DS3000 Storm Event Predictions by Weather Features

Luke Abbatessa, Daniel Gilligan, Ruby Potash, Megan Putnam

Northeastern University: Khory College of Computer Sciences

ABSTRACT

Our project investigated predicting tornado storm property damage based on storm event properties (associated injuries, deaths; Fujita scale rating; ground length [mi] and ground width [ft]; beginning/ending month of the event) using data from the NOAA National Centers for Environmental Information Storm Events Database during 2012-2022. The best performing machine learning model out of the three algorithms tested was a hyper tuned, 30 tree random forest regressor, which had a coefficient of determination value of 0.98 and a mean squared error (MSE) value of 8.07e+14. Despite the high MSE, this model showed little evidence of either under or over fitting and explained a significant degree of variance as indicated by the strong R2 value. The most important features were associated deaths, injuries, and tornado length.

INTRODUCTION

In 2021 alone, the United States experienced 20 weather/climate-related disasters of $1 billion dollars or more, and 338 of these billion-dollar disasters have occurred since 1980. The total cost of these 338 disasters is $2.295 trillion—an amount that could pay off all of the outstanding student loan debt in America 1.5 times over. Given the extent of damages, we thought that it was of interest to develop a forecasting model for storm damage prediction, more specifically monetary property damage. A prediction model that could create such a forecast based on certain storm data (e.g. wind speed, grade of severity, time of year) would add a more human aspect to weather forecasting and would allow people to truly grasp the scale of severe weather.

Moreover, prediction of storm damages could help communities focus their preparatory efforts more efficiently and accurately. It is evident that the vast damages associated with these weather events are not just limited to financial: from 1980-2022 there have been 15,689 fatalities associated with billion-dollar natural disasters, with an average of 365 fatalities per year [2]. Recently, trends are not improving either. Between the years 2017 and 2021, there were 89 billion-dollar disasters and 4,557 associated fatalities [2]. It is clear from these statistics that an extensive amount of money goes into storm event response and reparations—not to mention the additional environmental destruction and loss of human lives that occurs with these events. Therefore, it is in the best interest of public safety and storm effect prevention to generate a model of weather event damages that works to predict adverse impacts of storm events based on their intrinsic properties (e.g. type of event, time of year, etc). This way, given a particular storm's characteristics, communities can have a better idea of the ways in which a storm may affect their landscape and population. In addition, this could underscore the situation's severity and encourage people to take proper precautions, as described above.

As such, our **goal was to create a predictive model based on the historical data patterns of tornadoes**—these being an optimal example of severe and devastating storm events—that would function to **predict subsequent property damages**, thereby addressing the significant issues described above.

RELATED WORK

Related to our planned analysis of properties of storm events and their relationship to both physical and financial damages, one publication developed a generalized linear model (GLM) relating meteorological conditions to damages incurred to Public Service Electric Gas in New Jersey, finding that lightning information and wind duration were two missing predictors of damages [3]. Another paper used the same dataset as our project looking at the meteorological intensity and impacts of winter storms from 2001 to 2014 on coastal counties in Connecticut, New Jersey, and New York [4]. They found that floods were responsible for the highest losses, but that no relationship was found between the number of storms that hit a county and the damage the county faced. A third paper reviewed climatological influences on major past storm events in the North-east Atlantic [5]. They found that major storm impacts are associated with positive North Atlantic Oscillation phases.

METHODOLOGY

The storm event data were first extracted from the NOAA (National Oceanic and Atmospheric Administration) National Centers for Environmental Information Storm Events Database [1]. The Storm Events Database has an archive of data csv files detailing the occurrence of storms and significant weather events each year from 1950 to 2022. We chose to analyze storm information from the last decade (2012 to 2022). For each year, there were three different csv files: storm details, storm locations, and storm fatalities, with each event assigned a unique storm ID that was consistent across these three file types. As such, we import each of these three files for all of the years in our desired time frame. Then, we merged storm detail, location, and fatality data on unique storm ID and integrated them into a compiled data frame containing all data type information for storms during years 2012-2022.

Following this data extraction, we isolated records pertaining to tornado storm events only. Next, we inspected the data for missing data patterns, invalid data, duplicate records, and other data cleaning needs. Data cleaning steps included filtering variables of interest (those pertaining to tornado storm properties and storm damages), splitting dates into year and month integers, and converting to correct data types. Additionally, we dropped all records that were not classified as tornado events. Since we were focusing our analysis on prediction of damages given a limited set of tornado storm properties, and there was widespread data missing from tornado records, we decided to delete any records that were missing more than one of our variables of interest. The final features selected for model generation were tornado injuries, deaths, Fujita scale rating, length and width of tornado, and beginning/ending month of the storm event.

