Tumbling with Tornadoes:

A Tornado Simulation Project

Jesus Olivera and Pujita Ravichandar

DAV 5300: Computational Math and Statistics Final Project

Yeshiva University Katz School of Science and Health

Abstract

This project explores historical tornado tracks data from the National Oceanic and Atmospheric Association (NOAA), the National Climatic Data Center (NCDC), the Storm Prediction Center (SPC), the National Weather Service (NWS), and the Homeland-Infrastructure Foundation Level Data (HIFLD). The goal of this project is to model the magnitude of tornadoes, create a mathematical model of the angular velocity, and model the economic losses caused by tornadoes of various magnitudes in the Fujita Scale. The best resulting logistic regression model for the magnitude prediction yielded an accuracy score of 0.707, the angular velocity simulations followed the expected relationship trends described by the laws of circular motion, and the linear regression model used for the economic loss model yielded an RSME (Root Mean Square Error) of 18.5 and an R2 (Coefficient of Determination) of 0.31. This study is a solid introduction into using data science, computational math/statistics, and machine learning in the field of storm and weather studies.

Introduction

About Tornadoes

Tornadoes can be briefly defined as “violently rotating columns of air” that extend from a thunderstorm to the ground. Tornado systems can be hard to detect because air is invisible. They start to become more visible when condensation forms and when they pick up other dirt and debris.

These storms are formed when two great air fronts, often one hot and moist and one cool and dry, travel against each other. The formation of these storms occurs within many processes that happen on the storm-scale around the mesocyclones Some studies of these storms suggest that the temperature difference between the fronts is what determines the severity of the storm, after the mesocyclone forms. Although, some strong tornadoes have formed without a large temperature difference. There are still many unknowns related to the specifics of tornado formation, and much more is left to be researched.

Tornadoes happen not only in North America, but across the globe in Australia, South America, Asia, Africa, and South America. Bangladesh and Argentina have the two highest concentrations of tornadoes, outside of the United States. The U.S. has about 1,200 tornadoes annually. Tornado records only date back to 1950, and tornado observation methods have changed greatly over the years. The recent formal observations of these storms make it difficult to provide an accurate number for how many tornadoes to expect per year.

The area where most tornadoes occur in the U.S. is nicknamed “Tornado Alley”. This is a region in the central continental United States; most of the states that fall into this category are in the Midwest. This region is faced with an updraft of warm moist air from the Gulf of Mexico and a downdraft of cool dry air from the Rocky Mountains, creating the perfect atmosphere for tornado formation. However, this can be misleading as not all tornadoes occur in this region. There have been tornadoes recorded in all 50 states, as well as Puerto Rico. Additionally, tornado season varies per region. In the U.S., tornado season shifts toward the Southeast in the cooler months and toward the South, the Central Plains, and the Midwest in the early summer.

There are many aspects that go into calculating the strength of a tornado. The strength is not only dependent on the massive scale of the tornado, but also how much damage it caused, along with some other metrics. The Fujita Scale (F-Scale) and the Enhanced-Fujita scale (EF-Scale) are the conventions used to measure the damage of tornadoes. The F-Scale is the original scale, but the EF-Scale takes into account more variables, not just the wind speed, when dictating a magnitude value. There is a conversion from the F-Scale to the EF-Scale to preserve and continue the use of the existing historical data.

Our Motivations

Tornadoes are one of the most violent and damaging atmospheric storms we experience. There is still much that is unknown about them. They are rare, deadly, difficult to predict, and can cause millions of dollars in damage per year. The U.S. has the highest concentration of tornado events in the world. Understanding how these storm systems work and using data to help improve prediction methods is integral in gaining more knowledge about these great storms. This exploration can be used to help tornado forecasts, warnings, damages, and other natural disaster damage control.

About the Dataset

This dataset contains information about historical tornado tracks in the continental United States, Alaska, Hawaii, and Puerto Rico from 1950 to 2015. The data is from the Homeland Infrastructure Foundation-Level Data (HIFLD). The values are gathered from the Storm Prediction Center (SPC), Nation Centers for Environmental Protection (NCEP), and the National Oceanic and Atmospheric Association (NOAA). Link to the dataset: https://hifld-geoplatform.opendata.arcgis.com/datasets/historical-tornado-tracks

A further description of the features in this dataset can be found in the Appendix A.

