Predicting Tornadoes

Abdullah Robins
Simon Angel
Alvaro Carrascosa
Frazier Horn

The Problem

Tornadoes are dangerous, deadly, and unpredictable—they leave decimated communities in their wake and perplex scientists. More tornadoes hit the United States than any other nation on the planet. Despite the fact that technological advancements in climatology and meteorology have increased tenfold over the last decade, forecasts and warnings available to the general public with regards to tornado detection remain only a few minutes out from occurence.

Our Approach

When thinking about tornadoes, the first place that generally comes to mind for someone wanting to study this transient phenomena is the Great Plains of the America—better known as Tornado Alley. Gathering data for this monumental task was difficult because of data inadequacy and access, so we decided to approach the problem of prediction in a completely different manner than does the National Weather Service. The National Weather Service currently uses the Tornado Detection Algorithm which collects Doppler radar velocity data. Since we did not have access to this data nor any historical Doppler radar data, we decided to analyze developing trends and patterns in weather leading up to tornado occurrences. Our goal for the data was to study and better observe features near the ground such as temperature, wind speed, surface solar radiation, and surface pressure, to name just a few. The outcome of our project is to provide people in tornado ridden areas with a tool that gives them the probability of tornado occurring at different time frames with high reliability.

Data Collection

Our data collection capabilities were limited due to the fact that most live weather radar API services are incredibly expensive. This limitation forced us to work with common weather parameters that are often found online and made available via free API trials. This type of data was often composed of several weather parameters arranged in a time series.

Data requirements

In order to optimally train and validate the accuracy of our proposed models, we decided to obtain the following two types of data:

a) Weather data leading up to Tornado events
b) Weather data unrelated to Tornado events

API

In order to obtain weather data we decided to use the OikoLab API[1] due to the fact that it offered a free trial and a great number of useful weather related features. This enabled us to retrieve the following weather parameters for any given day and location for the past 10 years.

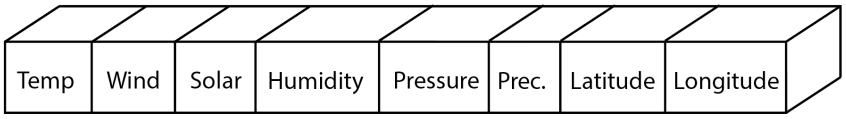
Available features

- **Date** *(MM-DD-YYYY | HH-MM)*
- **Temperature (***Measured in Celsius)*
- **Wind Speed (***Measured in m/s)*
- **Solar Surface Radiation (***Measured in W/m^2)*
- **Relative Humidity (***Measured in g/m^3)*
- **Surface Pressure (***Measured in atmospheres)*
- **Precipitation (***Measured in l/cm^3)*
- **City (***Name of the city)*
- **Latitude** *(degrees | minutes | seconds)*
- **Longitude** *(degrees | minutes | seconds)*

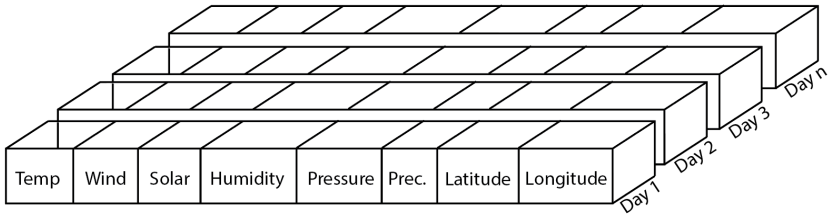
Data structure

In order to easily understand and visualize our data structure, we decided to gather and structure the required data the following way.

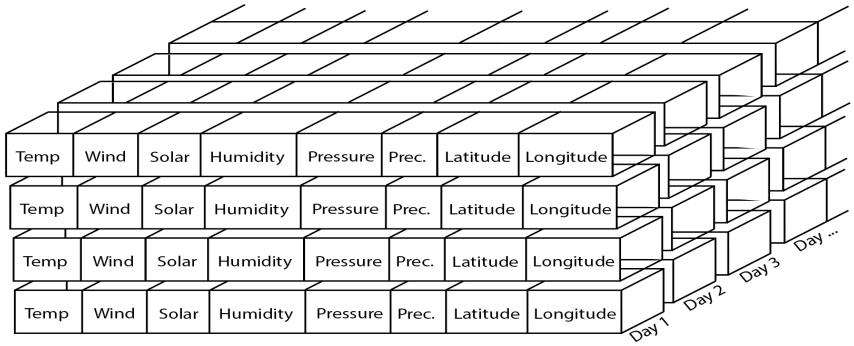
- This illustration represents the data structure received by a call to our API



- The model will be trained with time serieses of data, therefore the following structure represents the data for a tornado over a series of n days.



