Final Project - Team America

Divya Rajendran, Ethan Violette, Pramod Duvvuri, Wenjuan Sang

13 April 2018

Introduction and Statement of Goals

Are storms in the Tropics of the Americas increasing? With hurricanes staying at the forefront of the news over the past couple of years, the idea of an increase in number of dangerous storms is fairly frightening. We aim to utilize past data of tropical and subtropical storms to answer this question, as well as see the trends in storm frequency for each category of storm. For our choice of data set, we utilized the Department of Homeland Security's Storm Tracking data[1].

Data Description

The data contains 59,228 objects of storm instances, with simple features such as day/month/year of the storm, basin of origination (either the Northern Atlantic or the Eastern Pacific), Wind Speed (in knots), Pressure, and Category of Storm. Location data, as in latitude and longitude, was also included, but because of the spherical nature of the globe (apologies for the contraversial statement), this information is extremely difficult to interpret in the context of a model, and beyond our expertise. We will then instead just use latitude and longitude for nice visual plots.

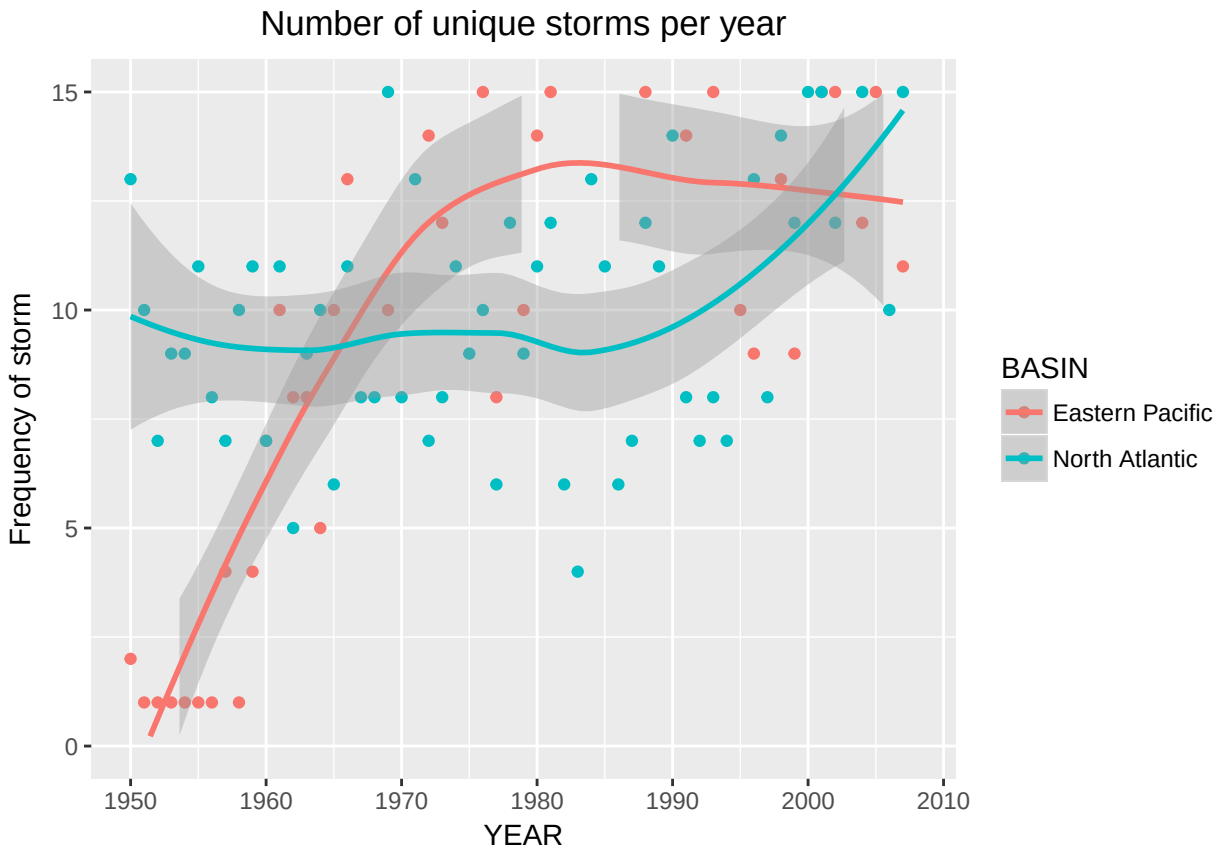
When we check the summary of our Historical storm data, we are interested in descriptive fields like YEAR, MONTH, DAY, WIND_KTS, PRESSURE, CAT, LAT, LONG.

We see that the storms in the data are divided into two regions 1. Eastern Pacific and 2. North Atlantic and that the data ranges for the years 1851 to 2008. Unique storm names were given from the year 1950, and so therefore we are only considering the data from 1950-onwards in our project.

We transformed latitude and longitude unique pairs of data to get 51 unique locations to understand how storms vary in each of these locations.