Exploratory Data Analysis (EDA)

The dataset used in this analysis contains 60,114 rows. Each row represents a tornado instance. The original data frame contains 13 columns. For this analysis only magnitude (mag), injuries (inj), fatalities (fat), economic losses (loss), length (len) and width (wid) will be used. The dataset includes values for tornadoes category F0 to F5 ranging from 1950 to 2015.

The insights derived in this EDA section are only a part of the formal EDA done on this dataset. The details of this process can be found in the Python code for this project. The EDA uncovers that as the magnitude of a tornado increases, its width and length also increase. Most tornadoes are in category F0 and F1, with the least number of tornadoes being in category F4 and F5 respectively. Further exploration showed that Texas experienced the highest number of tornadoes from 1950 to 2015, and Oklahoma and Alabama experienced the highest number of tornadoes of magnitude F5 during this period. Additionally, most tornadoes occurred during the afternoon and evening from March to August, fitting the research in the Introduction. This is illustrated in Figure A.

Figure B shows the change in tornado occurrences over time. A possible reason for this effect might be attributed to the development of new tracking and tornado measuring technologies. Figure C shows the average economic loss caused by each magnitude of tornadoes. The trend uncovers that, on average, tornadoes of category F5 cause the most economic losses.

An analysis of the outliers was conducted to further explore the trends in the data. This led to the conclusion that all of the outliers represent real instances of tornadoes with extreme values in the metrics that were measured. Detailed information about these outliers can be found in the referenced Python script for this project.

Linear Regression Models

Logistic regression

The correlation analysis uncovered the numerical features that inspired the logistic regression that is described in this section. The technicalities of this approach are documented in the Python script for this project. This logistic regression meets all the required assumptions and analyzes tornado data and strives to establish a dependence of a tornado's magnitude on the width and length. The prediction variables used from the data are width and length, with a response variable of tornado magnitude. In this case, we will be applying a logistic regression to a multiclass problem (magnitudes F0 - F5). Each class will be treated as a binary classification problem in the "one vs all" method.

The first step is to standardize the features by removing the mean and scaling the unit variance, so the machine learning estimators do not behave incorrectly when the individual features do not have a standard normal distribution. Following this, the data is equally split into a "train" and "test" set where 30% of the data is used for both. After that is completed, the model is run. The coefficient of determination R^2 of our logistic regression model score is 0.67 which provides a sturdy leg to stand on with a bit of room for improvement (1 is the best possible score). The specific results can be found in the (Appendix C).

Results

The regression best prediction was on tornadoes with magnitude of F0. For the remaining tornadoes magnitude, the regression precision was under 50% which indicates the regression might classify the majority of those tornadoes as tornadoes with different magnitudes. See Appendix C for the full Classification Report.

Random Forest

The Random Forest Regression is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averages to improve the predictive accuracy and control over-fitting. All the required assumptions for this regression were met. The dataset for Logistic Regression and the Random Forest Regression is identical.

The primary step is to establish the classifier and measure its baseline accuracy. The initial model's accuracy was 0.67. We calculated the value utilizing "cross_val_score" from the sklearn package where the cv parameter was 5; this further split the training set into 5 subsets called folds. Then, the model can be fit, and the accuracy is tested. The accuracy increased from 0.6712 to 0.6752.

To optimize the model, the parameters of the classifier are adjusted. To optimize the function that measures the quality of the split, a parameter that establishes the maximum depth of the tree was added, and the number of features to consider when looking for the best split was adjusted. After that, the model was fit again, providing a new accuracy score.

Results

The accuracy for this model increased from 0.6752 to 0.707. The overall model was improved from its baseline and first iteration. The specifics of this model can be seen in Appendix C.

Models and Simulations

The logistic regression shows that numerical features like radius can be used in relation to magnitude. These attributes can be used in a more physical mathematical simulation outlined in this section. There are two phases in the modeling and simulation of the angular velocities of tornadoes. The first one is a model of the angular velocity based on the radius and designated velocity measurements by magnitude. The second also models the angular velocity but includes probability calculations regarding the radius. The details of these models are in the upcoming sections.

