Climate Change in MN

Ivana

4/20/2021

A basic introduction:

Climate change is happening globally and locally. In this project, I will be exploring how the climate is changing in Minnesota through a few basic visualizations and through modelling the severe weather events that Minnesota is prone to. Namely, this is tornadoes.

Tornadoes are extremely difficult to predict. But, they are expensive and difficult to prepare for. I pulled federal climate data from Minnesota to try and see how climate trends might help predict the number of tornadoes. With this information, it would be easier to prepare for property damage or tornado injuries.

General setup:

Data tables were read-in and tidied.

fullTbl has 1526 rows with 7 variables: Year, Month, Precipitation, MinTemp, AvgTemp, MaxTemp, and PDSI. fullTbl covers years from 1895 onwards.

Next, I need to tidy the Tornado table.

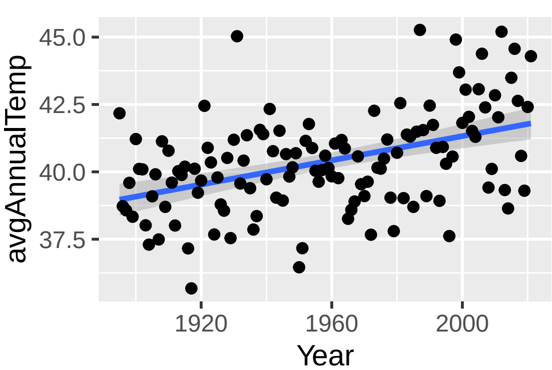
TornadoTbl has 1058 rows and 8 variables: Year, Month, Mag, CrD, PrD, Dth, Inj, and County. TornadoTbl includes tornado data from 1950 onwards.

The two tibbles are merged to produce fullTornadoTbl. fullTornadoTbl has the variables from fullTbl and TornadoTbl, but the years start from 1950 onwards to 2021.

How is the temperature changing in Minnesota? And an introduction to variables:

Let's explore some of the trends in the variables of the imported datasets.

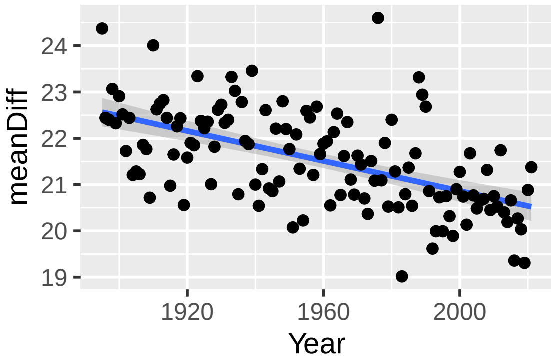
First, this is a basic visualization of the average monthly temperature by year:



There appears to be a slight increase in the average temperature by a bit over 2 degrees Fahrenheit since 1895.

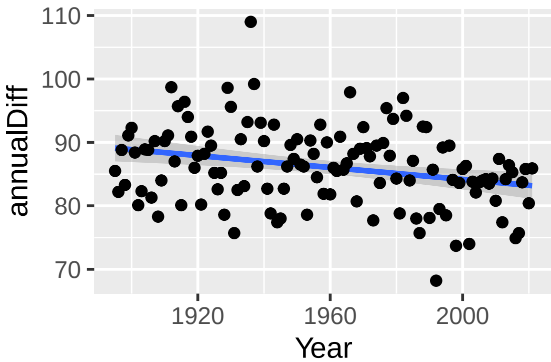
Climate change is not just a matter of warmer weather on average - it matters how the temperature is increasing and how this may impact other climate events.

This next graph is a basic visualization of the mean difference in monthly average maximum and minimum temperatures since 1895.



It appears that monthly temperature variation within months will decrease, on average, by about 2 degrees Fahrenheit.