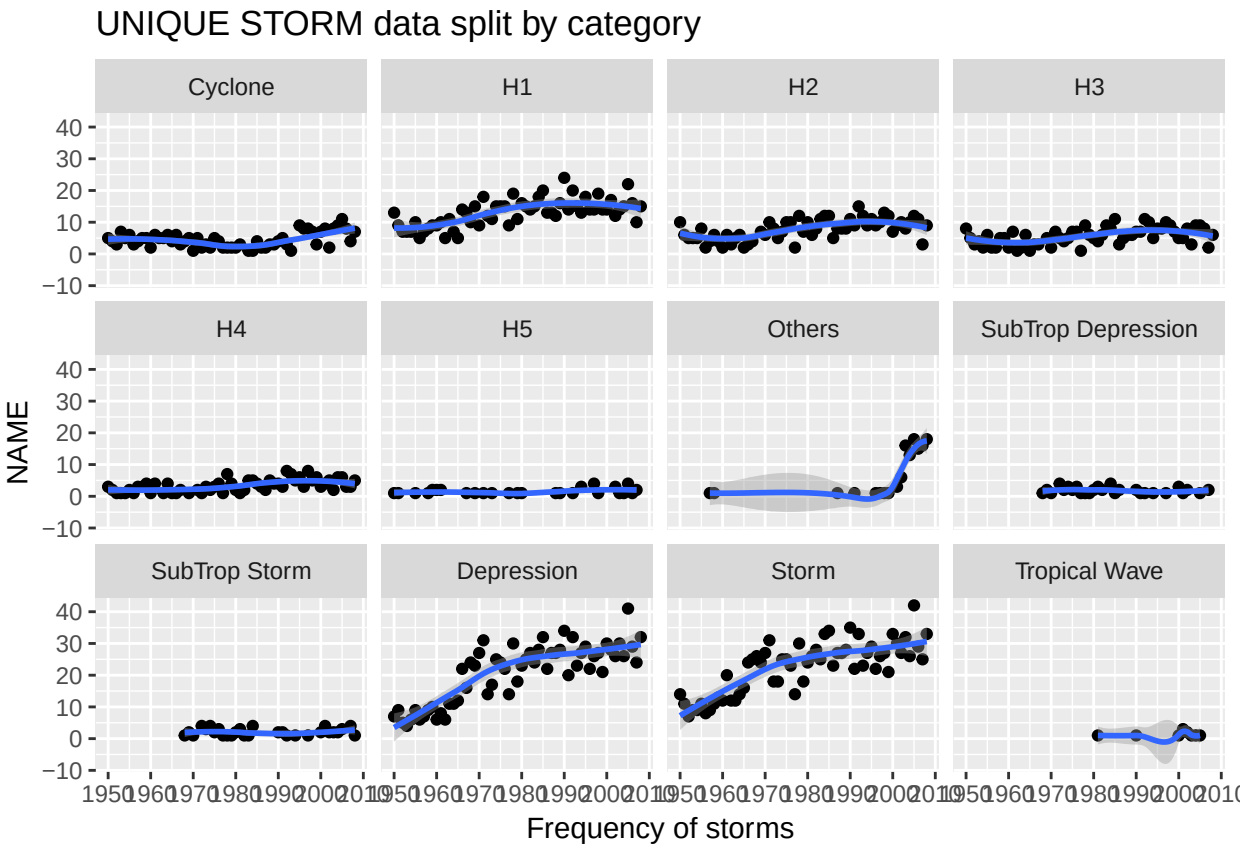
```
##       YEAR                  BASIN         NAME
##  Min.   :1950   Eastern Pacific:59   Min.   : 1.00
##  1st Qu.:1964   North Atlantic :59   1st Qu.: 8.00
##  Median :1979                       Median :11.00
##  Mean   :1979                       Mean   :11.69
##  3rd Qu.:1994                       3rd Qu.:15.00
##  Max.   :2008                       Max.   :28.00
```

1. A bivariate relationship between number of unique storms per year



We first used the loess method to plot data. We found that the patterns of storm frequency over the years are differentiated by their originations. Overall, we observe an increase in storm numbers in both basins. It is interesting that, when we split the data by basin, the plot shows that storms increase really fast from 1950 to about 1980 in Eastern Pacific area, during which time, storm number almost has no change in North Atlantic area. After 1980, storm number slowly decreases in the Eastern Pacific area (almost a flat line). However, it increases quickly in North Atlantic area, suggesting that we might want to split the data by Basin.

2. A trivariate relationship between Number of unique storms, category of the storm and the year.



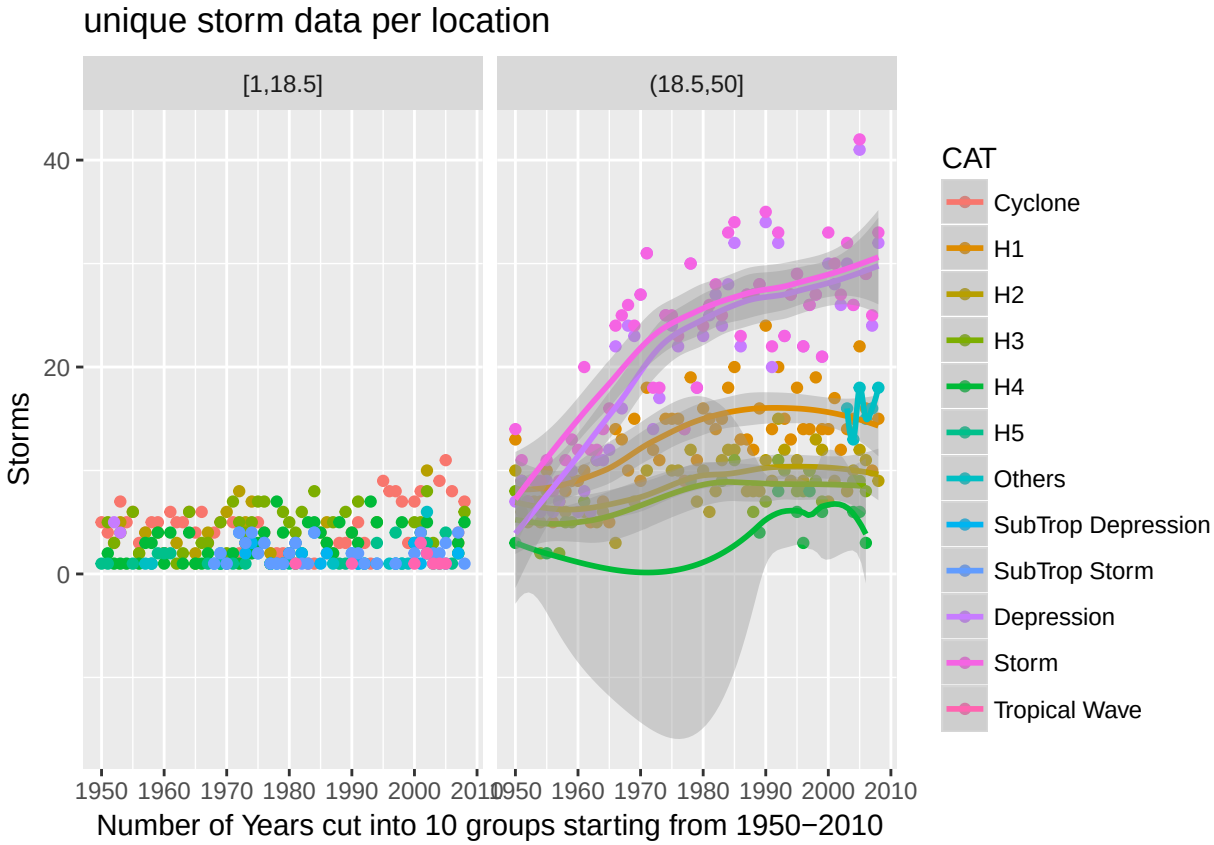
The numbers of each category of storms are highly varied. Some of the types, like tropical wave, only have few samples, while the others have fairly large samples. Depressions and Storms are constantly increasing over the years and the others do not have much change longitudinally. We see that there were huge number of TD and TS storms; with an averagely increasing trend for TD and TS storm categories, implying that storms of type TD and TS were more common over the years.