Finally, we selected our three algorithms for model generation and optimization. Since our project aimed to predict tornado property damages based on tornado storm features, our model needed to be equipped to predict continuous property damage values. This signified that we needed to implement regression machine learning algorithms, and subsequently led us to our final selection: **1) multiple linear regression, 2) k-nearest neighbor regression, and 3) random forest regressor algorithms**. The training/test split was 75/25 given the smaller set of data for model generation, and we conducted grid searches with 5-fold cross validation for KNN regression and random forest regressor models to hyper tune k neighbor and tree estimator values, respectively. Each model's performance was evaluated using mean squared error (MSE) and coefficient of determination (R2) values to be compared with the other models.

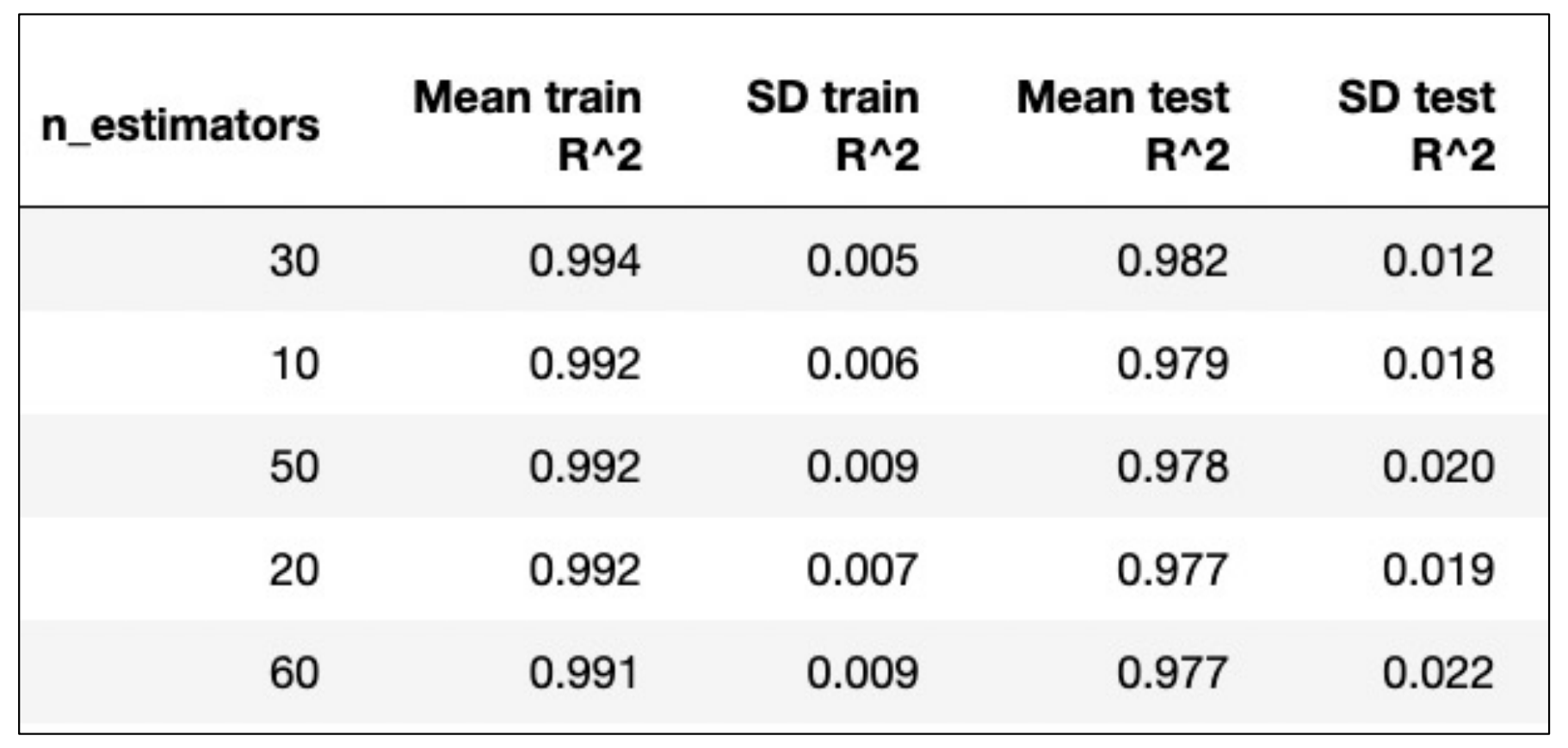
RESULTS AND EVALUATION

Multiple Linear Regression: The multiple linear regression model performed with an MSE value of 3.02e+16 and an R2 value of 0.09. The very high MSE value and low correlation coefficient indicated that the multiple linear regression model was not a good fit for the data and was not the best choice for predicting tornado damages based on the input features.