It's not enough to just look at the average temperature: temperature fluctuations also matter. To determine if annual temperature fluctuations are also growing, this graph plots the annual difference between the average annual min temp and average annual max temp:

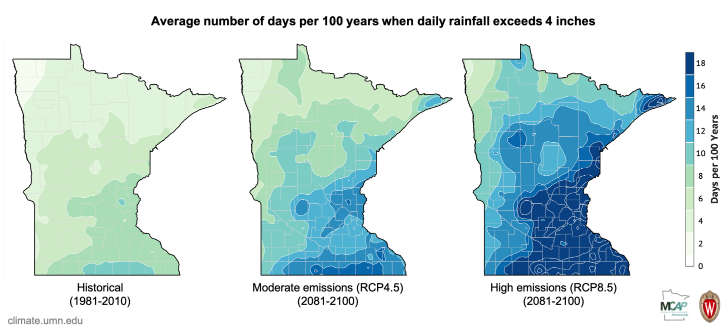


It appears that annual average temperature fluctuations among the 12 months will decrease by about 5 degrees Fahrenheit.

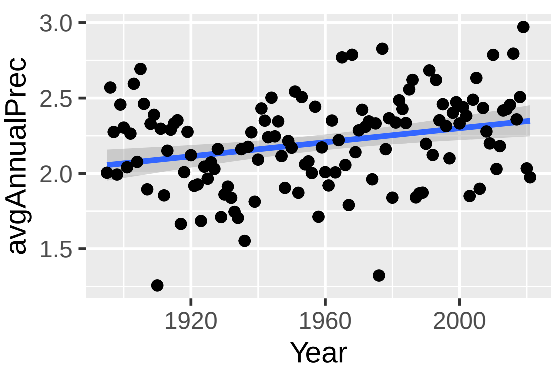
How is Precipitation changing in Minnesota?

According to the UMN's page on the "Minnesota Climate Adaptation Partnership", "extreme events, like flooding, drought, and heatwaves, will likely become more frequent and more intense with climate change."

One phenomenon that has been occuring in Minnesota is an increase in "mega-rain" events which is where at least 6 inches of rainfall accumulates over 1000 square miles within a day.



This plot is the average monthly precipitation per year:

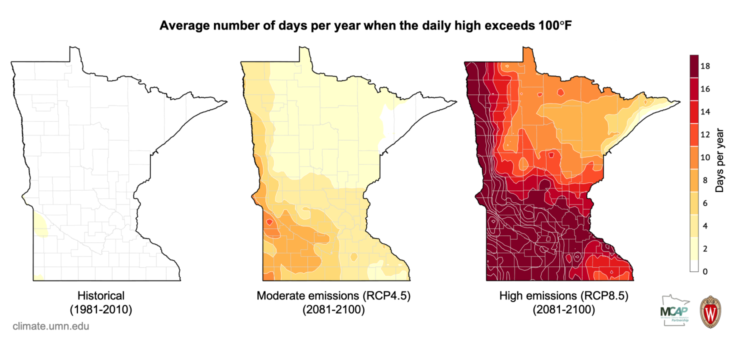


It appears that the average annual monthly precipitation has increased by a bit over a quarter inch.

According to the UMN's climate article, there will be an increase in mega-rain events. Historical trends show this may be accompanied by a slight increase in annual monthly precipitation as well.

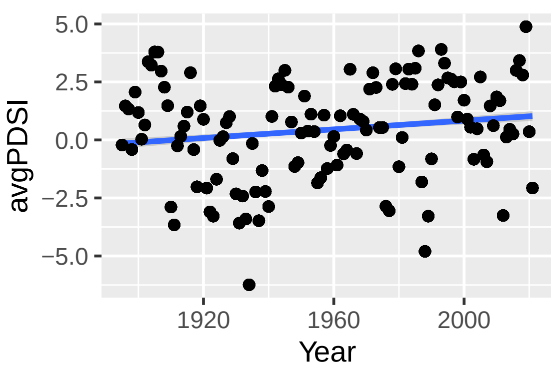
What about Heatwaves and Droughts in Minnesota?

