[handwritten annotation: precip is @ county level, right?]

[handwritten annotation: or $\frac{\#migrants_t}{population_t} - \frac{\#migrants_{t-1}}{population_{t-1}}$]

Daniel Malone
Elliane Hall
Md Sayemul Islam

[handwritten annotation: Effects of]

Methodology for Estimating Precipitation on Migration Rates

To evaluate the effect of precipitation on population migration rates in counties in the U.S., we will implement an Ordinary Least Squares (OLS) regression. We will utilize state-year and county fixed effects to adjust for variation *[handwritten: in Y]* across states s and ~~across~~ time t, ~~and~~ variation across *[handwritten: as well as for fixed]* counties c ~~that remain unchanged during our observation period~~. Further, we use a linear model for our regression as it allows for ease in our interpretation of our estimates. Our OLS regression model is as follows:

$$Mig_{ct} = \beta_0 + \beta_1 Precip_{t-1} + \beta_2 Drought_{ct} + \beta_3(Precip_{t-1} \times Drought_{ct}) + \alpha_c + Y_{st} + \varepsilon_{ct}$$

Where,

[handwritten: c]

Mig_{ct} = Population migration rate divided by total population in county c and time t

c = County

t = Time in years

$Precip_{t-1}$ = Precipitation in inches

$Drought_{ct}$ = Dummy variable for drought

$Precip_{t-1}$ x $Drought_{ct}$ = Interaction term to measure precipitation in drought states

s = State

[handwritten annotation: ? $\rightarrow \frac{\#migrants_t - \#migrants_{t-1}}{population_t}$]

[handwritten annotation: counties?]

The beta estimates are our regression estimates, where β_0 is our constant. Our β_1 estimate provides the expected migration, as a percent of total population, explained by precipitation in year *t-1* (one year before the year of observation) in counties that are not experiencing drought. Precipitation is in *t-1* years since changes in precipitation will likely have a delayed effect on migration. Drought is defined in our model as five years of continuous decrease in precipitation year over year, in *[handwritten circle: county c]*. Further, drought is in year t instead of year *t-1* since we want to observe what happens to migration as the state is entering a drought. Thus, our β_2 estimate is the estimated migration change, as a percent of the total population, in a county that is in a drought (*Drought_{ct}* will be equal to one when the county is in a drought, and zero otherwise). Lastly, our β_3 estimate provides the predicted difference in migration, as a percent of the total population, as

[handwritten annotation: awkward wording — come up with a simpler description]

a result of a change in precipitation in year *t-1* for counties that are considered to be in a drought, compared to counties that are not considered to be in a drought.

Since drought is defined in our model as five years of continuous decrease in precipitation, year or year, we will not be able to run our regression until 2005 (5 years after the start of the observation period). There is a slight concern in dropping five years of data, so we will run robustness checks on drought when it is defined as three years of continuous decrease in precipitation and five years of continuous decrease. If we find that using three years in our model is robust, we will use three years of declining precipitation to define drought so that we can keep more observations.

Our county fixed effect, α_c, adjusts for variation across counties that are unchanged across our observation period. α_c represents the dummy variables for all the counties of observations (except for county 1). Thus, we use county fixed effects to try to control for omitted variables that vary across counties that could affect our regression estimates, such as geographical differences across counties, county-level laws that exist throughout the period of observation, universities that exist throughout the period of observation, etc. Our county fixed effects, and adjust for variables that vary on a county level that are unchanging for all years of observation. These are important to adjust for since people move because of their geographical preferences, schooling preferences, etc.

Our state-year fixed effects, Y_{st} , adjust for variation across time periods *t* and state *s* (except for time period 1, which is 2000). We use state-year fixed effects to try to control for omitted variables that could affect our regression estimates, such as economic crises like the Housing Crisis of 2007, laws that change within the panel time that are made on a state level (income tax, property tax, abortion rights, etc) and states that experience higher natural disasters than others during the panel time. Thus, our state-year fixed effects adjust for variables that vary on a state-level, that also change within our observation period. These are important to adjust for since people might move as a result of a new state-level law, higher number of natural disasters, etc.