K-nearest neighbor regression: The k-nearest neighbor regression (KNN) performed best with a k value of 2. This model yielded an MSE value of 2.25e+14 and a coefficient of determination value of 0.864. The MSE was smaller than that of the linear regression model, but it was still a relatively large value. This could be partially accounted for by the large spread and high variance of the property damage data (mean value of $60,507,658.23 and median value of $2,000,000.00). The higher R^2 indicates a better fit of the model to the data. However, given the difference between test and training scores—suggesting the possibility of overfitting—and the large MSE value, it was concluded that this model is not optimal for predicting tornado property damages.

SELECTED MODEL

Random Forest Regressor: The random forest regressor model performed best with a tree count of 30. This model yielded an R2 value of 0.982 and a MSE of roughly 8.07e+14. Again, this was a high MSE value, yet the R2 value indicated a good fit for the data. Importantly, it has been argued that R2 can be more informative/truthful than other evaluation metrics such as symmetric mean absolute percentage error (SMAPE), mean absolute error (MAE), MAE percentage variant (MAPE), root mean standard error (RMSE), and MSE in evaluation of regression models [6]. Therefore, given the 30 tree random forest regressor model's lowest MSE and highest R2 values, we concluded that this was the best model for predicting tornado property damages given the input features. Further analysis of features demonstrated that deaths and injuries from a tornado storm event had the highest level of relative importance in the model followed by tornado length (Figure 1). Furthermore, relatively high R2 values and small differences in mean and SD scores between train and test sets indicate lack of over and underfitting of the model (Table 1).



n_estimators	Mean train R^2	SD train R^2	Mean test R^2	SD test R^2
30	0.994	0.005	0.982	0.012
10	0.992	0.006	0.979	0.018
50	0.992	0.009	0.978	0.020
20	0.992	0.007	0.977	0.019
60	0.991	0.009	0.977	0.022

Table 1. Results from random forest regressor hyper tuning of tree number as measured by coefficient of correlation mean and standard deviation values during testing and training. Relatively high R2 values and low SD between train and test sets indicate lack of over and underfitting of the model. n_estimators = number of trees; R^2 = coefficient of determination; SD = standard deviation.