Next, heatwaves, "defined as a series of at least 4 days with air temperatures that would normally only occur about once every 10 years", "are projected to become more frequent. . . [,] stronger[, and] longer". Heatwaves keep track of abnormally hot days, but the temperature cutoff for an abnormally hot day will increase as the average temperature increases.



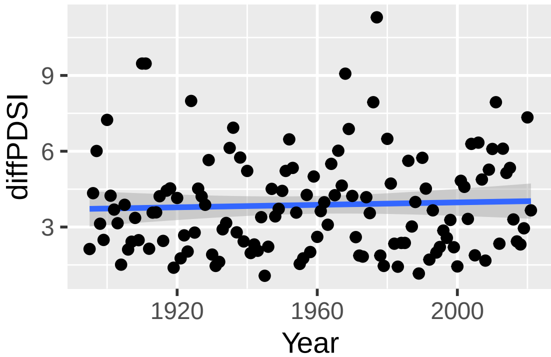
Another measure of climate change in Minnesota is through the drought index. The PDSI is a "standardized index based on a simplified soil water balance and estimates relative soil moisture conditions. The magnitude of PDSI indicates the severity of the departure from normal conditions. A PDSI value>4 represents very wet conditions, while a PDSI < -4 represents an extreme drought."

This plot visualizes the PDSI index by year:



It appears that the average PDSI value started off at about neutral (0). It has since increased to about 1, on average, which indicates wetter conditions overall.

How about annual fluctuations in the PDSI index?



There appears to not be much fluctuation in maximum versus minimum PDSI index by year: just a slight increase. This indicates that fluctuations between the PDSI index have remained relatively constant by year and may continue to do so. Combined with the overall increase in PDSI index, this may indicate a fairly consistent and wetter climate in the future.

Tornadoes

“Much about tornadoes remains a mystery. They are rare, deadly, and difficult to predict, and they can deal out millions or even billions of dollars in property damage per year.” Minnesota is part of Tornado Alley, so tornadoes are a very relevant problem here.

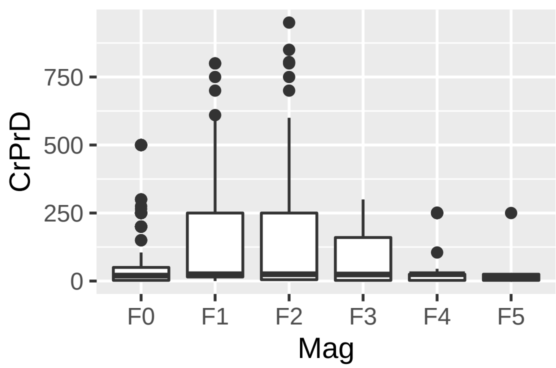
First, a foreword on this subset of data.

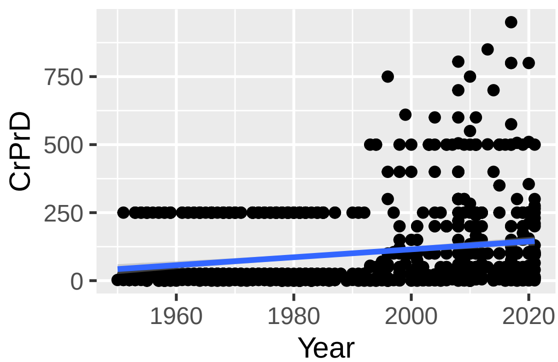
“About 1,200 tornadoes hit the U.S. yearly. Since official tornado records only date back to 1950, we do not know the actual average number of tornadoes that occur each year. Plus, tornado spotting and reporting methods have changed a lot over the last several decades, which means that we are observing more tornadoes that actually happen.”