- By combining different Tornado events, we will arrive at the following three dimensional data structure.



Fetching the data

In order to obtain the required data in the structure previously discussed, we did the following:

- Weather data leading up to Tornadoes

  We retrieved a list of past Tornadoes and their corresponding times and locations from the NOAA Database[2].

  Consequently, we used the times and locations of each event to retrieve a series of 60 days of weather data leading up to the time of the event through the OikoLab API[1].

- Random weather data

  By exclusively using the location of the Tornadoes, we retrieved a large number of random serieses of 60 days of weather data using the same OikoLab API[1].

→The file *weather_data_scripts.ipynb* contains a set of different functions that we built in order to retrieve data in different ways. It allows us to generate more instances of random weather data, as well as, retrieve weather data with custom parameters or sets of different days.

Storing the data

- By making use of these functions and gathering the data according to our predefined steps, we built two different files to split the storage of tornado weather data and random weather data. We decided to store our data in *.csv* format, where each entry represents a day of weather data. Each 60 consecutive days then represent the time series data for each tornado or random weather set.

  The file containing the data for tornado events is *all_new_data.csv*

  The file containing the data for random weather is *random_weather_allnew.csv*

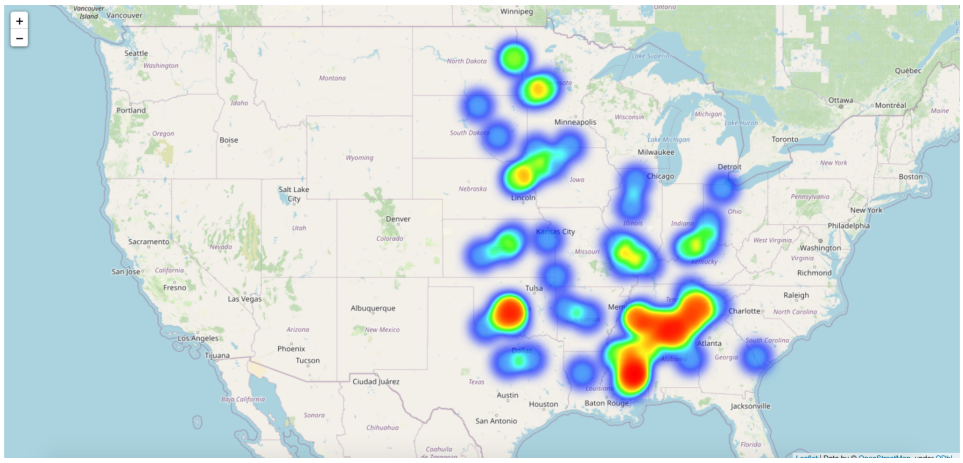
Data Preparation and Cleaning

After gathering the data from our sources we needed a succinct way of grouping the historical data for both the tornadoes as well as the random data. To this goal we added event-ids to the random historical data which allowed us to group the random data alongside the tornado data. To the data we added a column for outcome and if the data pertained to tornadoes we put 1 and 0 otherwise. The format of the data in the end before adding additional features looked like this:

	datetime	temperature	wind_speed	surface_solar_radiation	relative_humidity	surface_pressure	total_precipitation	city	event_id	latitude	longitude	outcome
0	2017-04-30	12.18	5.76	181	0.60	98916.0	0.03	Birmingham	0	NaN	NaN	0.0
1	2017-05-01	11.01	3.41	191	0.79	99275.0	0.21	Birmingham	0	NaN	NaN	0.0
2	2017-05-02	10.88	3.09	184	0.76	100772.0	0.03	Birmingham	0	NaN	NaN	0.0
3	2017-05-03	9.70	5.75	189	0.72	101186.0	0.00	Birmingham	0	NaN	NaN	0.0
4	2017-05-04	10.98	5.86	248	0.66	101182.0	0.00	Birmingham	0	NaN	NaN	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
12592	2020-07-04	26.76	3.30	331	0.61	97023.0	0.00	NaN	898307	46.11	-95.9	1.0
12593	2020-07-05	24.55	1.61	233	0.69	97004.0	0.11	NaN	898307	46.11	-95.9	1.0
12594	2020-07-06	24.38	1.65	287	0.68	96966.0	0.01	NaN	898307	46.11	-95.9	1.0
12595	2020-07-07	25.59	3.91	267	0.68	96431.0	0.50	NaN	898307	46.11	-95.9	1.0
12596	2020-07-08	31.23	1.21	290	0.56	96143.0	0.01	NaN	898307	46.11	-95.9	1.0

