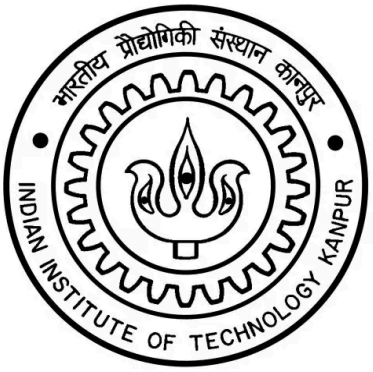
# INDIAN INSTITUTE OF TECHNOLOGY KANPUR




## DMS-672
## End-Term Project

HURRICANE INTENSITY FORECASTING

**Members:**
Stuti Gupta (211062)
Mridul Nambiar (210633)
Rohan Virmani (210871)
Ravi Kumar(210832)
**Instructor - Dr. Faiz Hamid**

# Index

# INTRODUCTION

Hurricanes are unpredictable and highly destructive storms that can strike with little to no warning, causing significant damage to both lives and property. Accurate predictions of a hurricane's intensity and trajectory are vital for effective disaster management. By anticipating the storm's path and strength, authorities can take proactive measures to protect vulnerable communities and minimize the impact of the storm. With advancements in technology and machine learning, it is now possible to create more reliable models to forecast hurricane behaviour, thereby providing the necessary time to prepare.

However, accurate forecasting is not just about predicting the storm's physical characteristics—it is equally important to consider the socio-economic and geographical factors that determine how different regions will be affected. Integrating demographic data such as age, gender, income levels, and population density with geographic information like infrastructure and land use patterns can offer deeper insights into the storm's potential impact. This allows decision-makers to better understand which areas will be most vulnerable and prioritize emergency responses accordingly. Such a holistic approach ensures that resources are allocated where they are most needed, potentially saving lives and reducing the overall devastation caused by the hurricane.

# ABSTRACT

This project focuses on Hurricane Ian, which struck the U.S. in September 2022, causing significant devastation. In the first phase of the project, we extracted demographic data for the regions impacted by the storm and conducted exploratory data analysis (EDA) using key measures such as population density, income levels, and infrastructure. The goal was to identify the most vulnerable areas and better understand the socio-economic and geographical factors that could influence the severity of the storm's impact. By analyzing these factors, we aim to provide actionable insights that can aid in resource allocation and disaster response efforts.

In the second phase, we developed an LSTM (Long Short-Term Memory) model to predict the future behaviour of Hurricane Ian using temporal data. By leveraging the storm's present state—such as its position and expected maximum wind speed—we were able to forecast the hurricane's path and intensity. This model processes sequential data and provides advisories based on the predicted behaviour, helping to improve the accuracy of the forecasts and provide more timely information for emergency response planning. The combination of demographic analysis and advanced machine learning models forms the foundation for more effective hurricane preparedness and response strategies.

# PROJECT METHODOLOGY

The project was divided in two phases each one withs its own objective

## PHASE-1

DATA EXTRACTION & EXPLORATORY DATA ANALYSIS

### *(i) DATA COLLECTION:*

In this project, data was collected from three primary sources to enable hurricane path prediction and vulnerability analysis:

1. **Hurricane Cone and Track Data**:
    - Data Source: National Hurricane Center (NHC)
    - Description: This source provided hurricane track and cone files for each advisory issued during Hurricane Ian in 2022. The track and cone data are essential for understanding the projected path of the hurricane and the areas potentially impacted over time.
2. **Population Demographic Data**:
    - Data Source: US Census Bureau
    - Description: This dataset includes demographic information such as age, sex, vehicle ownership, and other characteristics. This information is vital for identifying vulnerable populations in the hurricane's path based on demographics that may affect evacuation, shelter needs, and resource allocation.
3. **Population Shape Files**:
    - Data Source: US Census TIGER/Line Shapefiles
    - Description: These shape files include detailed census tracts that map the spatial distribution of the population. This geographic information enables the visualisation and analysis of population data within specific regions affected by the hurricane, which we performed using QGIS.

**Procedure**: The hurricane cone data was integrated with population demographic information by creating intersections within QGIS. For each advisory cone, we overlaid the population shape files and performed spatial joins to obtain demographic data specific to regions within the cone. The resulting intersected data for each advisory was then exported for further analysis.

*DATA-FORMAT:*

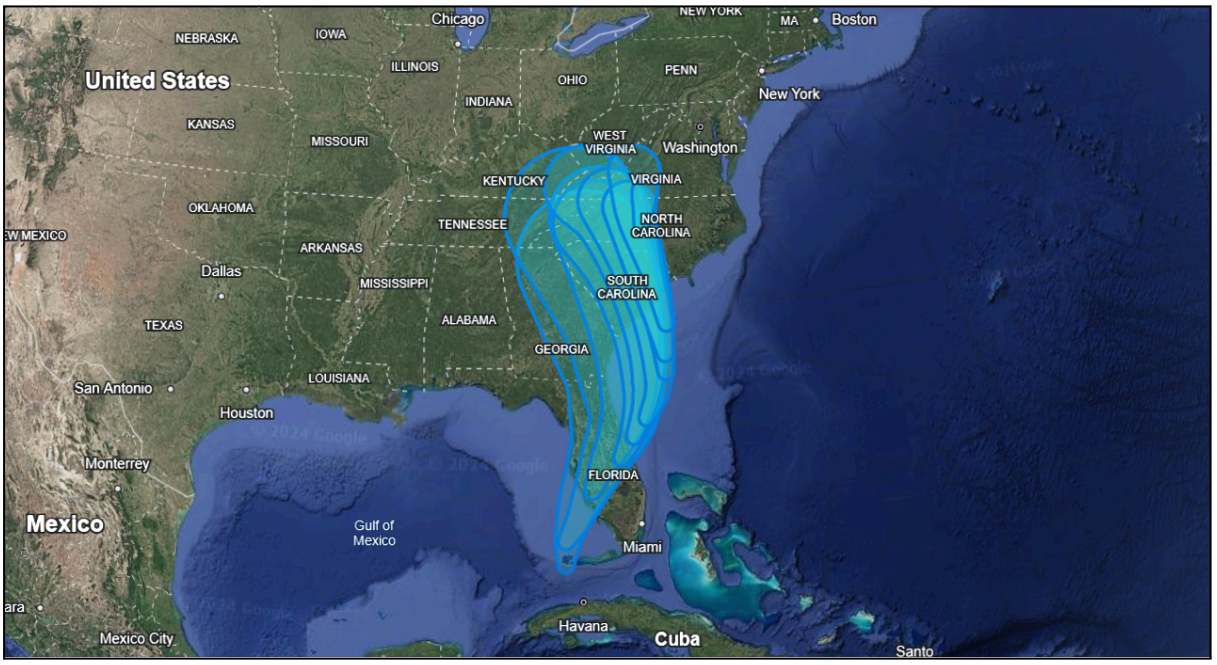
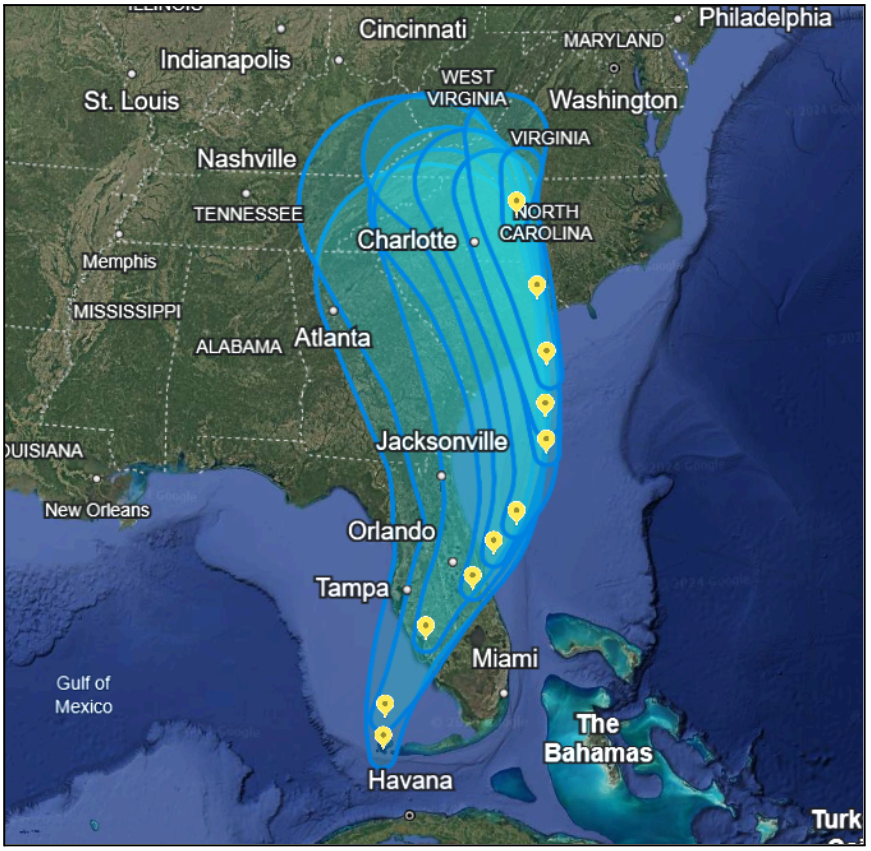