To help combat this, I removed tornadoes without any crop damage, property damage, injuries, or deaths in my tibble. In doing this, the way these tornadoes are detected is less dependent on detection by increasingly advanced weather technology.

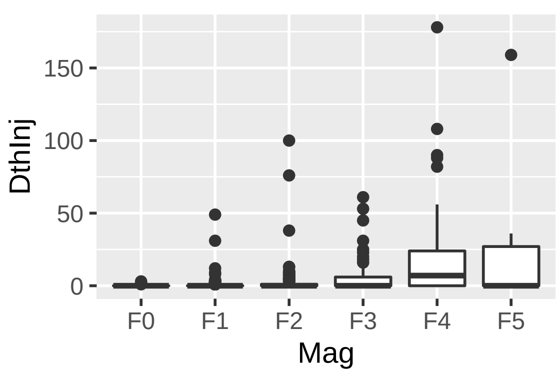
TornadoTbl is a tibble of tornadoes in Minnesota that have caused some harm or damage since 1950.

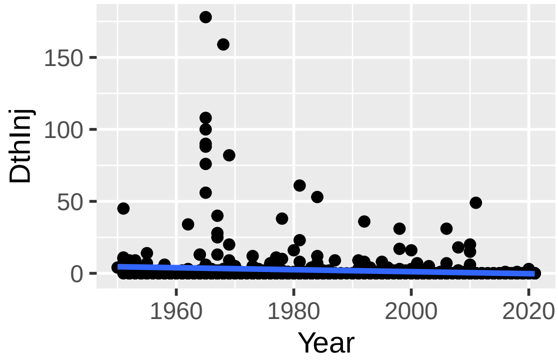
Plots of property damage and crop damage in damaging tornadoes:





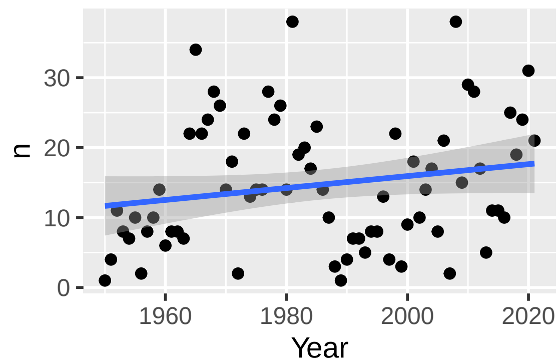
Plots of injury or death in damaging tornadoes:





Clearly, tornadoes are expensive and dangerous. Since they are difficult to predict, they are definitely worth studying. With increasing knowledge of tornadoes, financial losses and deaths and injuries can be reduced. This weather phenomenon will be tested in various models to determine which machine learning model may be used for predicting the number of tornadoes per year as the climate changes.

This is the number of damaging tornadoes per year:



The number of damaging tornadoes per year has increased from about 12 to 16. Again, definitely a topic worth studying.

Linear model:

Tornadoes are a result of multiple weather factors. They come from strong thunderstorms, caused by surface heating, moisture, or "areas where warm and cold, or wet and dry air bump together."

Since tornadoes are difficult to predict, I will attempt to create a model that predicts the number of tornadoes per year using a selection of climate variables.

My tibble, fullTornadoTbl, contains some variables that may not independent enough (such as maxTemp and minTemp), so I will not use all of them. The ones I will be using include Year, Month, Precipitation, AvgTemp, County, and PDSI.

I finalized fullTornadoTbl for running in a linear model and named it "modelTbl".