Mathematical Model

Physics of Tornadoes

Tornadoes can be briefly described as rotating columns of air. The air masses that form tornadoes are rotating and can be described by the laws of centripetal motion. With regards to rotation, these laws state that the smaller the radii (from the center of the tornado), the faster the air masses spin. (excluding the "eye of the storm") Understanding the fundamentals of rotational motion is integral to studying and simulating the rotational metrics of this natural phenomena.

The three main metrics are used in our simulation are radius, wind speed (velocity), and angular velocity. For this study it is assumed that tornadoes are just spinning masses of air. Weather is a very complicated system, and this assumption is far from the physical reality; however, for creating an initial simulation this assumption is useful to simplify the system. The following formula describes the relationship between the radius (r), wind speed (v), and angular velocity (ω).

$$v = r\omega \text{ or } \omega = v/r$$

(Formula I)

It is important to note that the radius values are measured with respect to the center of the tornado. Although we are assuming uniform circular motion, the angular velocity at different distances from the center of the tornado

vary. These concepts will be the basis of how the angular velocity is simulated for each magnitude tier in the F-Scale.

Process

This section outlines the process taken to model the angular velocities for tornadoes of magnitude F0 through F5. The basic steps include transforming the data, creating matrices of the angular velocity and radii values for each magnitude, and plotting the resulting graphs.

The relevant attributes from the dataset are the magnitude (‘mag’) and radius (‘wid’). The first step is to isolate these variables and separate them by magnitude. The magnitude is listed by the Fujita Scale (F-Scale), which provides a range of wind speeds for each magnitude value. These are listed in Figure D. This range of wind speeds is what provides the velocity information. With this velocity and the given radii, there is sufficient information to calculate the angular velocity.

For this calculation, there are two types of arrays used: one of the radii and one of the velocities. These are used to create matrices of the angular velocity. For each magnitude of tornado, the minimum and maximum radii are identified. Then an array of evenly spaced radii values within these bounds is created. Another similar array of evenly spaced velocity value is created using the lower and upper bounds identified by Figure D.

From these arrays, two matrices are created to plot our model: one of angular velocities and one of radius information. The radius and the velocity arrays are used to calculate a matrix of angular velocities. The radius identifies the row values, and the velocity identifies the column values. The angular velocities are calculated using Formula I. The radius matrix is the same dimension as the angular velocity matrix, where each row is the radius array. This matrix is necessary to plot the calculations. The radius matrix is plotted against the angular velocity matrix to create a plot of the angular velocities (y-axis) vs. the radii (x-axis) for each magnitude of tornado (F0 - F5). These plots are shown below in Figure E.

Results and Use Case

The relationship shown by the plots in Figure E fit what is expected. Formula I highlight that the angular velocity and the radius are inversely proportional, which translates to what is seen on the models. It is interesting to see that the F2 - F4 tornadoes appear to be more linear than the F0 and F5 plots. This is because the radius range for F0 (0-73 mph) and F5 (>261 mph) are much larger than the other magnitudes. See Figure D for more details on the velocity ranges per magnitude. The bounds for the F5 matrix were confined by the maximum radius tornado seen in this magnitude: 1760 ft. These results are a good basis to expand our modeling and understanding of tornado angular velocity, which in turn can provide information about the force that the tornado applies on objects like infrastructure, wildlife, etc.

Probabilistic Simulation

Process

This section outlines the process taken to add probability to the model of angular velocities for tornadoes of magnitude F0 through F5 as seen in the previous section. The basic steps include creating probability density histogram for the radii of each F-Scale magnitude, creating matrices of the angular velocity and radii values for each magnitude, and plotting the resulting graphs.

The same attributes, magnitude (‘mag’) and radius (‘wid’) are used in this probabilistic simulation. The dataframes of width are isolated by magnitude, similarly to above. The primary step is to convert the radii values into miles to match the velocity units. Following this, the probability density histogram is created by summing the total tornadoes in each bin and dividing by the total number of tornadoes in the respective magnitude tier. This

results in the probability that a tornado falls within a particular range of radii. The probabilistic histograms are shown below in Figure F.
The probabilities identified by these plots are used to simulate tornado radii. Using the random package in Python, 50 tornado radii are simulated for each magnitude tier. These tornadoes are then used to plot angular velocity vs. radius plots to build upon the plot in Figure E.

