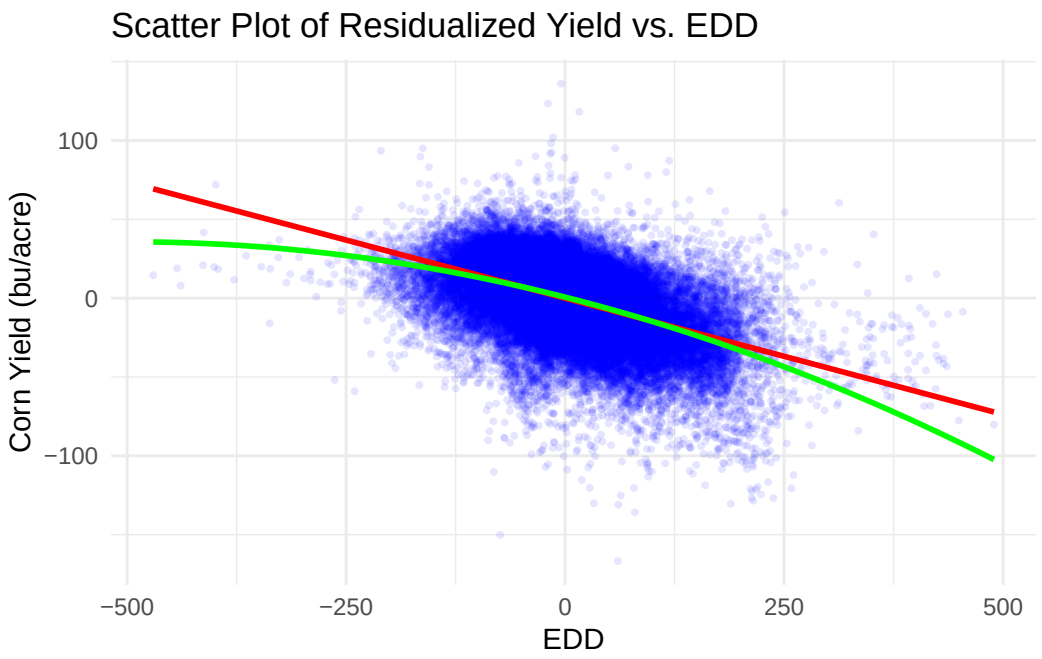
```
OLS estimation, Dep. Var.: CornYield
Observations: 76,702
Fixed-effects: Year: 43,  GEOID: 2,023
Varying slopes: Year (GEOID): 2,023
Standard-errors: Clustered (Year)
          Estimate Std. Error  t value           Pr(>|t|)
gdd       0.023057   0.010743  2.14634 0.0376686293235542 *
edd      -0.143734   0.014853 -9.67732 0.0000000000029597 ***
ppt       0.061948   0.016550  3.74301 0.0005462919335757 ***
I(ppt^2) -0.000046   0.000012 -3.83629 0.0004131998584641 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 17.2     Adj. R2: 0.815089
             Within R2: 0.158299
```

To understand how fixed effects work:

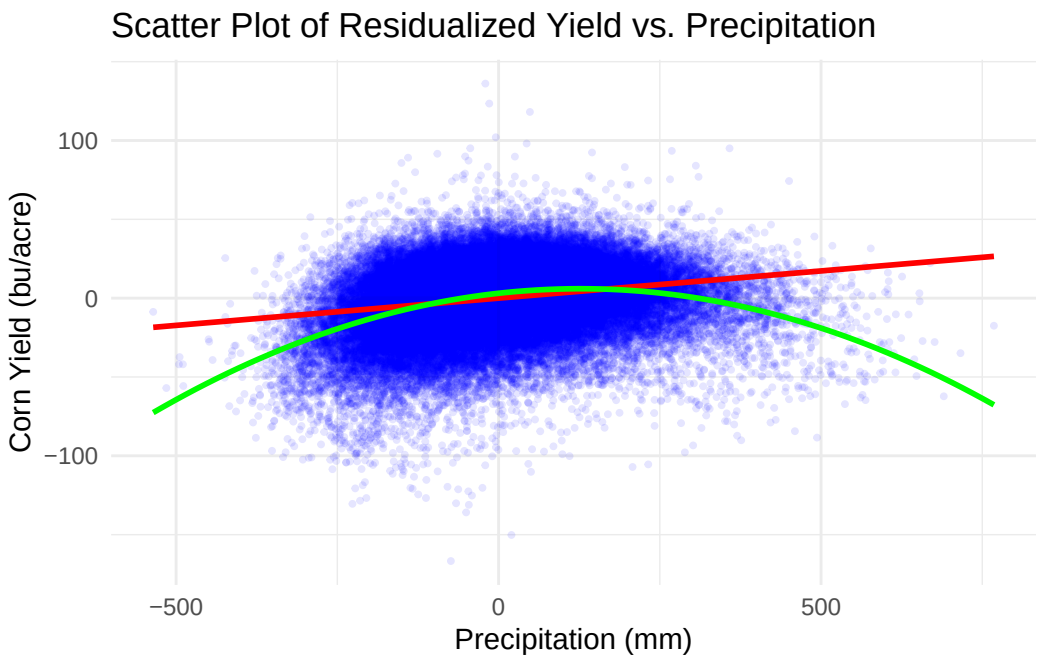
- Run a fixed effects regression for corn yield, EDD, and precipitation, removing county (GEOID) and year (Year) effects, and the county-specific time trend.
- Extract the residuals, which capture the variation in yield and precipitation not explained by fixed effects.

- Create a scatter plot of the residuals, adding a linear trend line to visualize their relationship.

```
# Regress Yield and Precipitation separately on fixed effects
yield_fe <- feols(CornYield ~ 1 | GEOID + GEOID[Year], data = corn_data)
edd_fe <- feols(edd ~ 1 | GEOID + GEOID[Year], data = corn_data)
precip_fe <- feols(ppt ~ 1 | GEOID + GEOID[Year], data = corn_data)

# Extract residuals
corn_data <- corn_data %>%
  mutate(resid_yield = resid(yield_fe),
         resid_edd = resid(edd_fe),
         resid_ppt = resid(precip_fe))

# Scatter plot of residuals
scatterplot3 <- corn_data %>%
  ggplot(aes(x = resid_edd, y = resid_yield)) +
  geom_point(color = "blue", alpha = 0.1, size = 0.7) +  # 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 Residualized Yield vs. EDD",
       x = "EDD",
       y = "Corn Yield (bu/acre)") +
  theme_minimal()

scatterplot3
```




```
# Scatter plot of residuals
scatterplot4 <- corn_data %>%
  ggplot(aes(x = resid_ppt, y = resid_yield)) +
  geom_point(color = "blue", alpha = 0.1, size = 0.7) +  # 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 Residualized Yield vs. Precipitation",
       x = "Precipitation (mm)",
       y = "Corn Yield (bu/acre)") +
  theme_minimal()

scatterplot4
```