To specify, Tropical Wave occurances are fewer in number, implying that either the identification of such storms was only possible from the year 1980 onwards, or the first recorded occurance is in year 1980. The same can be said of storms of the category Other, which observed a sudden increase through the years 2000-2008. The storms of category Subtropical Storm, H5, Subtropical Depression are less frequent in numbers ranging between 0-5 each of the years. The storms of category Cycle, H2, H3, H4 are very frequent in numbers ranging between 0-12 each of the years. Also, we see that the number of storms of categories: H1, Tropical Depression and Ttropical Storm have increased during the period 1985-1995.

3. The relationship between number of unique storms in each year per location facetted by category.

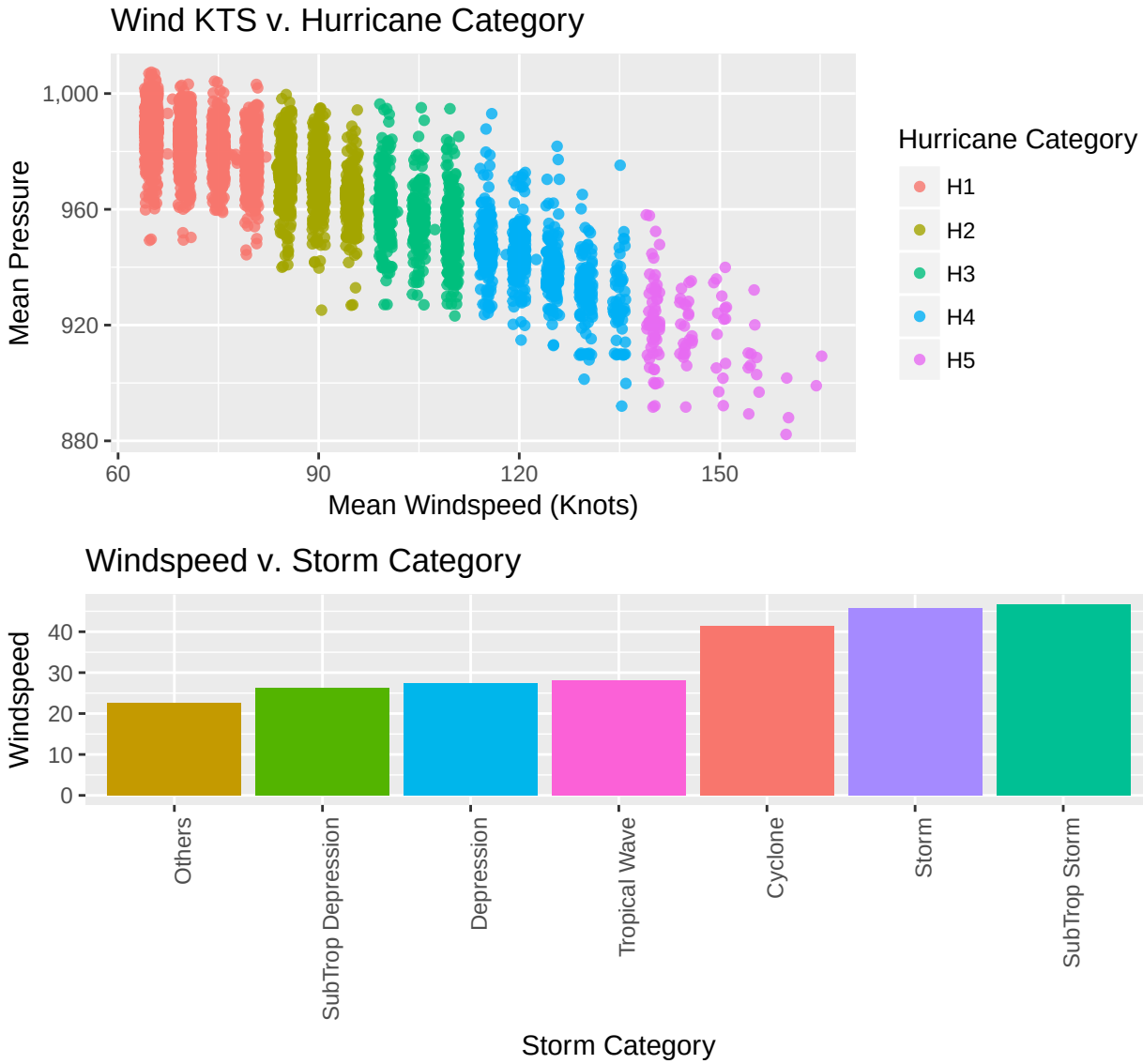
```
##       YEAR             CAT           NAME        loc_grpID
##  Min.   :1950   H1        : 59   Min.   : 1.0   Min.   : 1.00
##  1st Qu.:1967   Depression: 59   1st Qu.: 3.0   1st Qu.: 7.00
##  Median :1981   Storm     : 59   Median : 6.0   Median :18.50
```

```
##  Mean   :1981   H2         : 58   Mean   : 9.4   Mean   :21.66
##  3rd Qu.:1996   H3         : 58   3rd Qu.:13.0   3rd Qu.:37.00
##  Max.   :2008   Cyclone    : 57   Max.   :42.0   Max.   :50.00
##                 (Other)    :162
```