The two matrices that are created to plot our model are reevaluated to account for the radius probabilities. The radius array used to calculate a matrix of angular velocities, is based on the simulated tornado radii. The velocity ranges are the same as before. However, they are split into 50 steps, to calculate and plot the angular velocity matrices. As before, the angular velocities are calculated using Formula I. The radius matrix is plotted against the angular velocity matrix to create a plot of the angular velocities (y-axis) vs. the simulated radii (x-axis) for each magnitude of tornado (F0 - F5). The new plots are shown below in Figure G.

Results and Use Case

The relationship shown by the plots in Figure G follow a similar pattern to the plots in Figure E, following expectations. The inverse proportionality between the radius and the angular velocity is still seen on these models. Similarly, to the previous model, the F2 - F4 tornadoes appear to be more linear than the F0 and F5 plots because of the smaller radius range. See Figure D for more details on the velocity ranges per magnitude. The bounds for the F5 matrix were confined by the maximum radius tornado seen in this magnitude: 1760 ft.

These results provide interesting insight on what radii are most likely to occur, along with the expected angular velocities. The most probable radius ranges are blocked off accordingly and may suggest some sort of physical phenomena that gravitate towards a particular radius range. This information is useful when predicting the impact and reach that a tornado has on infrastructure, wildlife, etc.

Economic Loss Model

The mathematical simulation provides information on the angular velocity. This is useful for predicting damages caused by a tornado, providing motivation to explore the economic attributes in the data. The economic model meets all the required assumptions and was designed with the purpose of providing insight to community leaders, business and individuals regarding building more resilient budgets that plan for possible economic losses due to damages caused by tornados.

The first step in building the model consisted of selecting and standardizing the relevant numerical features. In this case, the model uses the width, length, fatalities, injuries, and economic losses from each tornado. The predictor of the model (y) is economic losses, and the response variables (x) are the remaining variables used.

The next step consists of dummifying the categorical feature, the tornado magnitude (F0 - F5), and creating a new dataframe of the appropriate values. To better understand the distribution of economic losses represented, a histogram was used. This plot is shown in Figure H. The histogram illustrates that the data is strongly skewed. Because of this, the median will be used as a baseline for the model instead of the mean.

Creating the Model

The data was split between train and testing and a baseline was calculated.

```
Baseline RMSE:   22.3109329267541
Baseline R2:     -0.003842043891185032
```

The square root of the variance of the residuals (RMSE) indicates the absolute fit of the model to the data and how close the observed data points are to the model's predicted value. In this case, the value is very high, indicating that the model does not fit the data properly. The R^2 measures the strength of the relationship between the model and the dependent variable (economic loss) on a 0 – 100% scale. The larger the R^2 the better the

regression model fits the observations. In this case we observe a negative R^2 which tells us that our model is not a good one.

The next step is to develop a linear regression and compare the results with the baseline previously established. The data is fitted to the model and new results are obtained.

```
Linear Reg 1 - RMSE:                    18.495819925158944
Linear Reg 1 - R2:                      0.31011437291787824
```

In the second iteration, both underline metrics showed some improvements, but the model is still far from being a good one.

To further improve the model a third iteration is performed. In this case a linear regression using Lasso regularization is performed. This method minimizes the sum of squared errors, with a bound on the sum of the absolute values of the coefficients. Basically, the values are shrunk towards a central point, like a mean.
After the data is fitted the results of the model are obtained.

```
Linear Reg 2 - RMSE:                    18.37598455888654
Linear Reg 2 - R2:                      0.31902502157457124
```

As shown in the results above, better results were obtained, but the model can still be further improved.

Conclusion

This exploration into historical tornado track data strives to uncover interesting insights about the magnitude and economic impact of these natural disasters. Through the exploratory data analysis, there are many patterns that do arise with regards to tornadoes, including the regions with the greatest number of tornadoes ("Tornado Alley"), the number of tornadoes per magnitude, and the damage caused by each magnitude of tornadoes, among other analyses.