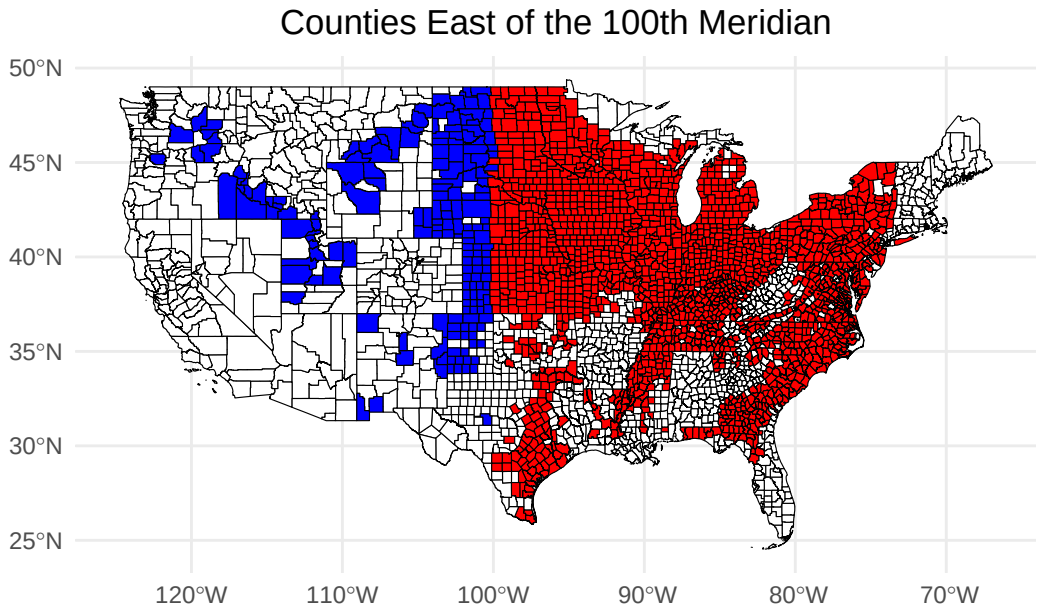
Spatial Heterogeneity

Restricting to counties east of the 100th meridian ensures a more homogeneous sample, reducing noise caused by irrigation, land use, and climate variability.

- Climate and agricultural practices: East of the 100th meridian has sufficient rainfall for rain-fed agriculture, while the west relies on irrigation.
- Crop differences: Eastern counties grow more corn and soybeans; western counties focus on range land and drought-resistant crops.
- Historical settlement: Denser populations and smaller farms in the east, larger, irrigated farms in the west.
- Policy and programs: Many federal agricultural policies (e.g., crop insurance, EQIP) target rain-fed regions differently from irrigated areas.

Map of Counties Included in the Regression East of the 100th Meridian

```
|
|                                                                                |  0%
|
```



East of 100th Meridian Line

```
fe_model <- feols(CornYield ~ gdd + edd + ppt + I(ppt^2) |
                  Year + GEOID[Year],
                  data = corn_data[corn_data$east_dummy==1, ])

summary(fe_model)
```

```
OLS estimation, Dep. Var.: CornYield
Observations: 69,146
Fixed-effects: Year: 43,  GEOID: 1,798
Varying slopes: Year (GEOID): 1,798
Standard-errors: Clustered (Year)
           Estimate Std. Error  t value          Pr(>|t|)
gdd        0.021023   0.011999  1.75214 0.087049209532693 .
edd       -0.151737   0.016663 -9.10636 0.000000000016998 ***
ppt        0.076309   0.015895  4.80073 0.000020277158093 ***
I(ppt^2) -0.000058   0.000012 -4.97231 0.000011646537237 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 16.8     Adj. R2: 0.817636
```

```
              Within R2: 0.174532
```

West of 100th Meridian Line

```
fe_model <- feols(CornYield ~ gdd + edd + ppt + I(ppt^2) |
                  Year + GEOID[Year],
                  data = corn_data[corn_data$east_dummy==0, ])

summary(fe_model)
```

```
OLS estimation, Dep. Var.: CornYield
Observations: 7,556
Fixed-effects: Year: 43,  GEOID: 225
Varying slopes: Year (GEOID): 225
Standard-errors: Clustered (Year)
             Estimate Std. Error   t value       Pr(>|t|)
gdd       0.02930561   0.010085  2.905742 0.005824011958 **
edd      -0.09001536   0.013799 -6.523438 0.000000070119 ***
ppt       0.01992034   0.018889  1.054577 0.297647906050
I(ppt^2)  0.00000989   0.000028  0.354048 0.725074336303
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 17.8     Adj. R2: 0.830369
              Within R2: 0.061924
```

Questions :

- How do the effects of EDD differ between east and west?

EDD has a stronger negative impact in the east (-0.152, $p < 0.0001$) than in the west (-0.090, $p < 0.0001$), indicating that extreme heat is more damaging in the east, likely due to less reliance on irrigation.

- How does precipitation and its squared term influence the outcome in each region?

This implies that water availability is a key factor for crops in the east, while irrigation buffers its effect in the west.

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