Figure: Ten most relevant hurricane cones plotted on GoogleEarth for reference

The figure on the Right elaborates more on the above image, also highlighting the actual path of the hurricane. Note that for each cone, the compressed section implies higher probability. Hence the locus of such sections is the actual path of the hurricane. This is shown by the yellow dots on the map.



### *(ii) Data Preprocessing:*

Following the data collection, the merged dataset required extensive cleaning to ensure that only relevant information was retained for analysis. The raw dataset contained numerous extraneous columns, including:

- **Unnecessary Demographic Details**: Columns such as "margin of error" and various age group breakdowns (e.g., population between 1–4 years, 4–14 years) added complexity without contributing meaningful insights for this project.

**Cleaning Process**:

- We streamlined the dataset by consolidating age data into broader categories: **Child, Youth, Adult**, and **Senior**. This classification helped simplify age demographics while retaining essential information for vulnerability analysis.
- Additionally, only essential **household income** and **vehicle ownership** data were preserved, as these factors significantly impact evacuation potential and resource needs during a hurricane.

The result of this data preprocessing was three clean, structured datasets focused on core demographic attributes relevant to hurricane vulnerability analysis.

### *(iii) Intersection Data Retrieval:*

This step involved two key components: learning QGIS functionalities for spatial intersections and linking intersection data with census data for meaningful analysis.

1. **Learning QGIS for Intersection Layers**:
    - To prepare for intersection-based analysis, we familiarized ourselves with QGIS—a Geographic Information System tool that allows spatial data manipulation and visualisation. QGIS was essential for creating intersection layers between the hurricane cone data and population shape files.
    - Using QGIS, we identified regions within each hurricane cone and extracted these intersections, which represent the geographic areas potentially impacted by each advisory.

2. **Linking Intersection Data to Population Data**:
    - For meaningful analysis, each intersection layer was linked to population data using unique identifiers or **census tract keys**. This allowed us to associate each hurricane cone intersection with corresponding demographic data.
    - Focusing on the state of Florida, where Hurricane Ian had significant impact, we linked each intersection to census tracts and then to

**county-level data**. By grouping the data at the county level, we made the analysis more interpretable and actionable, enabling insights on county-specific vulnerability.