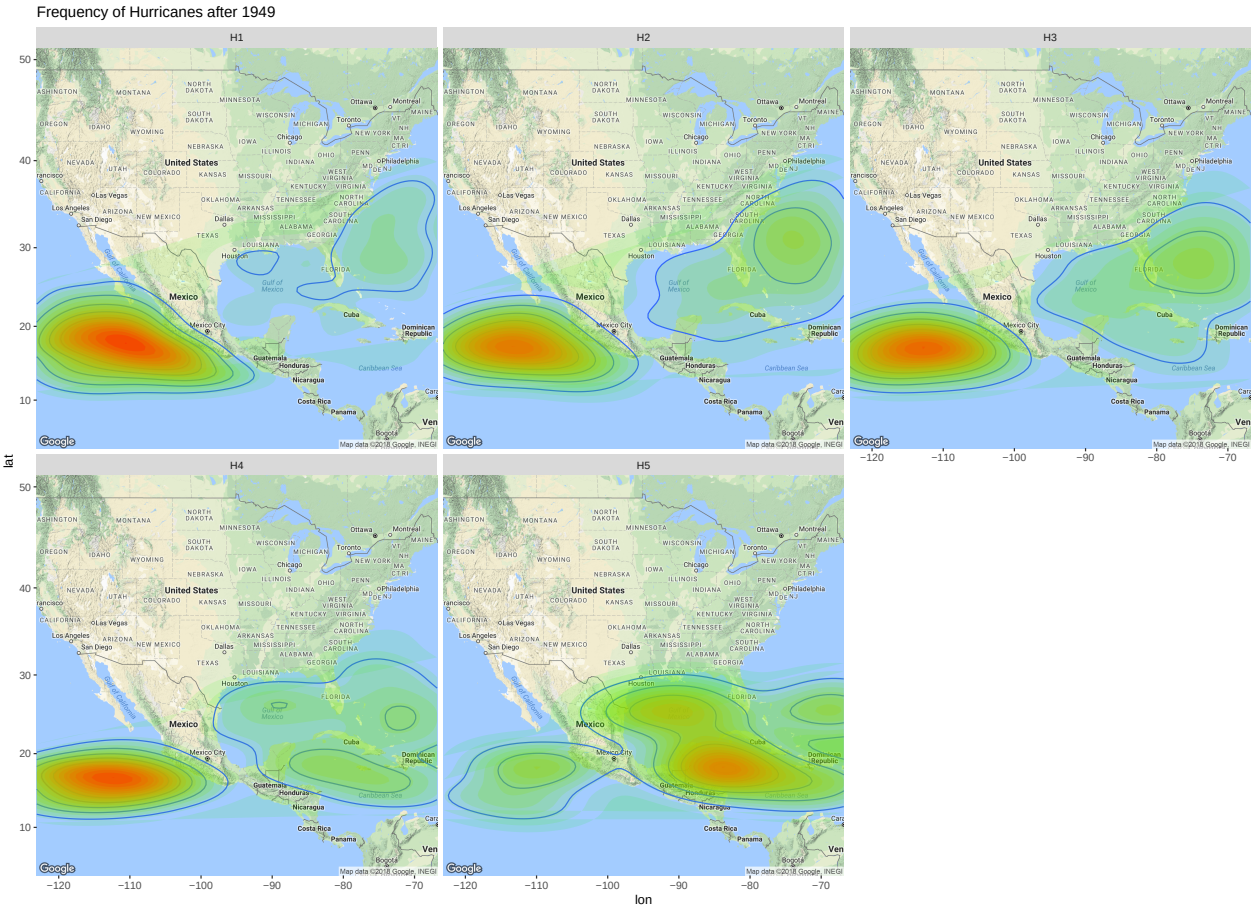
From the above plot, we can see that there is a significant relationship between the number of storms and unique locations. However, since a location id was obtained by combinations of latitude and longitude, we cannot surely comment about the relationship, due to limitations regarding coordinates specified in the introduction. This also indicates the difficulty in coding the combinations of longitude and lattitude merely by data. In order to make the location variable more meaningful, we will plot them on maps.