First, create training and testing datasets:

```
tidymodels_prefer()

set.seed(54321)
tornado.split <- initial_split(modelTbl, prop=0.8)
trainTbl <- training(tornado.split)
testTbl <- testing(tornado.split)
```

Then, run the model:

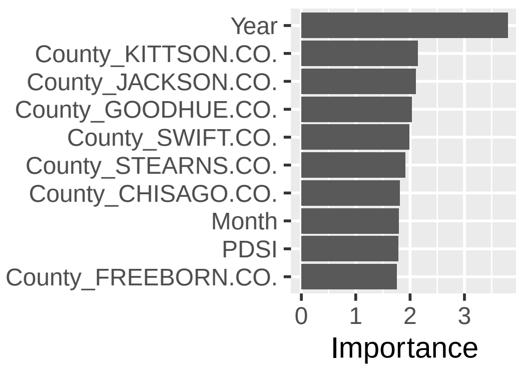
```
recipe <-
  recipe(formula=numTornadoes ~ ., data=trainTbl) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_normalize(all_nominal_predictors())
lm.model <- linear_reg() %>%
```

```
  set_engine("lm")
lm.wflow <- workflow() %>%
  add_recipe(recipe) %>%
  add_model(lm.model)
lm.fit <- fit(lm.wflow, trainTbl)
tidy(lm.fit)
augment(lm.fit, new_data = testTbl) %>%
  rsq(truth = numTornadoes, estimate = .pred)
augment(lm.fit, new_data = testTbl) %>%
  rmse(truth = numTornadoes, estimate = .pred)
```

The r^2 value is 0.0523. It's pretty bad!

rmse is 8.83. Also a pretty high value.

Variable importance:



It seems the number of tornadoes, per year, depends on what year it is. Then, the importance drops off to a selection of counties, the month, the PDSI, and another county.

A linear model is a very basic model, but it will provide a baseline for comparing r^2, rmse, and variable importance in more advanced models.

In addition, it is ideal to have independent variables for a linear model. It would be better to use a model like random forest or boosting. A random forest model does not assume that variables are fully independent. I will also test a boosting model because it tends to perform similarly to a random forest. I will then decide which model is the best for future prediction of numTornadoes.

Random Forest Model for predicting numTornadoes per year:

The training and testing tibbles are reused from above for better rmse, r^2, and variable importance comparisons.

I find the optimal parameters mtry, min_n, and trees using 10-fold cross validation.

```
set.seed(12345)
tissue.folds <- vfold_cv(trainTbl, v = 10)
```

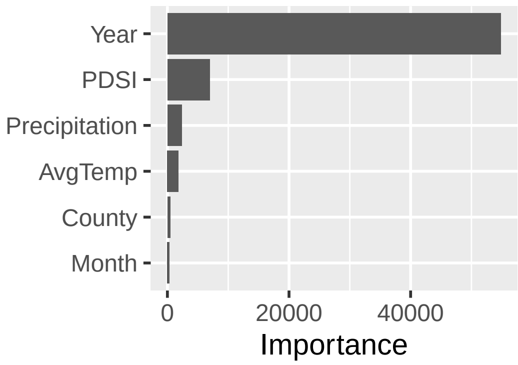
```
best.penalty
```

```
## # A tibble: 1 x 4
##    mtry trees min_n .config
##   <int> <int> <int> <fct>
## 1     6   500     1 Preprocessor1_Model105
```

mtry ranged from 1-6, min_n ranged from 1-10, and trees ranged from 100-500. Using rmse as the metric, the best values are mtry = 6, trees = 200, and min_n = 1.

r^2 = 0.960, rmse = 1.81

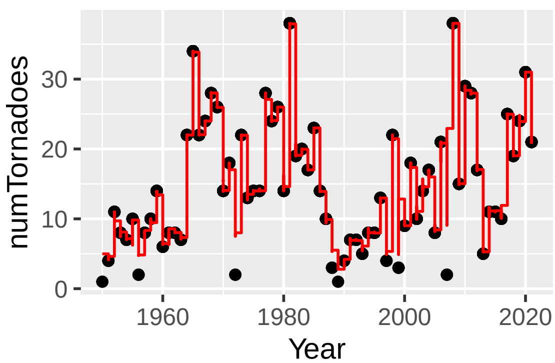
This model is a significant improvement from the simple linear model as the rmse is much lower.