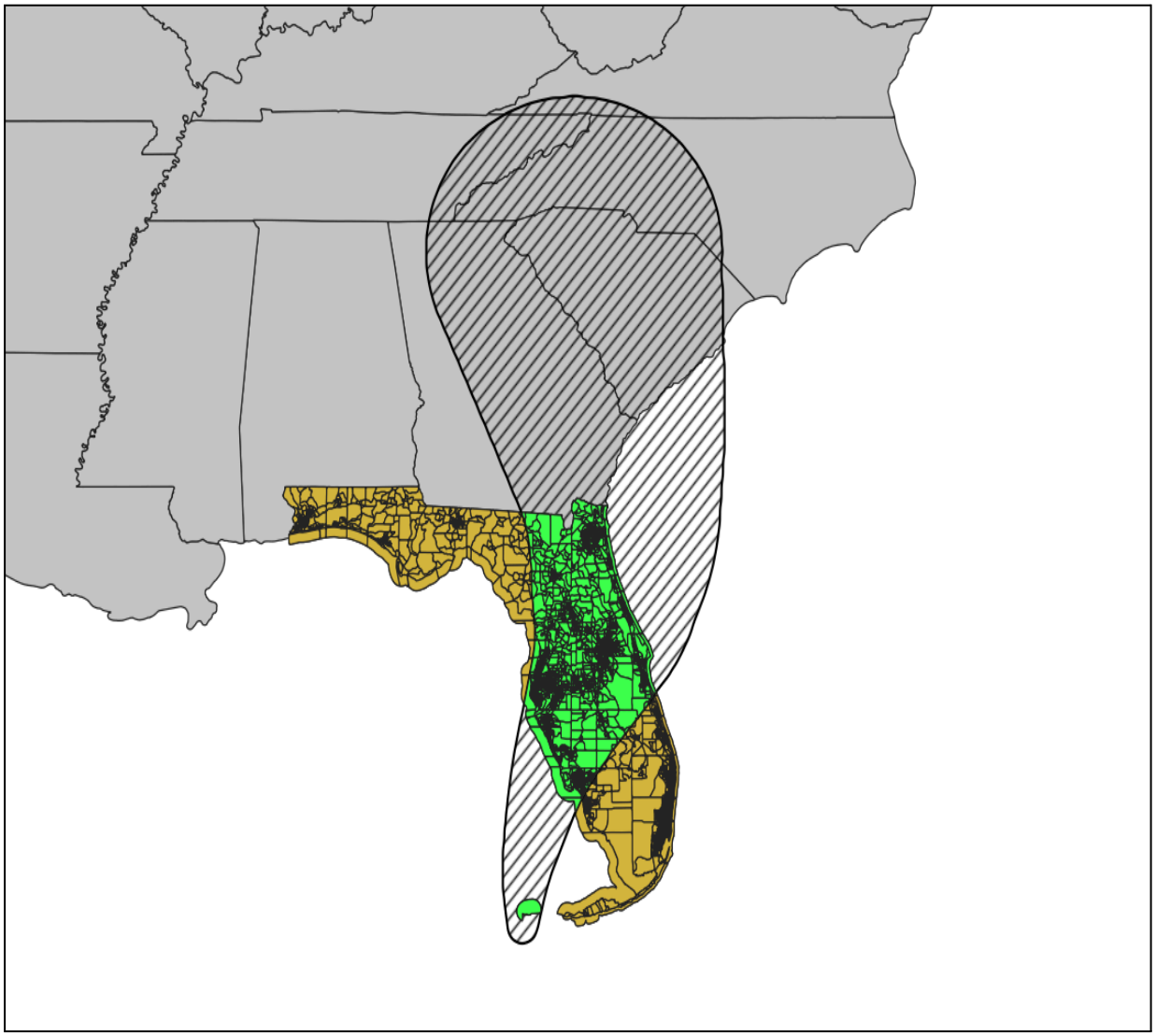


Figure: Intersection data preparation for a particular cone (Adv Num 19A here)

This intersection and linking process provided a cohesive dataset that mapped hurricane impact areas to relevant demographic data, facilitating county-level vulnerability assessment.

### *(iv) Vulnerable Counties Detection:*

After preprocessing the data, we moved to the critical step of creating features that would allow us to assess the vulnerability of different counties and calculate their aid scores. This stage involved the development of several metrics to quantify the needs of each county based on the factors of vehicle availability, income, and population dynamics.

**Vehicle Availability Index**: We started by constructing the **Vehicle Availability Index**, which is essential for understanding a county's ability to evacuate residents during a hurricane. The index was designed by considering the distribution of vehicle ownership among households.

- **Vehicle Ownership Categories**: The following vehicle availability categories were considered:
    - **No vehicle available**: Households that do not have any vehicle.
    - **1 vehicle available**: Households with one vehicle.
    - **2 vehicles available**: Households with two vehicles.
    - **3 or more vehicles available**: Households with three or more vehicles.
- For each county, and each household size, an intermediate vehicle index was calculated which assigned larger weights to no vehicle available and least weight to 3 or more vehicles available.
- **Household Size Weighting Formula**: To incorporate the household size into the **Vehicle Availability Index**, we assign the following weights to the intermediate household-wise vehicle index.
    - **1-person household**: Weight of **0.5**, as smaller households are more likely to evacuate efficiently.
    - **2-person household**: Weight of **0.7**, considering that larger households may need more logistical support.
    - **3-person household**: Weight of **0.9**, as these households may require more vehicles and resources.
    - **4 or more-person household**: Weight of **1.0**, reflecting the highest evacuation demand for larger households.
- **Vehicle Availability Index**: Calculated using the weighted sum of the number of vehicles available and the size of the families and scaled to be out of 100. The households with no vehicle and 4+ people would have the highest index since they need the most priority while evacuating.

| vehicle vailable | 2 vehicles available | 3 vehicles available | 4 or more vehicles available | 1-person household: | ... | 1-person household: - Weighted Sum of Fractions | 2-person household: - Weighted Sum of Fractions | 3-person household: - Weighted Sum of Fractions | 4-or-more-person household: - Weighted Sum of Fractions | 1-person household: - Weighted Sum | 2-person household: - Weighted Sum | 3-person household: - Weighted Sum | 4-or-more-person household: - Weighted Sum | Vehicle Need Score | Scaled Vehicle Need Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 44638 | 38552 | 13291 | 4353 | 38117 | ... | 0.749856 | 0.552651 | 0.474312 | 0.436041 | 0.187464 | 0.276326 | 0.355734 | 0.436041 | 1.255565 | 50.222600 |
| 2587 | 3277 | 1816 | 961 | 1539 | ... | 0.712313 | 0.523787 | 0.469350 | 0.333711 | 0.178078 | 0.261894 | 0.352012 | 0.333711 | 1.125695 | 45.027793 |
| 2823 | 3556 | 1362 | 694 | 2659 | ... | 0.726288 | 0.539145 | 0.429519 | 0.386032 | 0.181572 | 0.269572 | 0.322139 | 0.386032 | 1.159315 | 46.372610 |
| 91148 | 97550 | 33541 | 12900 | 75017 | ... | 0.725115 | 0.540278 | 0.442883 | 0.410161 | 0.181279 | 0.270139 | 0.332162 | 0.410161 | 1.193741 | 47.749643 |

**Vehicle Availability Index (Vehicle Need Score) Calculation Dataframe**

**Income-Based Vulnerability Scoring**: Next, we created an Income-Based Vulnerability Score to capture the socioeconomic vulnerability of each county. Counties with a higher proportion of low-income households were given a higher vulnerability score, as these households are often less equipped to prepare for and recover from natural disasters.

- **Define Income Ranges**: The income ranges are grouped into three categories: low , medium , and high income. For each group, specific columns in the dataset represent the number of households falling within each income range.
- **Summing Income Categories**: The total number of households in each income group (low, medium, and high) is calculated by summing the relevant columns across rows (i.e., for each county).
    - **Low Income = $0 to $49,999**
    - **Medium Income = $50,000 to $124,999**
    - **High Income = $125,000 to $199,999**
- **Calculating Income Fractions**: The fraction of households in each income category is calculated by dividing the number of households in that category by the total number of households in each county.
- **Poverty Rate Score**: The final Poverty Rate Score is calculated by weighting the fractions of households in the low, medium, and high-income categories. A higher weight was assigned to counties with a larger percentage of households with no vehicleThe weights are as follows:
    - Low-income fraction is weighted as **0.9**
    - Medium-income fraction is weighted as **0.65**
    - High-income fraction is weighted as **0.2**