4. Relation between wind speed and pressure

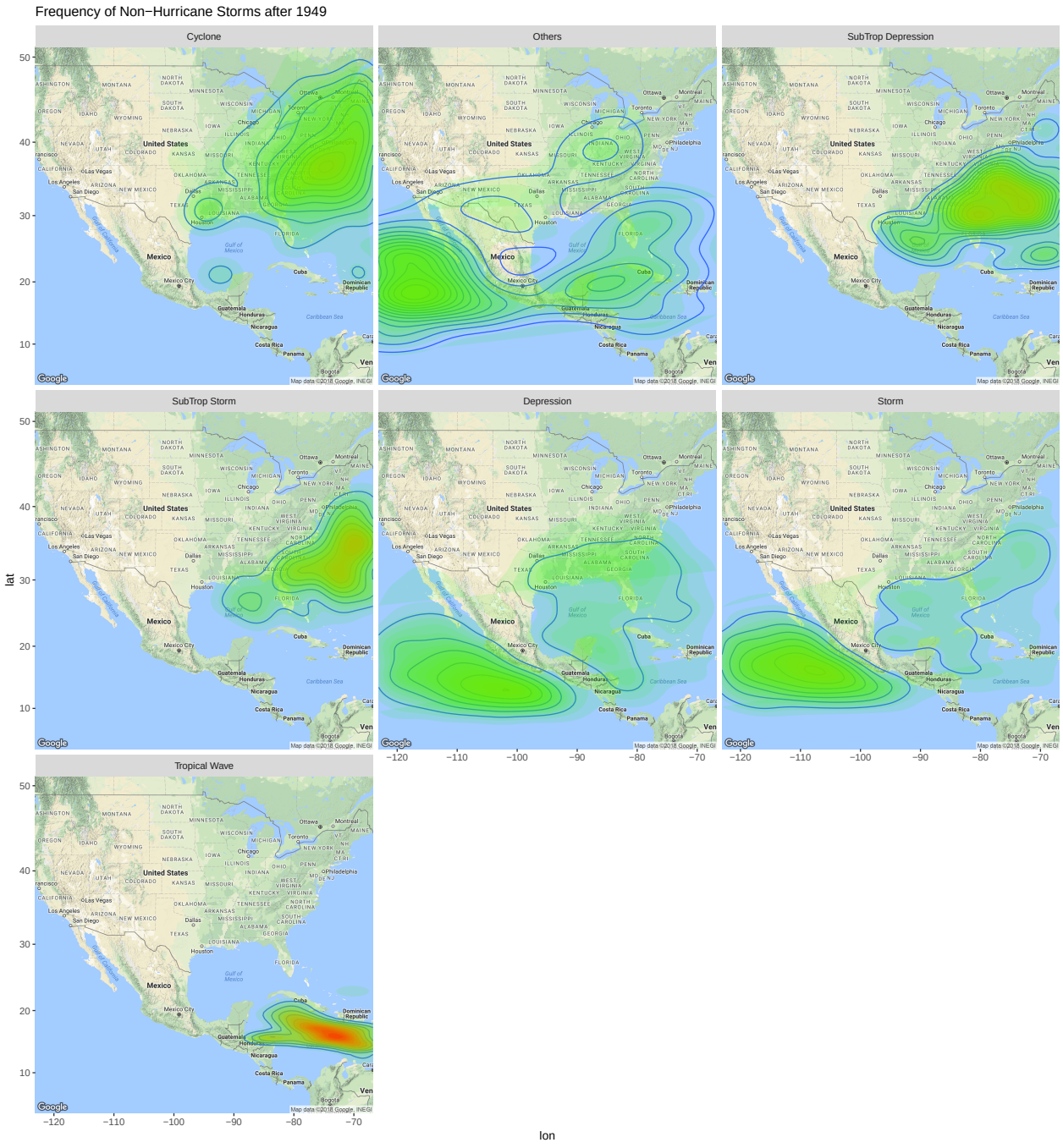


As we can see from above graphs, the five hurricane types are categorized perfectly by their wind speed and pressure. From H1 to H5, the pressure goes down, but the wind speed goes up, which means that the it becomes stronger and is able to deal more damage and destruction. For the storms that are not hurricanes, we see that Subtropical Storms and Tropical Storms have the highest wind speed (and therefore are the most dangerous). Subtropical Depressions and Depressions have relatively lower wind speed, but have the potential to become storms in their own right. Cyclones, due to their ability to turn directly into hurricanes, correspondingly have relatively high wind speeds. Tropical Waves, though not very dangerous, have an average wind speed amongst the types of non-hurricane storms. These two plots give us a clear idea about the relationship between storm category and wind speed and pressure.