The first goal is to create a logistic regression with the vision of formulating an algorithm to predict the F-Scale magnitude of a tornado given the width and the length. The optimized random forest model proved to be the best at categorizing tornado magnitudes based on their widths and lengths. We improved our accuracy score from 0.6752 to 0.707 after updating the parameters of the classifier. The model's precision returned the best results when predicting tornadoes in magnitude tiers F0, F1 and F3.

From here, simulations regarding the angular velocity can be calculated given the radii and the velocity ranges for the tornadoes in each magnitude tier. The initial model followed the basic laws of circular motion and illustrated an inverse relationship between the radius and angular velocity of the tornadoes. This model was expanded by including the probabilities of each tornado falling into a particular radius range. This resulted in a similar graph but with the most probable angular velocities highlighted. This could provide some insight about the mechanics of the tornado vortices and why certain radii are favored compared to others.

The process of trying to derive a prediction of economic losses based on the variables used in the analysis shows the complexity of the prediction. The third iteration of the linear regression using a Lasso regularization returned the best results, a RMSE of 18.5 and a R^2 of 0.31. The skewed distribution of the data might be in part the reason why the results are not optimal. Further model iterations can be performed to establish a more accurate prediction.

This tornado exploration can be expanded upon to create more models that strive to predict the magnitude of tornadoes given specific variables as input. The economic loss model can also be expanded to account for the specific crop, infrastructure, or wildlife loss, given the geographical track information in this dataset. There is still much to be learned about tornadoes and this exploration is just the beginning of many possibilities and projects to come.

References

https://www.nssl.noaa.gov/education/svrwx101/tornadoes/
https://www.nssl.noaa.gov/research/tornadoes/
https://www.nssl.noaa.gov/research/tornadoes/
https://www.nssl.noaa.gov/research/tornadoes/#:~:text=The%20U.S.%20typically%20has%20more,warnings%20to%20help%20save%20lives.

Appendix

Appendix A: Data Set Information

Dataset Variable Description:

- **yr**: Year, int
- **mo**: Month, int
- **dy**: Day, int
- **date**: Date, string
- **time**: Time, string
- **tz**: Time Zone, int
- **st**: State, string
- **mag**: Tornado Magnitude, int
- **inj**: Injured by Tornados, int
- **fat**: Fatalities by Tornados, int
- **loss**: Economic Loss in Millions, int
- **len**: Tornado Length in Miles, double precision
- **wid**: Tornado Width in Feet, int