| aphic Name | Estimate!! Total: | Less than $10,000 | 10, 000 to 14,999 | 15, 000 to 19,999 | 20, 000 to 24,999 | 25, 000 to 29,999 | 30, 000 to 34,999 | ... | 150, 000 to 199,999 | $200,000 or more | Low Income | Medium Income | High Income | Total Households | Low Income Fraction | Medium Income Fraction | High Income Fraction | Poverty Rate Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| achua ounty, lorida | 108597 | 10355 | 5808 | 4203 | 4719 | 5465 | 3562 | ... | 7794 | 8871 | 48480 | 37659 | 22458 | 108597 | 0.446421 | 0.346778 | 0.206801 | 68.588359 |
| Baker ounty, lorida | 9004 | 347 | 286 | 449 | 183 | 218 | 432 | ... | 827 | 629 | 3043 | 3870 | 2091 | 9004 | 0.337961 | 0.429809 | 0.232230 | 65.147712 |
| dford ounty, lorida | 9067 | 742 | 655 | 377 | 315 | 369 | 412 | ... | 601 | 227 | 4295 | 3516 | 1256 | 9067 | 0.473696 | 0.387780 | 0.138524 | 72.547700 |
| evard ounty, lorida | 246650 | 9338 | 7992 | 9033 | 9934 | 9958 | 9848 | ... | 19508 | 21326 | 85622 | 102361 | 58667 | 246650 | 0.347140 | 0.415005 | 0.237855 | 65.050030 |
| arlotte ounty, | 84671 | 3962 | 2610 | 3800 | 3838 | 3965 | 3683 | ... | 4660 | 4940 | 33310 | 36935 | 14426 | 84671 | 0.393405 | 0.436218 | 0.170377 | 69.349246 |

**Aid Need Score (Depending on age, sex, and population size):** Third, we calculated the Aid Need Score by combining demographic factors such as age dependency ratio, sex ratio, and population size to assess the relative vulnerability of each county.

- **Normalization of Population**: The population data is normalized to account for counties of varying sizes and densities. This ensures that the aid need score does not unfairly favour larger counties.

$$Normalized\ Population\ =\ \frac{Total\ population\ of\ county}{Total\ population\ of\ all\ counties}$$

- **Sex Ratio Adjustment**: The **Sex Ratio** is adjusted to reflect the proportion of males to females in each county. This is important because gender disparities can influence evacuation.

$$Required\ Sex\ Ratio\ =\ \frac{Male\ Youth + Male\ Adult}{Youth - Male\ Youth + Adult - Male\ Adult}$$

- **Age Dependency Ratio**: This ratio measures the proportion of dependent age groups (children and seniors) to the working-age population (youth and adults).

$$Age\ Dependency\ Ratio\ =\ \frac{Children + Seniors}{Working\ Age\ Population}$$

- **Aid Need Score Calculation**: The final Aid Need Score integrates these components.The formula for calculating the Aid Need Score is:

$$Aid\ Need\ Score\ =\frac{Age\ Dependency\ Ratio}{Required\ Sex\ Ratio}\ *\ Normalized\ Population$$

| ndency_Ratio | Children_Sex_Ratio | Youth_Sex_Ratio | Adult_Sex_Ratio | Senior_Sex_Ratio | Normalised_pop | Req_sexratio | Aid_Need_Score | Scaled_Aid_Need_Score |
|---|---|---|---|---|---|---|---|---|
| 41.795758 | 103.327607 | 93.408807 | 95.673112 | 79.399160 | 0.020805 | 0.946194 | 0.919013 | 17.204380 |
| 52.021959 | 111.970173 | 126.163724 | 124.881057 | 84.676705 | 0.002080 | 1.252724 | 0.086385 | 1.617178 |
| 51.950180 | 109.714549 | 160.433884 | 144.632977 | 72.768480 | 0.002069 | 1.487904 | 0.072234 | 1.352248 |
| 63.320684 | 105.169665 | 111.048562 | 96.960155 | 84.704105 | 0.045423 | 1.005164 | 2.861443 | 53.567669 |
| 99.365893 | 108.654402 | 115.188470 | 94.317793 | 92.502324 | 0.014124 | 0.986362 | 1.422847 | 26.636420 |
| 93.704749 | 110.424188 | 110.471262 | 93.052478 | 90.423629 | 0.011541 | 0.969417 | 1.115576 | 20.884155 |
| 54.006338 | 105.806256 | 106.213231 | 97.738785 | 84.081343 | 0.016337 | 1.000421 | 0.881912 | 16.509844 |
| 86.532801 | 104.230982 | 109.255835 | 98.857470 | 89.840410 | 0.028279 | 1.015265 | 2.410292 | 45.121876 |
| 61.502347 | 103.525692 | 121.462350 | 107.252469 | 89.006750 | 0.005194 | 1.110661 | 0.287605 | 5.384116 |
| 58.469794 | 81.653372 | 183.126338 | 147.313777 | 97.514885 | 0.002548 | 1.572653 | 0.094731 | 1.773415 |
| 50.625901 | 104.193867 | 101.702413 | 95.075583 | 75.179152 | 0.074057 | 0.970810 | 3.861916 | 72.297022 |

**Aid Need Score Calculation Dataframe**

**Final Aid Score:** It was taken considering vehicle availability index (VAI) to have a weight of 0.4, poverty rate score (PRS) to have 0.3, and aid need score (ANS) based on demographics to have 0.3. The weighted sum of the previously computed scores were added.

$$Final\ Aid\ Score\ =\ 0.4\ *\ VNI\ +\ 0.3\ *\ PRS\ +\ 0.3\ *\ ANS$$

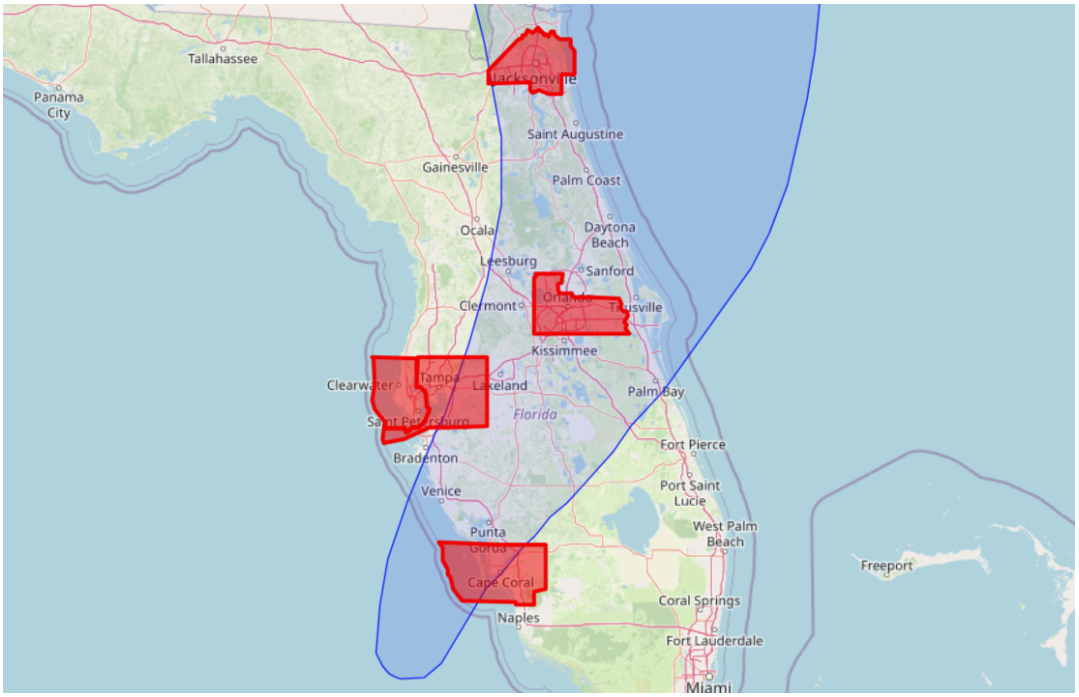
| | COUNTYFP | Scaled_Aid_Need_Score | Scaled Vehicle Need Score | Poverty Rate Score | Final_Aid_Score |
|---|---|---|---|---|---|
| **0** | 1 | 17.204380 | 50.222600 | 68.588359 | 45.826862 |
| **1** | 3 | 1.617178 | 45.027793 | 65.147712 | 38.040584 |
| **2** | 7 | 1.352248 | 46.372610 | 72.547700 | 40.719029 |
| **3** | 9 | 53.567669 | 47.749643 | 65.050030 | 54.685167 |
| **4** | 15 | 26.636420 | 48.304372 | 69.349246 | 48.117449 |
| **5** | 17 | 20.884155 | 48.126754 | 72.716762 | 47.330977 |
| **6** | 19 | 16.509844 | 46.922319 | 61.822368 | 42.268591 |
| **7** | 21 | 45.121876 | 49.832468 | 60.216116 | 51.534385 |
| **8** | 23 | 5.384116 | 47.080191 | 71.543329 | 41.910310 |
| **9** | 27 | 1.773415 | 52.912007 | 76.810373 | 44.739939 |
| **10** | 31 | 72.297022 | 51.878181 | 67.324579 | 62.637753 |