Exploration

The first attribute of the tornado data we explored was the geographical data. We plotted the location of the tornadoes in our data into a heatmap of the US to better understand where this phenomenon usually occurs.



When exploring these meteorological conditions we visualized how the data behaves ~60 days before an event, which could be a tornado (outcome = 1 on the graph) or simply a day without a tornado (outcome = 0 on the graph). The development of the

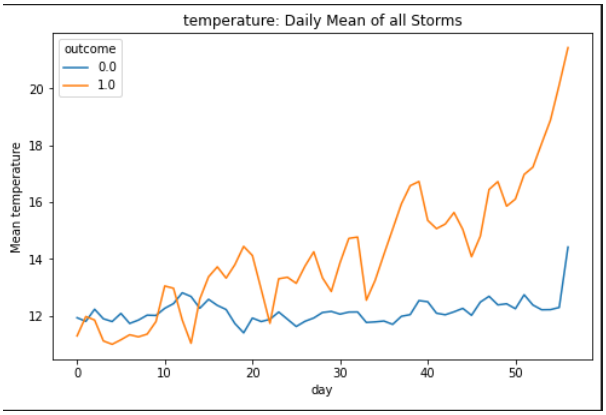
heatmap allowed us to determine the locations in the US in which there is a high frequency of tornadoes, we chose to sample our non-tornado samples (outcome = 0) from these locations.

In the following figures, the orange line represents samples with an outcome of 1, which end with a tornado, while the blue line represents those that end with a day on which a tornado did not occur.

1. Temperature

Figure 1

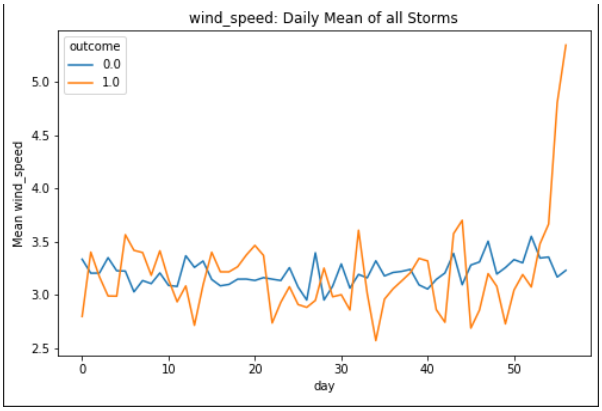
In figure 1, one can observe the average temperature for all samples on the 60 days before an event. We decided to continue analyzing temperature because of the higher rate of increase in temperature in the ~25 days before a tornado compared to the ~25 days without a tornado.



2. Wind Speed

Figure 2

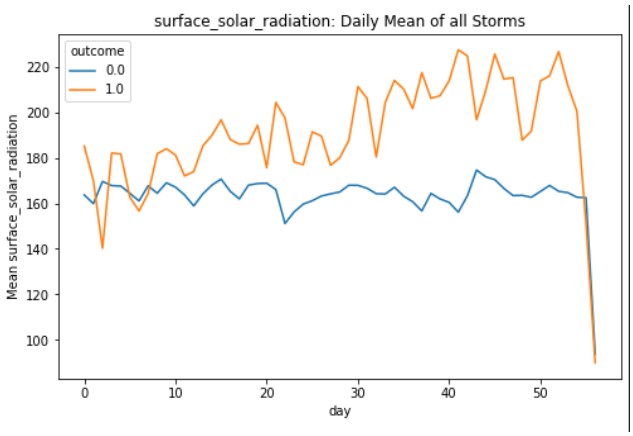
In figure 2, in the earlier 50 days of the samples we can observe there is much more volatility in the samples that end with a tornado (outcome = 1) when compare to the samples that did not end with the occurrence of a tornado.