Appendix B: EDA

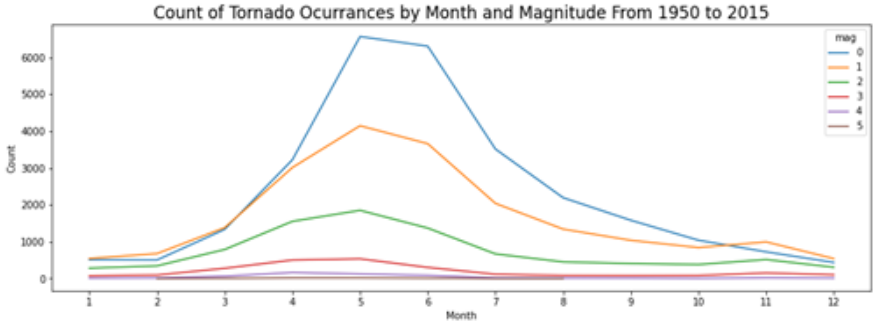

Figure A

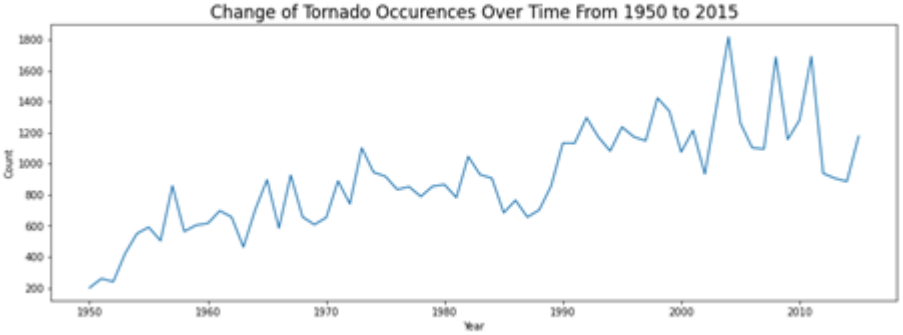

Figure B

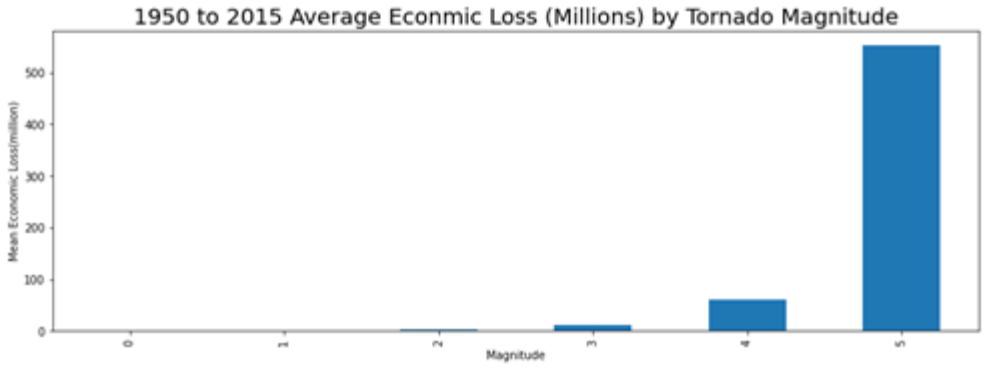

Figure C

Appendix C: Logistic Regression

Results

Classification Report Analysis:

	precision	recall	f1-score	support
0	0.72	0.96	0.83	4541
1	0.47	0.28	0.35	2060
2	0.26	0.04	0.06	590
3	0.53	0.05	0.09	172
4	0.00	0.00	0.00	38
5	0.00	0.00	0.00	4
accuracy			0.67	7405
macro avg	0.33	0.22	0.22	7405
weighted avg	0.61	0.67	0.61	7405

Precision: Percentage of correct predictions
Total Positive / (Total Positive + False Positive)

Recall: Percentage of positive instances
Total Positive / (Total Positive + False Negatives)

F1: Percentage of correct positive prediction
2*(Recall * Precision) / (Recall + Precision)

Support: Number of occurrences per category in a dataset

Classification Report A

Results

Classification Report Analysis:

	precision	recall	f1-score	support
0	0.80	0.89	0.84	4541
1	0.52	0.53	0.52	2060
2	0.34	0.09	0.14	590
3	0.52	0.19	0.28	172
4	0.33	0.05	0.09	38
5	0.00	0.00	0.00	4
accuracy			0.71	7405
macro avg	0.42	0.29	0.31	7405
weighted avg	0.67	0.71	0.68	7405

Precision: Percentage of correct predictions
Total Positive / (Total Positive + False Positive)

Recall: Percentage of positive instances
Total Positive / (Total Positive + False Negatives)

F1: Percentage of correct positive prediction
2*(Recall * Precision) / (Recall + Precision)

Support: Number of occurrences per category in a dataset

Random Forest Classification Report

Appendix D: Mathematical Model

F-scale Ranges (mph)