5. Mapping the frequency of storms

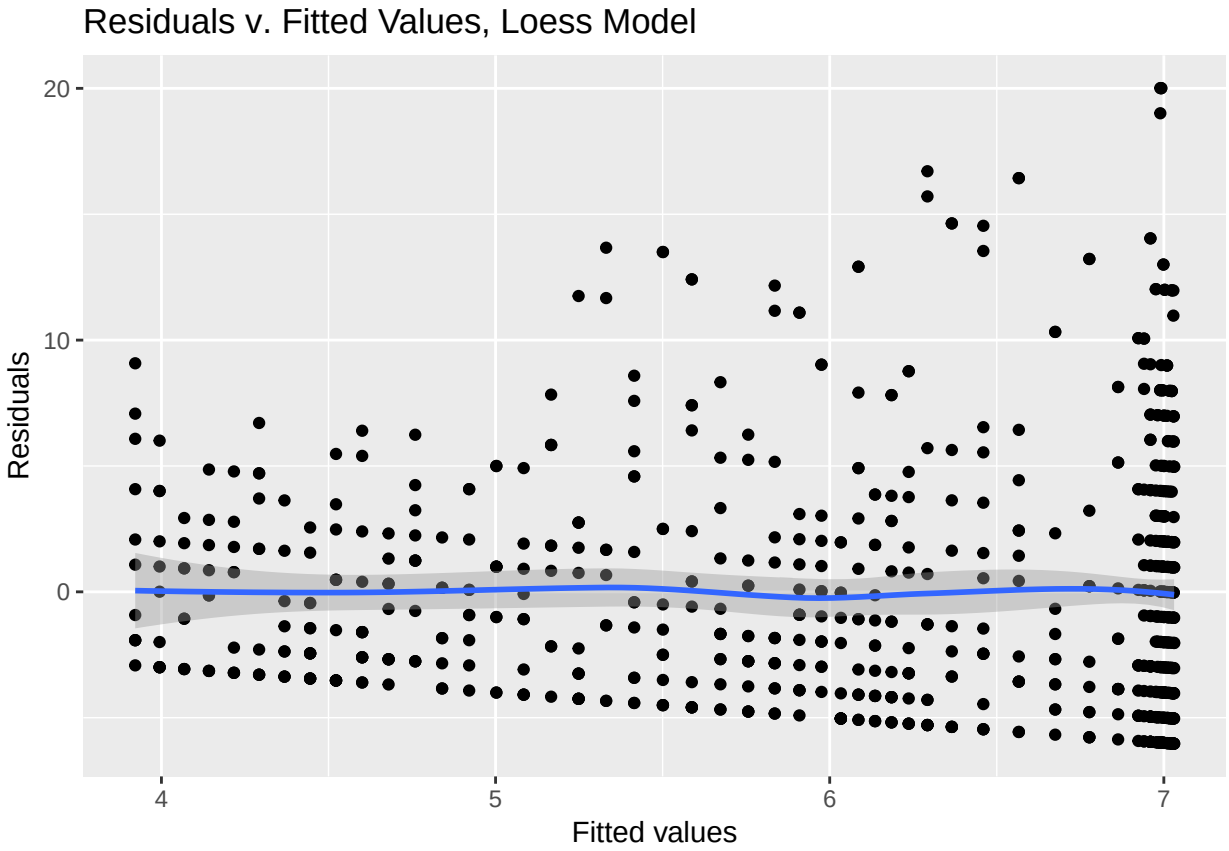


We used the package ggmap to plot the frequency of storms based on their originations. The color represents the density of storms, for which the red areas form the most storms and the light green areas form the least. It is clear that most hurricanes are formed in Eastern Pacific area, except H5, the most dangerous category. H5 hurricanes are much more often seen in North Atlantic area, rather than Eastern Pacific area.



As for the non-hurricane storms, we note a relatively even dispersion of storm frequency between the North Atlantic Basin and the Eastern Pacific Basin; the lone exception is the Tropical Wave, whose frequency is concentrated solely in the North Atlantic. This introduces an interesting contrast between the non-hurricanes' even dispersion and the hurricanes' frequency dispersion shift from Eastern Pacific to Northern Atlantic, as the hurricanes get stronger.