Also importantly, we can observe a stark increase in the wind speed for ~1 week before the occurrence of a tornado (outcome = 1). However, we find a rather constant rate of change when observing the average wind speed across all samples that led to a non-tornado day.

3. Surface Solar Radiation

Figure 3

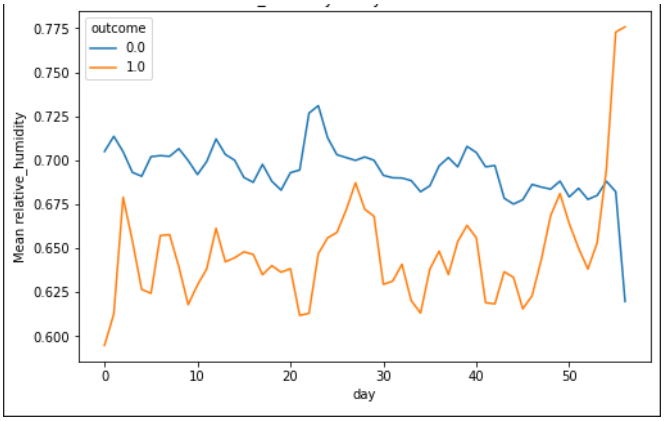


In figure 3, we observed that data of outcome = 1 had more volatility throughout the ~60 days before the event, when compared to the samples with outcome = 0.

Moreover, data of outcome = 1 had more of a linear increase in the ~ 50 days than the data representing samples that did not end in a tornado, which seems to observe a constant

4. Relative Humidity

Figure 4



In figure 4, one can observe that in the days leading to a tornado the average humidity is lower across all days, except 7-8 days before the tornado.

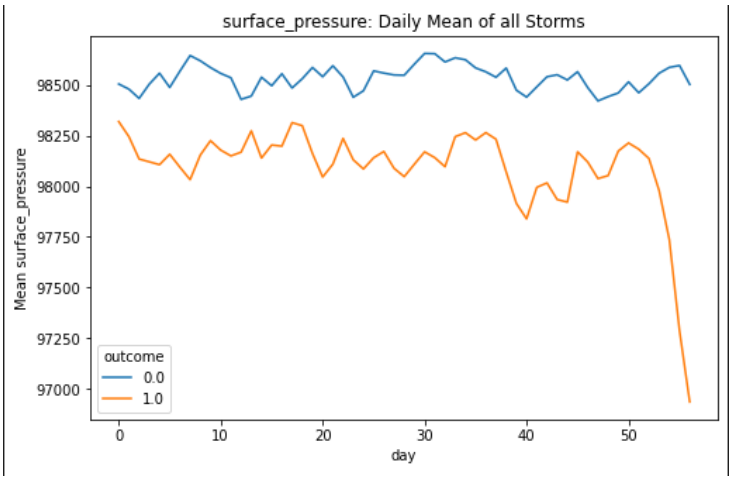
Moreover, we can observe much more volatility in the trend line representing the tornado samples.

Finally, we can observe that the tornado samples see a large spike in humidity in the ~ 1 week before the tornado, while in the samples ending with a non tornado day, there is no such phenomenon, in fact there is a small decrease.

5. Surface Pressure

Figure 5

In figure 5, one can explore that there is a lower average level of surface pressure in the samples that result in a tornado when compared to those that do not. Importantly, this reflects the scientific research that tornadoes are prevalent in time periods with lower pressure levels.

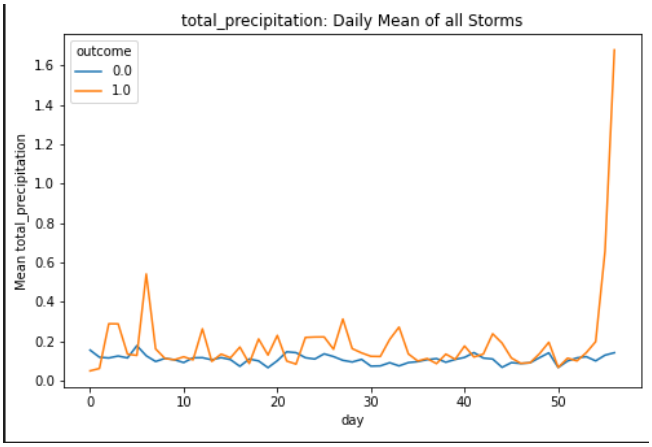


Also in ~10-8 days before a tornado, one can see there is an evident decline in the pressure, while in the days leading to a non-tornado day saw no such decline.