F0: < 73

F1: 73 - 112

F2: 113 - 157

F3: 158 - 206

F4: 207 - 260

F5: > 261

Figure D

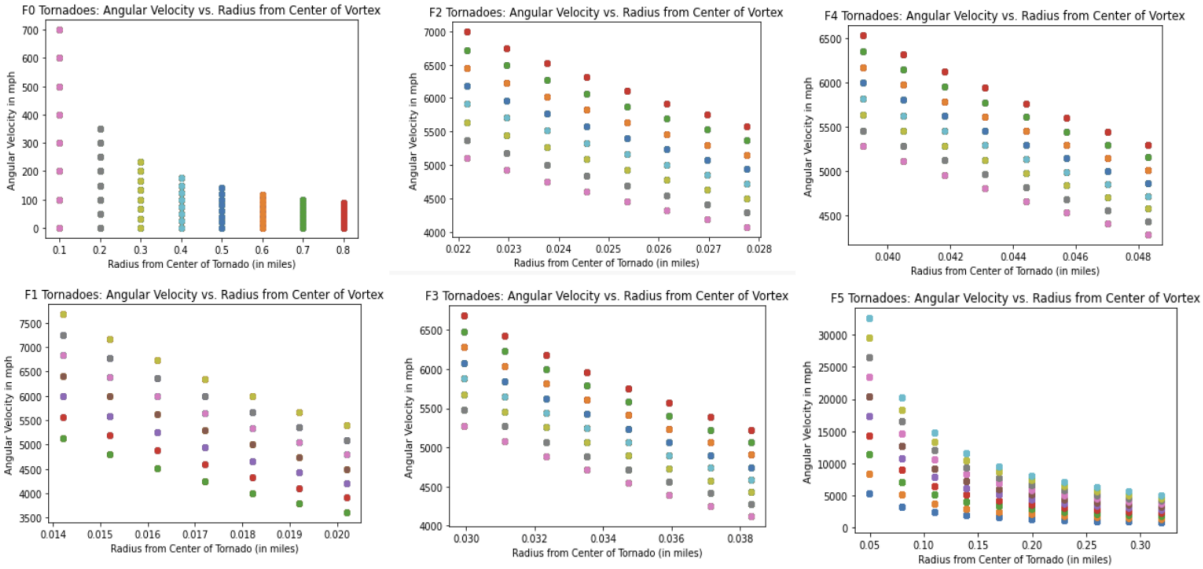

Figure E (a-f)