**Final Dataframe to Compute Weighted Sum of Each Type of Need Score**

After this, the highest 5 Final aid score counties were found and they were plotted on python using Geopandas and Folium. The process is listed below:

- **Identified High-Aid Counties**: From our merged dataset, we identified the top 5 Florida counties requiring aid based on their **Final_Aid_Score**. We generated unique **GEOID** codes for these counties by combining the Florida state code ('12') with each county's **COUNTYFP**.
- **Filtered County Data**: We filtered the county GeoDataFrame to keep only these top 5 **GEOIDs**, isolating the relevant counties for highlighting.
- **Initialized Folium Map**: Using Folium, we created a map centered on Florida with an optimal zoom level to display both the state and hurricane forecast cone.
- **Overlayed Hurricane Cone**: To show impacted areas, we plotted the hurricane forecast cone on the map in a transparent blue, highlighting regions likely to be affected.
- **Highlighted Top Counties**: We marked the top 5 counties with a distinct red border and thicker outline for visibility. A tooltip was added for each, showing the county name on hover.

The illustration is shown below:




**The top 5 affected counties (Duval, Orange, Lee, Hillsborough, and Pinellas), according to our Aid Need calculations, overlapping with the area of Cone 1**

Similarly, this process was repeated for all ten relevant cones. Finally we found that, out of all our predictions, those that were closest to the cones' heads matched very closely to the actual most affected counties in Florida. For example: Lee County (the region of Fort Myers), Hillsborough County, etc.

# PHASE-2

HURRICANE INTENSITY & PATH FORECASTING USING LSTM

## *(i) Data Collection and Temporal Data structure:*

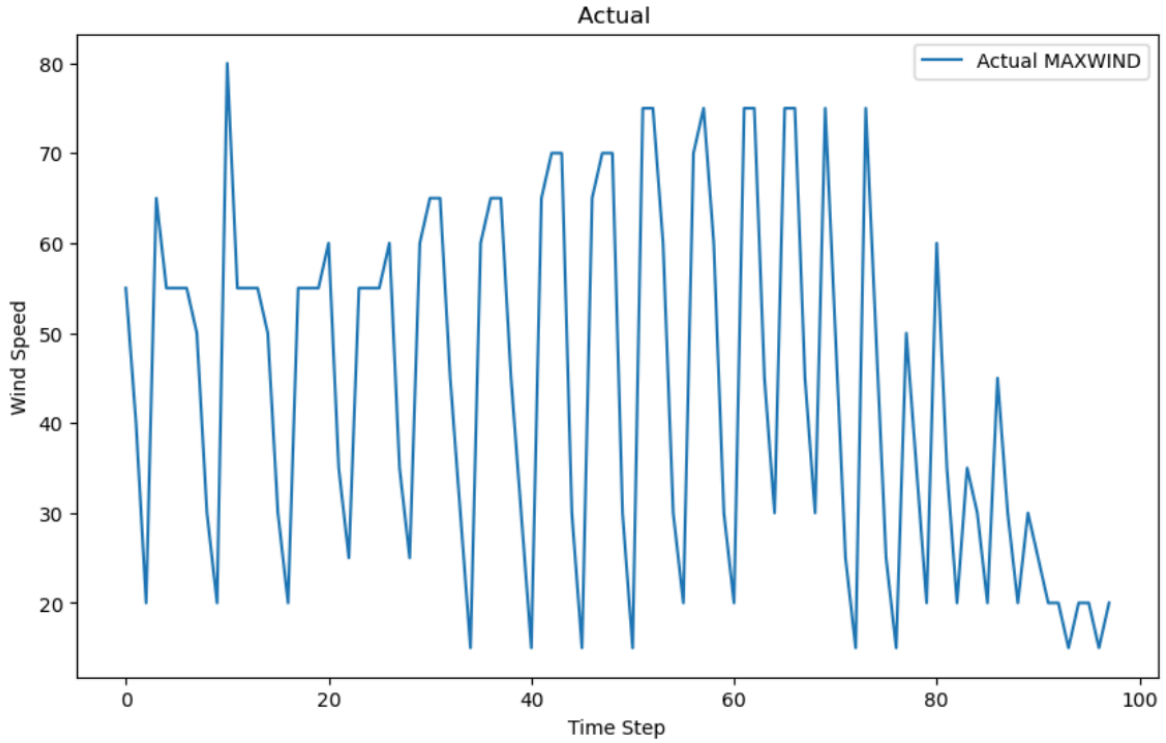
We merged the dbf files of all the available Cones of the hurricane data to get the following data frame format:

| ADVDATE | ADVISNUM | DATELBL | FLDATELBL | GUST | LAT | LON | MAXWIND | MSLP | SSNUM | TCDVLP | TAU | TCDIR | TCSPD | TIMEZONE | VALIDTIME |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 400 AM EDT Fri Sep 23 2022 | 1.0 | 5:00 AM Fri | 2022-09-23 2:00 AM Fri AST | 40 | 13.9 | -68.6 | 30 | 1006 | 0 | Tropical Depression | 0 | 290 | 11 | EDT | 23/0600 |
| 400 AM EDT Fri Sep 23 2022 | 1.0 | 2:00 PM Fri | 2022-09-23 2:00 PM Fri AST | 45 | 14.4 | -70.2 | 35 | 9999 | 0 | Tropical Storm | 12 | 9999 | 9999 | EDT | 23/1800 |
| 400 AM EDT Fri Sep 23 2022 | 1.0 | 2:00 AM Sat | 2022-09-24 2:00 AM Sat AST | 45 | 14.7 | -72.6 | 35 | 9999 | 0 | Tropical Storm | 24 | 9999 | 9999 | EDT | 24/0600 |
| 400 AM EDT Fri Sep 23 2022 | 1.0 | 2:00 PM Sat | 2022-09-24 2:00 PM Sat AST | 45 | 14.8 | -75.0 | 35 | 9999 | 0 | Tropical Storm | 36 | 9999 | 9999 | EDT | 24/1800 |
| 400 AM EDT Fri Sep 23 2022 | 1.0 | 2:00 AM Sun | 2022-09-25 2:00 AM Sun AST | 50 | 15.5 | -77.1 | 40 | 9999 | 0 | Tropical Storm | 48 | 9999 | 9999 | EDT | 25/0600 |

Explanation of each feature:

| Feature | Description |
|---|---|
| ADVDATE | Advisory date and time in EDT (Eastern Daylight Time). |
| ADVISNUM | Advisory number, indicating the sequence of advisories. |
| DATELBL | Simplified date and time label for the advisory forecast. |
| FLDATELBL | Full date and time label with time zone. |
| GUST | Wind gust speed in knots (maximum gust recorded). |
| LAT | Latitude of the storm's center during the advisory. |
| LON | Longitude of the storm's center during the advisory. |
| MAXWIND | Maximum sustained wind speed in knots. |
| MSLP | Mean Sea Level Pressure in millibars (lower values indicate stronger storms). |
| SSNUM | Storm system number, uniquely identifying the storm. |
| TCDVLP | Tropical Cyclone Development stage (e.g., Tropical Depression, Tropical Storm). |
| TAU | Forecast lead time from the initial advisory issuance in hours. |
| TCDIR | Tropical Cyclone Direction in degrees (9999 may indicate missing data). |
| TCSPD | Tropical Cyclone Speed in knots (9999 indicates missing data). |
| TIMEZONE | Time zone of the advisory (e.g., EDT). |
| VALIDTIME | Valid time of the forecast in day/hour format (e.g., 23/0600 for 23rd of month at 0600 hours). |

DATA FORMAT:




Every step down signifies a new advisory. (Above plot just shows a few percentage of the dataset). We found this format to be of less significance and hence did appropriate transformations.

## *(ii) Data Preprocessing and Feature Engineering:*

APPROACH A
Initial

**Grouping by ADVISNUM:**

- Grouped data by ADVISNUM (advisory number) to process each advisory separately.
- This enables us to analyze the storm's evolution within each advisory period, capturing distinct storm characteristics.

**Sorting Data by TAU:**

- Ensured the data is ordered by TAU (forecast lead time) within each advisory group.
- This step is essential to calculate time-based derivatives accurately, as TAU represents the temporal progression of storm conditions.

**Extracting Key Features for Each Advisory**:

- For each advisory group, extracted the first row's values of LAT (latitude), LON (longitude), MSLP (mean sea level pressure), and MAXWIND (W, initial maximum wind speed).
- These values provide a snapshot of the storm's initial conditions for each advisory, serving as a baseline for understanding subsequent changes.

**Calculating Maximum Wind Speed (WMAX)**:

- Computed WMAX, the maximum MAXWIND value within each advisory period.
- WMAX represents the peak wind intensity during the advisory, a critical indicator of storm strength.

**Calculating dW/dT (First Derivative of Wind Speed with Respect to Time)**:

- Calculated the rate of change of wind speed (dW/dT) across each advisory by taking the difference in MAXWIND over the difference in TAU.
- **Significance**: dW/dT provides insights into how quickly the storm is intensifying or weakening over time, which is crucial for predicting rapid intensification or dissipation.

**Calculating d2W/dT2 (Second Derivative of Wind Speed with Respect to Time)**:

- Calculated d2W/dT2, the rate of change of dW/dT over time.
- **Significance**: This second derivative highlights the acceleration or deceleration of storm intensification, offering a deeper understanding of storm dynamics and aiding in forecasting sudden changes in storm intensity.

**Creating a New Transformed DataFrame**:

- Constructed a new DataFrame with key features for each advisory, including ADVISNUM, LAT, LON, MSLP, W, dW/dT, d2W/dT2, WMAX, and Timestamp.
- **Importance**: This transformation condenses each advisory's data into a single row, simplifying analysis and enabling efficient exploration of storm behaviour over multiple advisories.

This preprocessing and feature engineering enhance the dataset's utility by providing insights into storm intensity and its rate of change, helping identify trends in storm growth, which is vital for forecasting purposes.

- **Converting ADVDATE to Datetime Format**:
    - Converted the ADVDATE column to a standardised datetime format (Timestamp), allowing for precise time-based calculations.

- **Importance**: This conversion ensures that time-related analyses, such as sorting or calculating time differences, are accurate. Consistent datetime formatting is crucial when working with time series data.

- **Sorting by ADVISNUM and Timestamp**:
  - Sorted the DataFrame by ADVISNUM (advisory number) and Timestamp to maintain the correct chronological order within each advisory.
  - **Significance**: This step preserves the temporal sequence of data points, which is essential for accurately calculating time derivatives (such as dW/dT and d2W/dT2). Proper sorting ensures that each advisory's progression over time is reflected in the correct order, enabling meaningful analysis of time-based trends.

These steps improve data integrity and enable accurate time-based analysis, which is vital for understanding and forecasting storm intensity changes over time.

| ADVISNUM | LAT | LON | MSLP | W | dW/dT | WMAX | TimeElapsed |
|---|---|---|---|---|---|---|---|
| 1.0 | 13.9 | -68.6 | 1006 | 30 | 0.520833 | 95 | 0.0 |
| 2.0 | 14.2 | -70.1 | 1006 | 30 | 0.598958 | 100 | 7.0 |
| 3.0 | 14.7 | -71.3 | 1006 | 30 | 0.625000 | 100 | 13.0 |
| 3.5 | 14.8 | -71.5 | 1006 | 30 | 0.625000 | 100 | 16.0 |
| 4.0 | 14.8 | -72.0 | 1005 | 35 | 0.598958 | 100 | 19.0 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 34.5 | 35.4 | -79.7 | 998 | 35 | -0.625000 | 35 | 190.0 |
| 35.0 | 35.7 | -79.8 | 1001 | 30 | -0.416667 | 30 | 193.0 |
| 36.0 | 36.4 | -79.9 | 1006 | 20 | -0.208333 | 20 | 199.0 |
| 37.0 | 36.8 | -78.8 | 1008 | 20 | 0.000000 | 20 | 205.0 |
| 38.0 | 37.3 | -78.2 | 1010 | 15 | 0.416667 | 20 | 211.0 |

This the Final Dataset format we used to train the model

APPROACH B
Improved

Incorporating the rate of range of Lat and Lon:

**Calculating deltaLAT (Change in Latitude over Time)**:

- Computed deltaLAT by finding the rate of change in latitude (LAT) with respect to TAU (forecast lead time).
- **Physical Significance**: deltaLAT represents the north-south movement of the storm per unit time. It provides insights into the storm's trajectory and directional movement, helping to predict the potential path and areas that might be affected.