The most important variable in predicting numTornadoes is Year. Unlike with the linear model, county factors drop off the list completely.

The next closest important variable is the PDSI. Other variables are close to having no importance.

A plot of the model on the original set of data:



Most points are accounted for. The model appears to have more trouble modelling years with a low number of tornadoes.

Next, a boosting model will be utilized.

Boosting Model for predicting numTornadoes

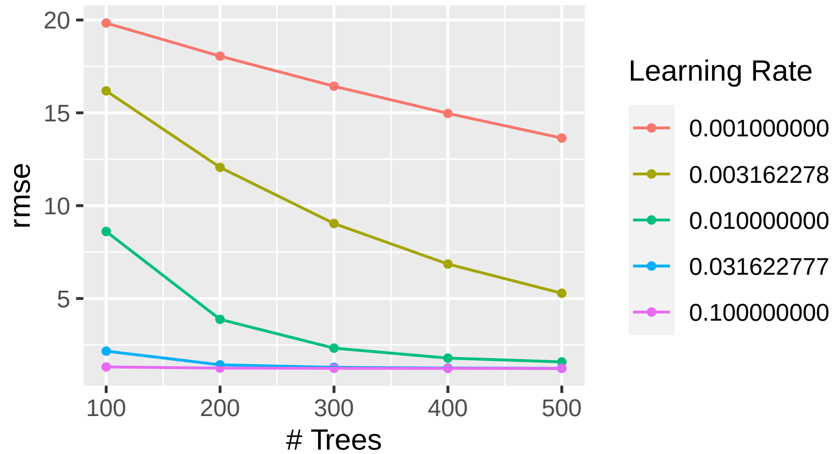
Next, I create a boosting model where I tune the number of trees and the learn_rate. I reuse my folds from the random forest model to compare the rmse (and r^2 values) between the three models.

I set trees and learn_rate as tuneable parameters in xgboost_spec. The range of trees to tune from is the same as for the random forest model.

```
xgboost_recipe <-
  recipe(formula = numTornadoes ~., data = trainTbl) %>%
  step_dummy(all_nominal_predictors())
  #step_zv(all_predictors())#add step normalize? what is step_zv?

xgboost_spec <-
  boost_tree(trees = tune(), learn_rate = tune()) %>%
  set_mode("regression") %>%
  set_engine("xgboost")

xgboost_workflow <-
  workflow() %>%
  add_recipe(xgboost_recipe) %>%
  add_model(xgboost_spec)

set.seed(24725)
xgboost_tune <-
  tune_grid(xgboost_workflow,
```

```
            resamples = folds,
            grid = grid)
```

```
autoplot(xgboost_tune, metric = "rmse", select_best=TRUE)
```



For learning rates of .1 and .03, rmse is minimized at about 100 trees with a comparable further decrease in rmse up to about 300 trees.

```
best.param
```

```
## # A tibble: 1 x 3
##   trees learn_rate .config
##   <int>      <dbl> <fct>
## 1   500     0.0316 Preprocessor1_Model20
```

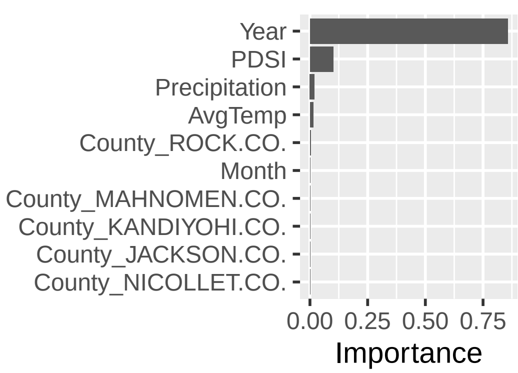
The best param that minimizes rmse has 500 trees with a learning rate of 0.0316.

rmse = 2.593176

r^2 = 0.9162263

In this case, the random forest model performed better. This model has a higher rmse value and a lower r^2 value. But, this model is still much better than the linear model; the rmse value is much lower.