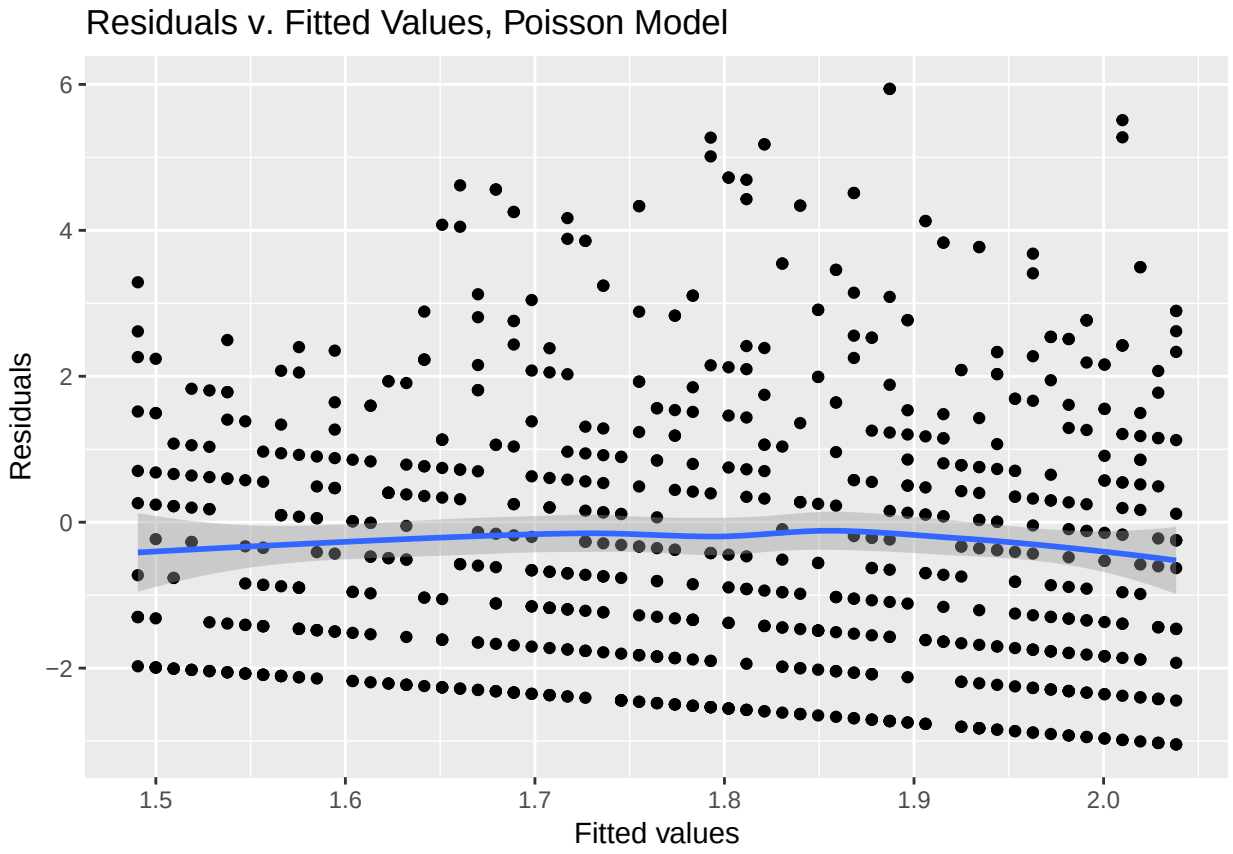
5 Modeling time and storms



First, we fitted a loess model to get a basic impression of storm number and year. The curve for the residuals just wiggles around 0, meaning that that our regression model specifies an adequate relationship between the outcome, storm number and the covariates, Year. Above EDA suggests that there are potential factors can be added to the model.
When we thought more deeply about out data, we knew that the data might be treated as count variable, instead of a continous one.

```
##
## Call:
## glm(formula = NAME ~ YEAR, family = poisson(link = "log"), data = model.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0460  -1.8612  -0.5747   0.9230   5.9388
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.693e+01  1.738e+00   -9.74   <2e-16 ***
## YEAR         9.444e-03  8.757e-04   10.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3144.3  on 796  degrees of freedom
```

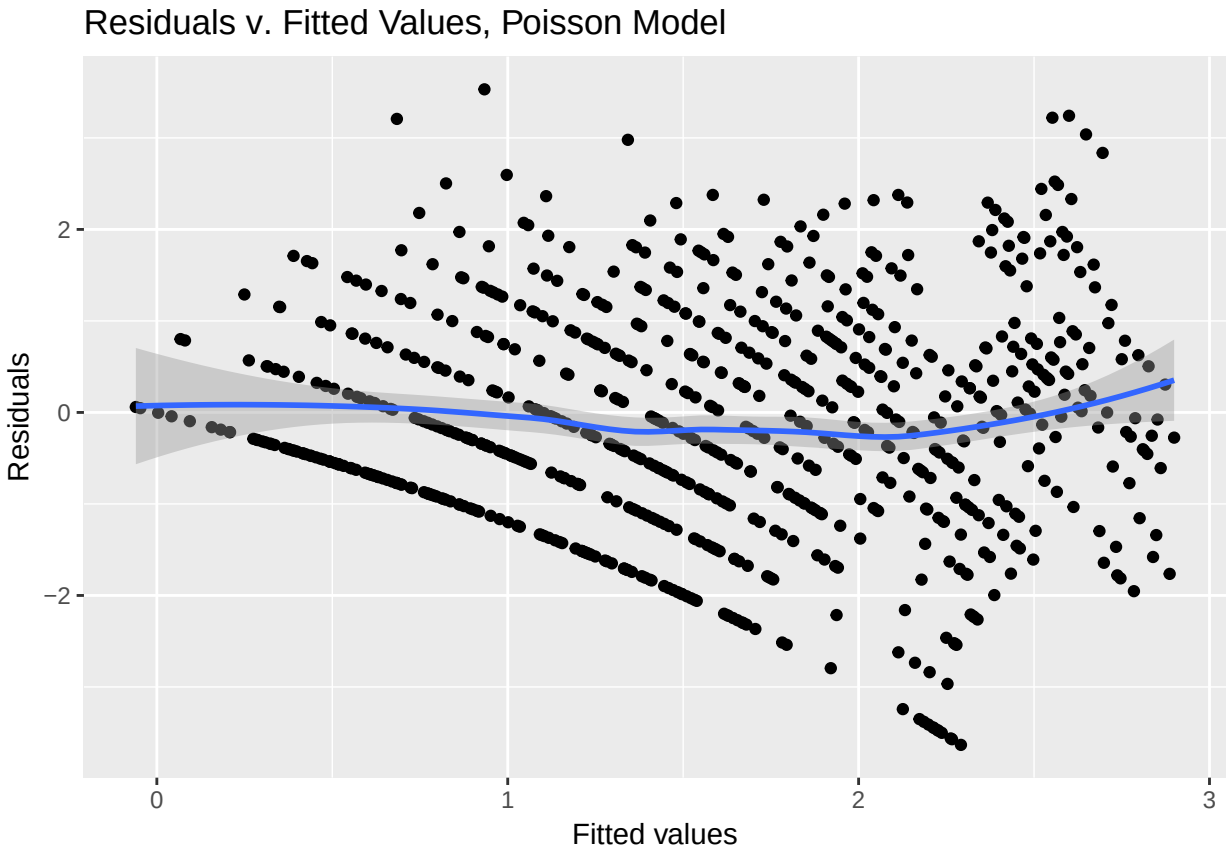
```
## Residual deviance: 3026.1  on 795  degrees of freedom
## AIC: 5674.5
##
## Number of Fisher Scoring iterations: 5
```