Lastly, our error term is the difference between the actual (population) values of the dependent variable, *Migration*$_{ct}$, and the predicted values of the dependent variables based on the regression model, the estimate of *Migration*$_{ct}$.

The interpretation of our β_0 estimate is migration, as a percent of the total population, that is not described by changes in precipitation or whether the county is in a drought for county 1 in year 1 (2000). Our β_1 estimate provides the predicted change in migration as a percent of the population with a one inch increase in precipitation in counties that are not in a drought ~~and~~ *at* time *t-1*, holding all other variables constant. Our β_2 estimate is the predicted change in migration, as a percent of the population, in counties that are in a drought, holding all other variables constant (not explained by ~~*current*~~ precipitation). Our β_3 estimate provides the predicted difference in migration, as a percent of the total population, between countries that are not in a drought and counties that are in a drought with a one inch increase in precipitation in time *t-1*. The predicted change in migration, as a percent of the population, that is explained purely by precipitation for counties that are in a drought is $\beta_1 + \beta_3$.

While county and year fixed effects control for several omitted variables that may affect the coefficient estimates, there could still be omitted variable bias in our model. We would have omitted variable bias if there is a variable not included in our model that is correlated with the independent variables (precipitation or drought) and correlated with the dependent variable (migration). A variable that would not be picked up by our model is something that cannot be explained by randomness in precipitation. An example of this would be cloud seeding, which, in practice, we only observe in areas with low rainfall (West coast states). Our estimates would be biased since cloud seeding occurs in counties with low precipitation (so the treatment of cloud seeding is not random) and it changes the random precipitation (so precipitation is no longer random).

We predict that cloud seeding would occur in areas that have increasing agricultural output (counties are likely more inclined to use cloud seeding to keep up with increases in agricultural output) and that cloud seeding is positively correlated with population (we can infer that agricultural output increases population migration). Further, cloud seeding is likely negatively

correlated with precipitation since counties with low precipitation would use cloud seeding to receive higher precipitation. Thus, the bias of omitting cloud seeding in our regression would be negative, and we would underestimate our coefficients. We are not especially worried about this potential omitted variable since cloud seeding rarely occurs in the U.S. during our observation years.

Variables that are not adjusted by county fixed effects are variables that vary across time and across counties. Examples of variables that would not be captured by county fixed effects are county-level laws/ policies that change within the observation period, unemployment rates in a county, county Gross Domestic Product (GDP), businesses entering and leaving a county, etc. The variables that we are most concerned about are county unemployment rates and county GDP. We will include these variables if we find that the regression estimates vary when unemployment rates and county GDP are added into the regression. There is no clear inference for the correlation between county GDP and precipitation, or county unemployment rates and precipitation.

Another important omitted variable that will not be captured by county fixed effects is county agricultural output, since county agricultural output varies over time. Higher agriculture output is likely to have a positive correlation with migration and a positive correlation with precipitation. Thus, if we omit agricultural output, we may have a positive bias on our estimates and, as a result, overestimate our estimates. the true relationship

Further, county house prices will not be captured by county fixed effects since house prices vary over time. Higher county house prices are likely positively correlated with migration, since house prices typically increase with increases in the population. However, there should be no correlation between county house prices and precipitation, since precipitation is random. Thus, we cannot infer the sign of bias. While we may not know the sign of the omitted variable bias, we would like to include this control in our regression since house prices likely have an impact on migration.

Precipitation is commonly used as a random variable since it is reasonably assumed to be, if not fully, random. Thus, it is used in many different studies as an instrumental variable. Further, the definition of a drought is debatable. Therefore, we have defined drought as five years of continuous decrease, year over year, for five consecutive years. Expressing migration as a percent of a population, which is already a proportion between 0 and 1, on a logarithmic scale makes little sense, and needlessly complicates the interpretation of the effects of the variables in the model. Thus, we leave our dependent variable in its original form. Expressing precipitation as a logarithmic function could be plausible; however, we would like to see the change in population migration, as a percent of the total population, for each one inch change in precipitation. There is no evidence to imply that the effect of rainfall on precipitation should be quadratic from the scatter plot; thus, we use a linear effect model.

GOOD JOB   A—

1) clearly define MIG   (show formula you're using)

2) come up w/ better term for "change in migration, as a percent of total population"

3) precip subscript .