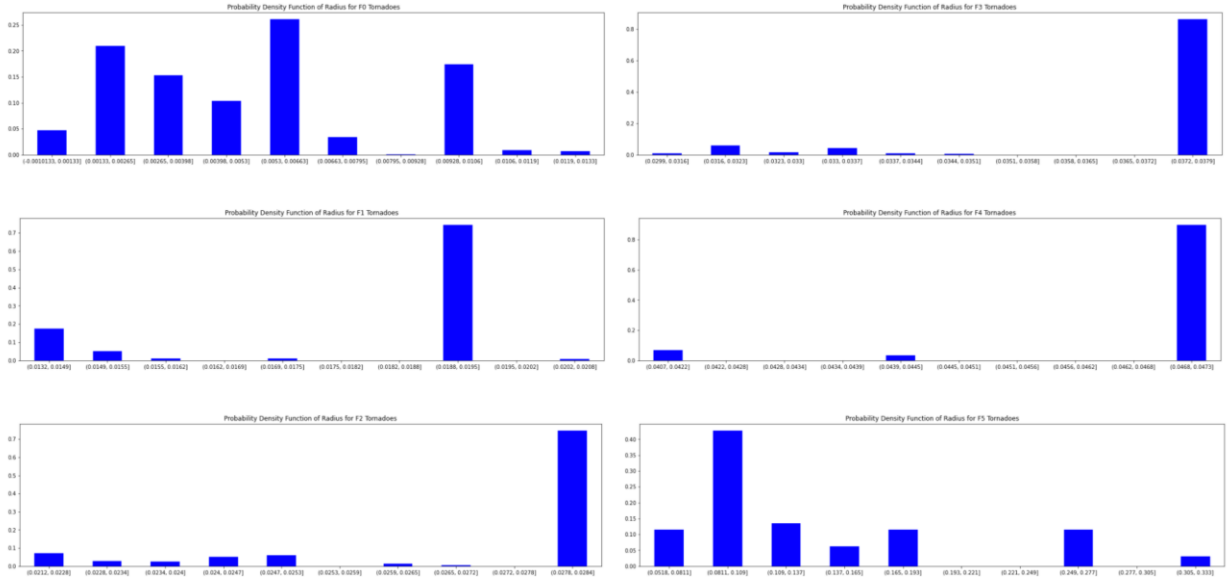

Figure F

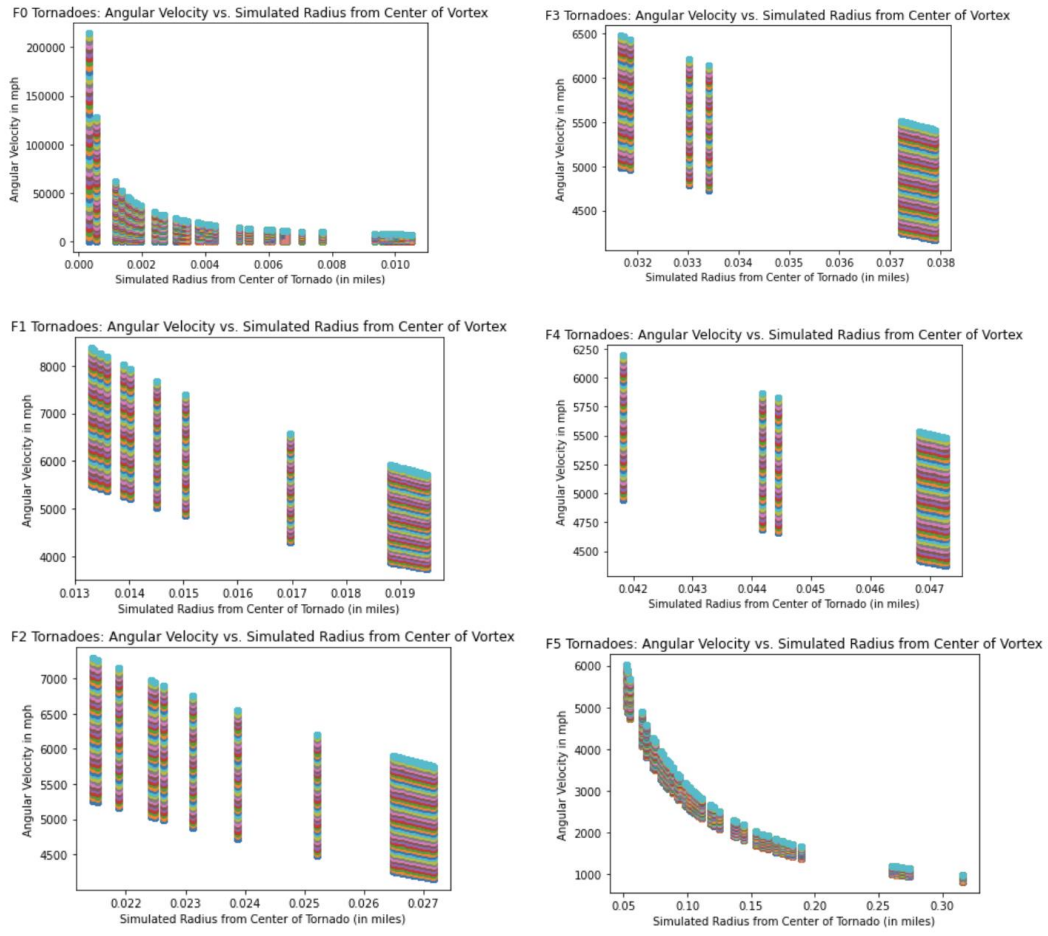

Figure G

Appendix E: Economic Loss Model

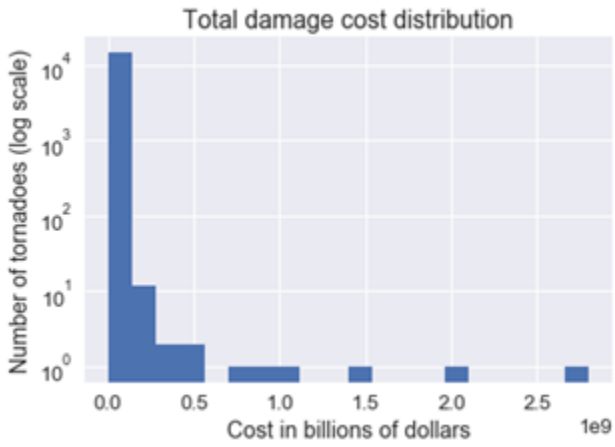

Figure H