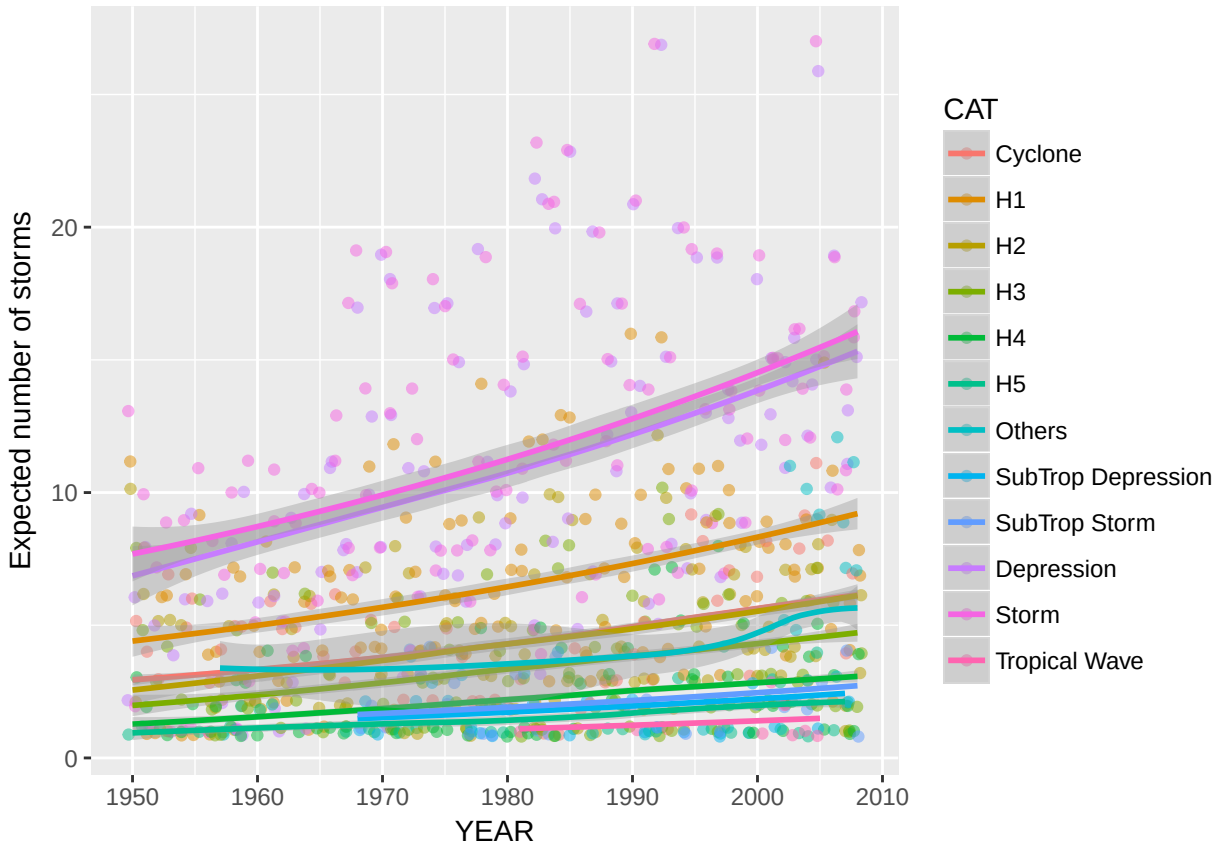
The curve for the residuals just wiggles around 0, meaning that that our regression model specifies an adequate relationship between the outcome, storm number and the covariates, Year. This supports our prediction that treating dependent variable as count variable fits the data. We can continue adding more covariates into the model.

```
##
## Call:
## glm(formula = NAME ~ YEAR + CAT + BASIN, family = poisson(link = "log"),
##     data = model.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6331  -0.8960  -0.1742   0.7462   3.5309
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -23.491331   1.771810 -13.258  < 2e-16 ***
## YEAR                    0.012733   0.000892  14.275  < 2e-16 ***
## CATH1                   0.265156   0.074163   3.575  0.00035 ***
## CATH2                  -0.143783   0.079899  -1.800  0.07193 .
## CATH3                  -0.393263   0.084987  -4.627 3.70e-06 ***
## CATH4                  -0.827841   0.097506  -8.490  < 2e-16 ***
## CATH5                  -1.137222   0.170325  -6.677 2.44e-11 ***
```

```
## CATOthers              -0.217200   0.114668  -1.894  0.05820 .
## CATSubTrop Depression  -0.914479   0.154753  -5.909 3.44e-09 ***
## CATSubTrop Storm       -0.814706   0.142689  -5.710 1.13e-08 ***
## CATDepression           0.774642   0.070679  10.960  < 2e-16 ***
## CATStorm                0.821981   0.070320  11.689  < 2e-16 ***
## CATTropical Wave       -1.378081   0.308311  -4.470 7.83e-06 ***
## BASINNorth Atlantic    -0.261370   0.030129  -8.675  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3144.3  on 796  degrees of freedom
## Residual deviance: 1201.6  on 783  degrees of freedom
## AIC: 3874
##
## Number of Fisher Scoring iterations: 5
```



Almost all variables are significant in predicting the number of storms, suggesting that we have obtained correct information from the EDA procedure. As the time increases by one year, about 0.01 more storms will be formed. This is not very exciting with year as the unit, but if we consider it monthly or daily, the number of storm increase becomes more impressive. Compared with Eastern Pacific area, less storms are formed in North Atlantic area. The curve looks fine for the full model with three covariates, but the residuals of the model show a bit of a pattern , especially in the bottom left area, which could mean that our model has limitations in its predictive accuracy.



Above graph shows the predicted number of storms. The graph indicates that the most storms are predicted for category Storm, especially if the storms are formed in more recent years. The lowest number of predicted awards is for those H5 hurricanes (excluding Tropical Wave that has few samples). The graph overlays the lines of expected values onto the actual points, although a small amount of random noise was added vertically to lessen overplotting.

Conclusions and Limitations

It appears that the number of storms is indeed increasing at a small rate as time goes on; this can be predicted with some degree of accuracy with features such as Basin of Origination, Category of Storm, and (of course) Year. Though we had sufficient tools at our disposal to conclude this using a fitted Poisson Model, we're limited in our predictive accuracy due to lack of additional features. Given more time, we would have merged this dataset with another that contained information about weather, number of man-made influencers of climate change, and other useful predictors, over time.

Sources

[1] https://data.world/dhs/historical-tropical-storm