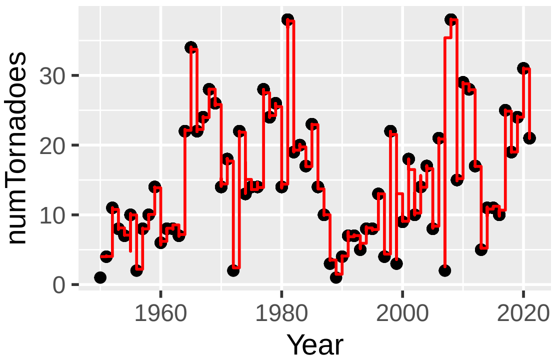
Importance image of variables:



Compared to the random forest model, Year is also an overwhelmingly important variable for predicting numTornadoes. PDSI is also the next most important variable, but not by much. All other variables are also negligible.

All three models heavily prioritize year in predicting numTornadoes. PDSI is included as a variable but with less importance. The random forest and boosting model do not put as much weight on variables besides Year and PDSI.

A plot of the model on the original data:



It appears that almost every datapoint on the graph is accounted for by the model. Although the random forest model performs better, this boosting model is better at mapping to years with a smaller numTornadoes value.

Conclusion

The random forest model is the most effective model for predicting numTornadoes using rmse as a metric. Moving forward, this model may be used to predict the number of tornadoes in a year.

However, some flaws with the model raise some questions. The most important variable in predicting the number of tornadoes per year is Year. PDSI, the next most important, was only slightly useful in predicting numTornadoes.

Part of the issue with the data is how the climate variables may not represent the necessary weather patterns that form tornadoes in enough detail. Since tornadoes form from clashing temperature or moisture fronts, better data would include things like the weekly number of tornadoes; the weekly range in temperature; PDSI; and data on longer, years-long weather patterns like El Nino/La Nina.

The number of tornadoes detected per year is also potentially reliant on what percentage of rural land is utilized for agriculture or living. A generally higher population density could result in more recorded human injury or crop/property damage. Data that is independent of population density would be useful. (Then again, no county was particularly significant in predicting numTornadoes.)

Although my data excluded tornadoes without any physical or monetary damage, weather technology also may skew how tornadoes are detected. Data that is independent of technological advances would be useful.

Another issue with modelling numTornadoes for an entire state is that this may result in irrelevant weather data for predicting numTornadoes. It may be better to rerun these models with weather data from just certain counties with a high number of tornadoes. Even so, no county stood out as particularly significant in predicting numTornadoes.

Predicting tornadoes is difficult, but my goal was to look at general climate trends to observe what might help people predict how climate change might influence the number of tornadoes. Then, things like the estimated cost of property damage or human health risks could be better anticipated.

As climate change continues, understanding how severe weather events will progress in Minnesota will become increasingly important.

Sources:

https://www.ncdc.noaa.gov/stormevents/listevents.jsp?eventType=ALL&beginDate_mm=01&beginDate_dd=01&beginDate_yyyy=1950&endDate_mm=01&endDate_dd=31&endDate_yyyy=2021&county=ALL&hailfilter=0.00&tornfilter=0&windfilter=000&sort=DT&submitbutton=Search&statefips=27%2CMINNESOTA

https://climate.umn.edu/our-changing-climate/extreme-events

https://www.drought.gov/data-maps-tools/us-gridded-palmer-drought-severity-index-pdsi-gridmet#:~:text=The%20PDSI%20is%20a%2

https://www.npr.org/2021/12/13/1063676832/the-exact-link-between-tornadoes-and-climate-change-is-hard-to-draw-heres-why

https://scijinks.gov/tornado/

https://www.nssl.noaa.gov/education/svrwx101/thunderstorms/

https://www.nssl.noaa.gov/research/tornadoes/