6. Total Precipitation

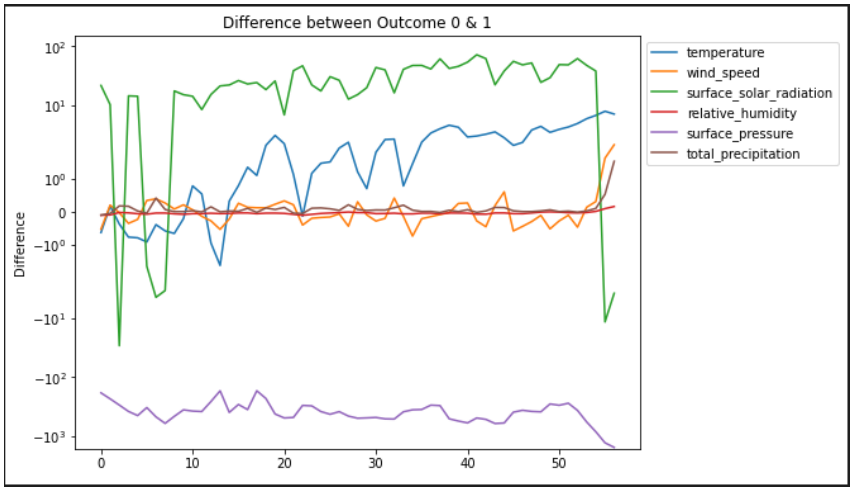
Figure 6

In figure 6, we must importantly note that there are spikes in precipitation in the samples ending in a tornado (outcome = 1), while those not ending in a tornado see constant precipitation at much lower levels throughout the observed 60 days.



Also, we can see the ~3-4 day period before a tornado (outcome = 1) has a enormous spike in precipitation, while those samples not ending in a tornado seem to continue their constant rate of precipitation.

Comparing the difference in magnitude between outcome=1 and outcome=0



In the figure above, we have plotted the difference between the two classes across the 60 days before the event. We have done this for each of the 6 types of data explained in the exploration. In this figure we subtract the average for each day for samples with outcome = 0 from those with outcome = 1. We did this to explore behavior; however, this highlighted the need for normalization as the difference in magnitudes would schew the way the models would interpret the data.

Data Generation

This section will refer to features we generated and experimented with, but did not necessarily include in our final models. The features of our final models will be discussed in the “Analysis of Models Section”.

Features

We attempted to create 4 features to represent the patterns in these aspects of the environment, temperature, solar radiation, humidity, wind speed; we will refer to these as “the aspects” in this explanation. We do the following for each sample:

1. For each aspect, find the data for the 21 days leading up to the event, we now have data for the 3 weeks before the event.
2. Use these 21 data points per aspect to compute 21 rolling averages such that each one is a rolling average of its predecessors.

    a. For example the 3rd average will be an average of the later 3 days, while the 21st will be an average of all 21 days leading up to the event.
3. Normalize each set of 21 rolling averages with respect to its predecessor rolling averages.
    a. So each aspect is normalized separately to avoid the difference in magnitudes between the aspects .
4. Process the data to compute 4 measurements per aspect reflecting the change in each aspect, so for example temperature would look like this.

temperature_0	temperature_1	temperature_2	temperature_3
0.382884	1.085619	0.159219	-1.627722

→ See AttemptedButFailedModel.ipynb for more information

We fed this data to a Naive Bayes and a logistic regression model which had a reliability of 68% and 52% respectively. We discarded this model and continued with the one outlined in the “Analysis of Models and Results” because of its superior reliability.

Ratios:

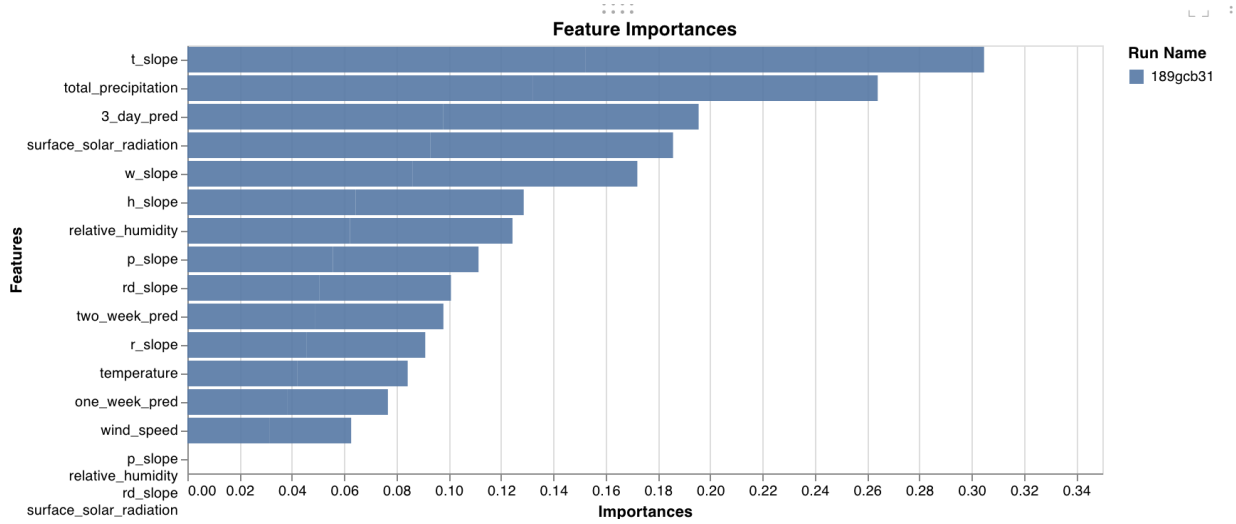
We developed functions to compute ratios between weather conditions such as change in humidity over pressure and experimented with these. However, the results were not reliable enough and we did not proceed with any ratios between different meteorological conditions in our final model.

Analysis of Models and Results

Features

The features we decided to explore were the slopes for each weather feature over time. We experimented with different time frames to find the optimal span of time to calculate the slopes for each feature. These features allowed us to capture the changes that happen in the environment for each span of time making our models even stronger. We also explored creating features to capture interval changes as well as ratios and polynomial calculated features.

After amassing the various features we ran the models and ranked the weight of the feature importance to the models. We removed the features with negative importance and narrowed our features down.



Model Initiation

We began our analysis by making this a binary classification problem with the outcome being a tornado or no tornado. To this end we decided to use logistic regression as well as Naive Bayesian for our models. Our models were trained using historical data for each Tornado as well as random historical data in the Tornado regions. The data was assigned a binary outcome to represent the events. As we progressed through the

project we decided upon doing four separate models for various prediction timeframes. The prediction timeframes were two weeks, one week, three days, and one day. For each model we used, we added the prediction from the last models as a feature. By using the models in succession we are able to build upon our models and provide more reliable results to our users.

Model Analysis

Scope/Reliability Table

Scope of Data (Time Period)	Time of Prediction (Days before Day of Tornado)	Reliability (Accuracy)	Layers
Uses earlier 14 Days of 28 Days before Tornado	14 Days	83.5%	N/A
Uses earlier 21 Days of 28 Days before Tornado	7 Days	85.7%	14 Day out Prediction
Uses earlier 11 days of 14 Days Before Tornado	3 Days	85.7%	14 & 7 Day out Prediction
Uses earlier 13 days of 14 Days Before Tornado	1 Day	89.0%	14, 7, & 3 Day out Prediction

We started off with a model that only predicted up until the day before if there was a tornado or not. This model had a high test accuracy of around 89% which was also higher than the train accuracy. After further analysis of the phenomenon we decided to add more data which ended up bringing the accuracy down for that model. From this point on we looked at creating more multi-dimensional features to better cover our time series data and added additional models to cover different prediction timeframes.

After retraining, tweaking, and optimizing our models we end up with our four models that range from 80-90% in their reliability. These models can be used in real life scenarios for people in the tornado regions. The goal of these models are to provide the user with a probability that there will be a tornado within the various time frames provided since the models allow the option of giving the prediction probability for each outcome.

Here are the results for our test data predicting a tornado the next day:

Tornado Probability	event_id
0.723268	658562
0.501502	52
0.800376	582231
0.664567	309267
0.194963	55
...	...
0.824176	505228
0.119353	12
0.585617	451535
0.824879	237666
0.438552	50

91 rows × 1 columns

References

1. OikoLab's API- https://oikolab.com
2. NOAA, National Centers for Environmental Information - https://www.ncdc.noaa.gov/stormevents/