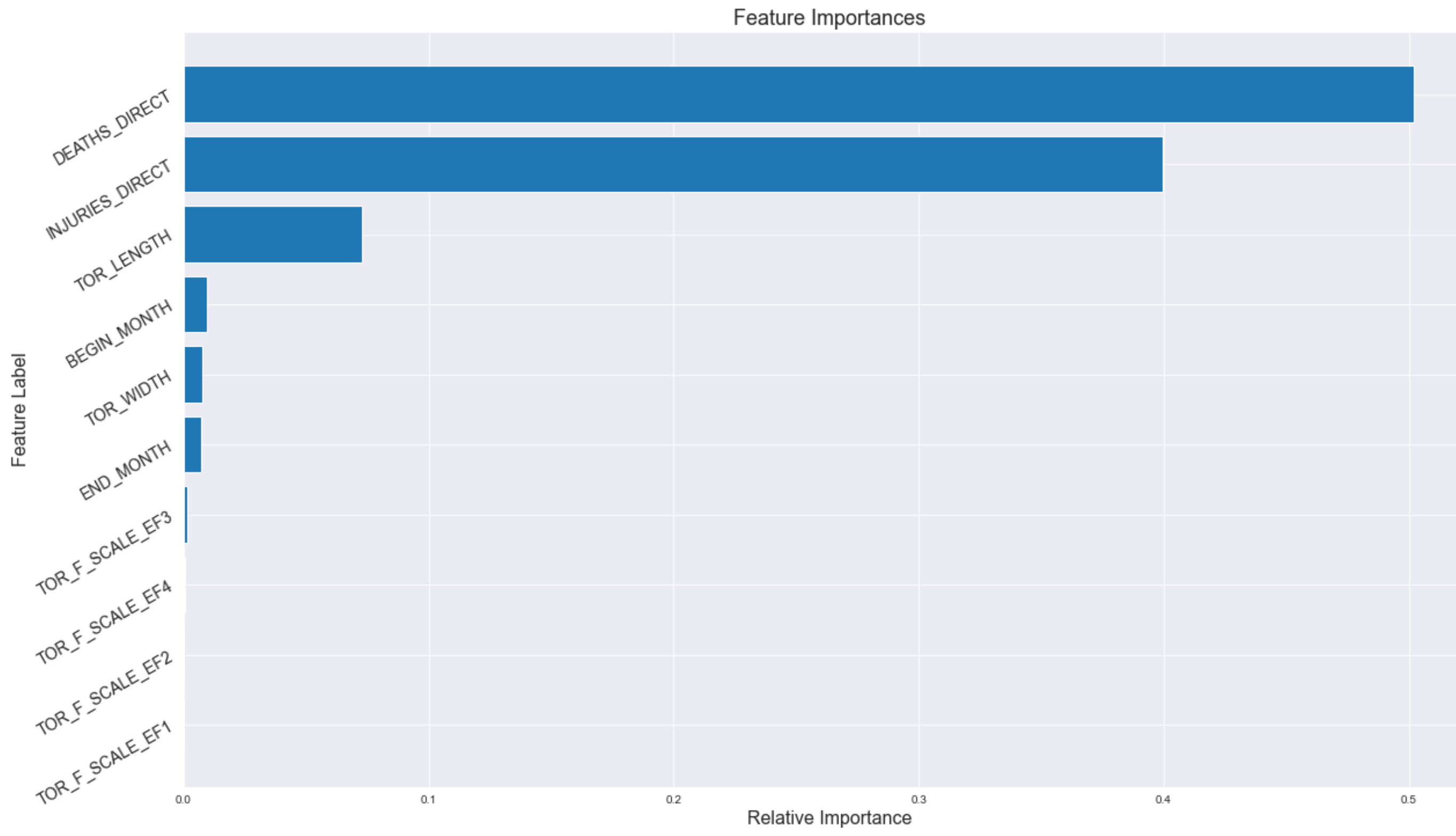


Figure 1. Relative feature importance for 30 tree random forest regressor model. Features in descending order of importance were tornado event associated deaths, associated injuries, tornado length, beginning month of storm, tornado width, ending month of storm,, and F scale rating of 3, 4, 2, and 1.

IMPACT

This project provides conceptual support for the development of better storm damage prediction models given intrinsic storm properties. While our chosen model had a relatively high MSE value (8.07e+14), it had a high coefficient of determination (0.982) and can still provide useful information about tornado damage prediction and preparation. This type of forecasting model has the potential to better equip communities for focusing their preparation efforts/fundings more efficiently by highlighting which tornado features make a storm more vulnerable to high property damages. For example, in our random forest regressor model, tornado length and width surprisingly had a greater feature importance than having rating on the Fujita scale. This could indicate that the physical size of the tornado had a greater impact on property damages compared to F scale rating alone, which should be taken into account when preparing for a given tornado's effect (e.g. organizations should not solely rely on F scale rating to determine precautionary efforts). Therefore, as mentioned, this project could benefit local communities as they allocate funds/resources for storm preparation and fortification. Additionally, a similar data analysis and model could be applied to other storm damage variables (i.e. crop damage, injuries, fatalities, etc.) or storm types (i.e. hurricanes, hail storms, etc.) to gain similar insights on storm damage prediction factors. This work is especially important as increased weather-related disasters are seen due to climate change [7].

CONCLUSION

In this project we explored the relationship between property damage and tornado storm properties including tornado-associated injuries, deaths, Fujita scale rating, length and width of the tornado, and beginning/ending month of the storm event. Our best performing model was a 30 tree random forest regressor, which had an R2 value of 0.98 and an MSE of 8.07e+14, showing little evidence of under or over fitting. The most influential model features were associated deaths, associated injuries, and tornado length in predicting storm damages. Further work can be done on investigating this relationship, perhaps with a larger dataset or exploring alternative predictive algorithms (especially given the large MSE yielded by the three models tested here). Additionally, it would be valuable to see similar analyses replicated to predict other storm effect variables, such as crop damage, or look at different types of storms, such as hurricanes.

REFERENCES

1. Storm events database | national centers for environmental information. (n.d.). Retrieved December 1, 2022, from https://www.ncdc.noaa.gov/stormevents/

2. NOAA National Centers for Environmental Information (NCEI) U.S. Billion-Dollar Weather and Climate Disasters (2022). https://www.ncei.noaa.gov/access/billions/, DOI: 10.25921/stkw-7w73

3. Cerruti, Brian J., and Steven G. Decker. "A Statistical Forecast Model of Weather-Related Damage to a Major Electric Utility." Journal of Applied Meteorology and Climatology, vol. 51, no. 2, 1 Feb. 2011, pp. 191–204., https://doi.org/10.1175/jamc-d-11-09.1.

4. Shimkus, Cari E., et al. "Winter Storm Intensity, Hazards, and Property Losses in the New York Tristate Area." Annals of the New York Academy of Sciences, vol. 1400, no. 1, 17 July 2017, pp. 65–80., https://doi.org/10.1111/nyas.13396.

5. Pouzet, Pierre, and Mohamed Maanan. "Climatological Influences on Major Storm Events during the Last Millennium along the Atlantic Coast of France." Scientific Reports, vol. 10, no. 1, 21 July 2020, https://doi.org/10.1038/s41598-020-69069-w.

6. Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ. Computer science, 7, e623. https://doi.org/10.7717/peerj-cs.623

7. World Meteorological Organization. (2021). Atlas of mortality and economic losses from weather, climate, and water extremes (1970-2019). Geneva, Switzerland :World Meteorological Organization.