**Calculating deltaLON (Change in Longitude over Time):**

- Calculated deltaLON similarly by finding the rate of change in longitude (LON) with respect to TAU.
- Physical Significance: deltaLON indicates the east-west movement of the storm per unit time. Together with deltaLAT, it helps track the storm's horizontal displacement and provides information about its travel speed and direction.

These features (deltaLAT and deltaLON) enhance the dataset by providing information about the storm's movement in both latitude and longitude directions, which is essential for understanding and forecasting the storm's path.

| ADVISNUM | LAT | LON | MSLP | W | dW/dT | d2W/dT2 | WMAX | deltaLAT | deltaLON | Timestamp | TimeElapsed |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.0 | 13.9 | -68.6 | 1006 | 30 | 0.520833 | 0.001240 | 95 | 0.089062 | -0.133333 | 2022-09-23 04:00:00 | 0.0 |
| 2.0 | 14.2 | -70.1 | 1006 | 30 | 0.598958 | 0.004960 | 100 | 0.094271 | -0.121354 | 2022-09-23 11:00:00 | 7.0 |
| 3.0 | 14.7 | -71.3 | 1006 | 30 | 0.625000 | 0.003720 | 100 | 0.095313 | -0.116667 | 2022-09-23 17:00:00 | 13.0 |
| 3.5 | 14.8 | -71.5 | 1006 | 30 | 0.625000 | 0.003720 | 100 | 0.094271 | -0.114583 | 2022-09-23 20:00:00 | 16.0 |
| 4.0 | 14.8 | -72.0 | 1005 | 35 | 0.598958 | 0.001240 | 100 | 0.102083 | -0.106250 | 2022-09-23 23:00:00 | 19.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 34.5 | 35.4 | -79.7 | 998 | 35 | -0.625000 | -0.034722 | 35 | 0.108333 | 0.020833 | 2022-10-01 02:00:00 | 190.0 |
| 35.0 | 35.7 | -79.8 | 1001 | 30 | -0.416667 | NaN | 30 | 0.091667 | 0.016667 | 2022-10-01 05:00:00 | 193.0 |
| 36.0 | 36.4 | -79.9 | 1006 | 20 | -0.208333 | -0.034722 | 20 | 0.012500 | 0.075000 | 2022-10-01 11:00:00 | 199.0 |
| 37.0 | 36.8 | -78.8 | 1008 | 20 | 0.000000 | NaN | 20 | 0.033333 | 0.050000 | 2022-10-01 17:00:00 | 205.0 |
| 38.0 | 37.3 | -78.2 | 1010 | 15 | 0.416667 | NaN | 20 | -0.008333 | 0.058333 | 2022-10-01 23:00:00 | 211.0 |

This the Final Dataset format we used to train the model

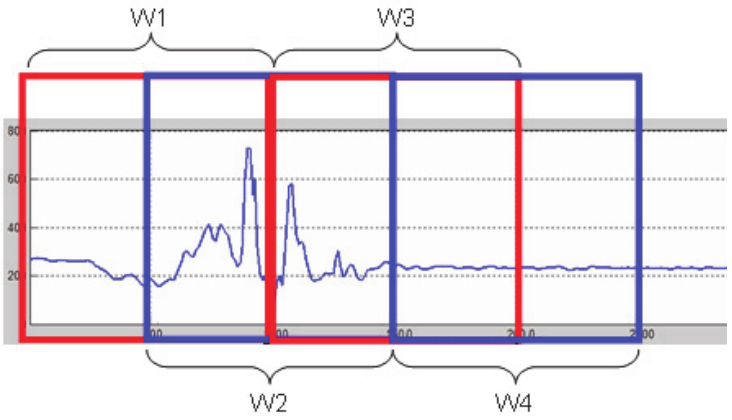
## *(iii) Model Architecture and Training:*

Why LSTM?

**LSTM (Long Short-Term Memory)**:

- LSTM is a type of recurrent neural network (RNN) designed to process sequential data and capture long-term dependencies.
- Unlike standard RNNs, LSTMs are capable of remembering information over long periods and can selectively retain or forget data through specialised "gates."
- **Key Components**:
    - **Input Gate**: Decides what new information should be added to the cell state.
    - **Forget Gate**: Determines what information should be discarded from the cell state.
    - **Output Gate**: Controls what information is output from the cell state.
- **Significance**: LSTMs are highly effective for time series data and tasks involving sequential patterns (e.g., speech recognition, stock price prediction) as they can learn dependencies over time, which regular neural networks cannot handle as efficiently.

DATA-PREPARATION:




- **Sliding Window Approach**:
    - Defined a function create_dataset to implement a sliding window approach with a specified time step (time_step=3).
    - **Purpose**: This approach allows the model to learn from a sequence of three previous time steps to predict the next time step.
    - **Process**: For each sequence of three time steps, the function creates input data (X) and targets (y), where each target is the next time step following the input sequence.
    - Very efficient for a temporal data format.
- **Splitting Data into Training and Testing Sets**:

    - Divided the dataset into 80% for training and 20% for testing.
    - **Importance**: Separating training and testing data helps evaluate the model's performance on unseen data and avoids overfitting.
- **Reshaping Data for LSTM**:
    - Reshaped X_train and X_test to fit the LSTM model's expected input shape: [samples, time steps, features].
    - **Reason**: LSTM networks expect 3D input data, where each sample contains a sequence (time steps) of multiple features.

MODEL- ARCHITECTURE:

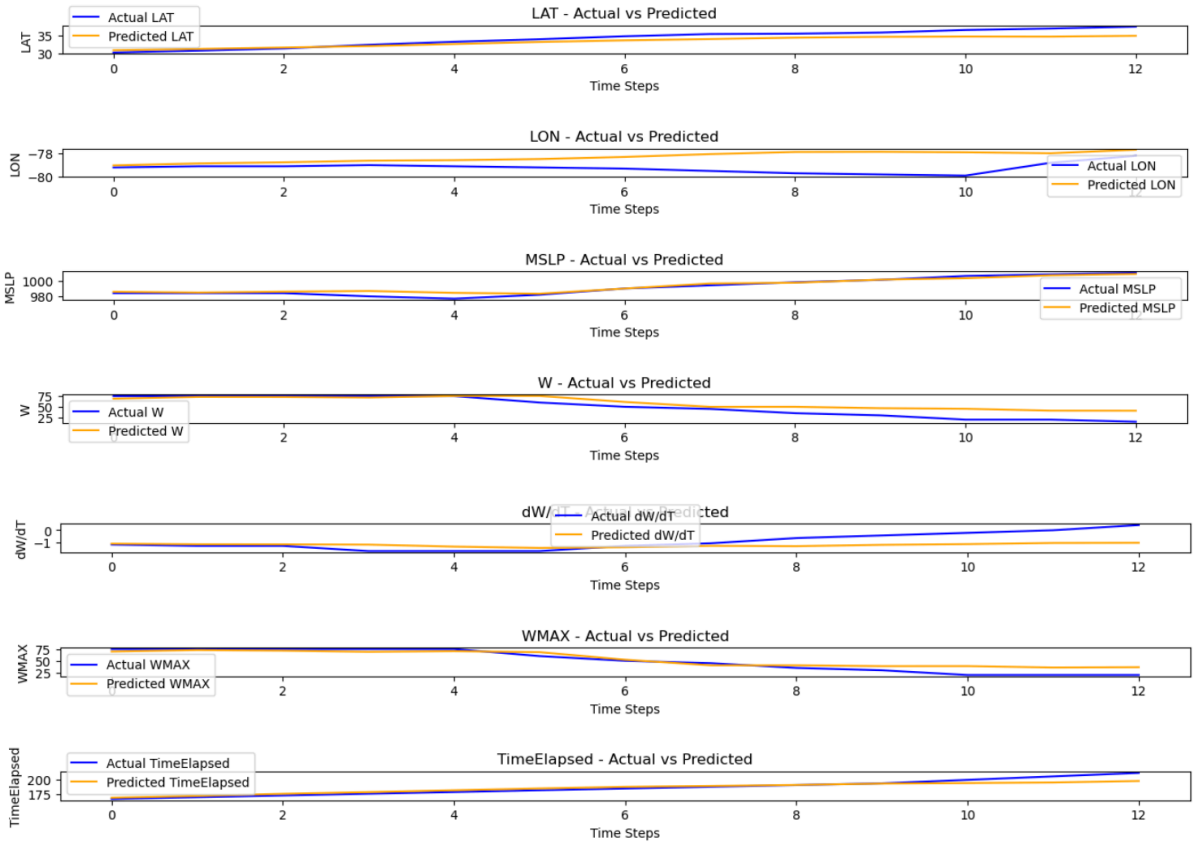
- **LSTM Layer**:
    - Added an LSTM layer with 100 units and return_sequences=False, meaning it outputs only the last time step's predictions for each input sequence.
    - **Purpose**: LSTM layers are ideal for sequential data, as they can capture temporal dependencies and patterns over time.
- **Dense Output Layer**:
    - Added a Dense layer with the number of units equal to the number of features (e.g., 7), to output predictions for each feature at the next time step.
    - **Significance**: This ensures that the model predicts the values for each feature at the next time step.

MODEL COMPILATION & TRAINING:

- **Compilation**:
    - Used the Adam optimizer and Mean Squared Error (MSE) as the loss function.
    - **Reason**: Adam is efficient for gradient-based optimization, and MSE is a standard choice for regression problems, penalising large errors.
- **Model Training**:
    - Trained the model for 500 epochs with a batch size of 16, using both X_train and y_train for training, and X_test and y_test for validation.
    - **Purpose**: Training over multiple epochs allows the model to learn patterns in the data, and using validation data helps monitor its generalisation capability.

# RESULTS

## Actual vs Predicted Features




APPROACH A

```
Mean Absolute Error for LAT: 1.0911216442401588
Mean Absolute Error for LON: 0.9199012169471147
Mean Absolute Error for MSLP: 2.229205791766818
Mean Absolute Error for W: 11.4795775780311
Mean Absolute Error for dW/dT: 0.5005340178807577
Mean Absolute Error for WMAX: 7.852262643667368
Mean Absolute Error for TimeElapsed: 3.950182401216945
```

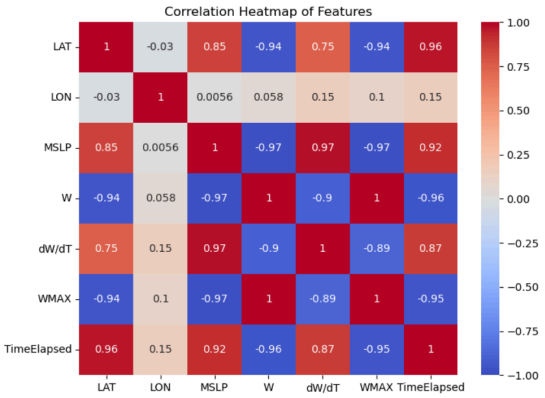
APPROACH B

```
Mean Absolute Error for LAT: 1.4213334303635814
Mean Absolute Error for LON: 0.6779948307917661
Mean Absolute Error for deltaLAT: 0.04789166121910775
Mean Absolute Error for deltaLON: 0.04732960927435477
Mean Absolute Error for MSLP: 7.224219689002395
Mean Absolute Error for W: 13.496373396653395
Mean Absolute Error for dW/dT: 0.49559488357641757
Mean Absolute Error for WMAX: 7.519734309269832
Mean Absolute Error for TimeElapsed: 4.9134603647085315
```

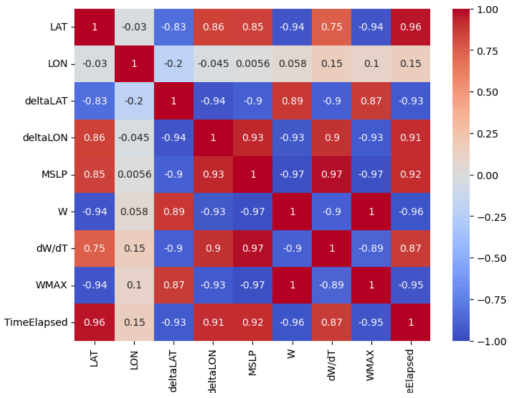
We can see that in approach B, the results are improving overall for position as well as the max predicted wind speed.

Correlation Heatmap of Features

APPROACH A




APPROACH B

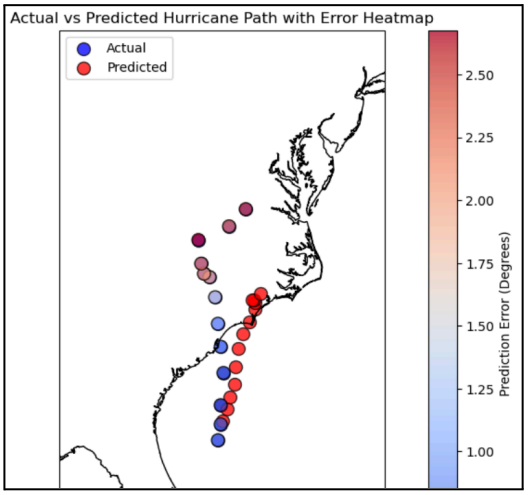



We see that for Wmax , almost all the features created hold a good relation.Here are a few inferences :

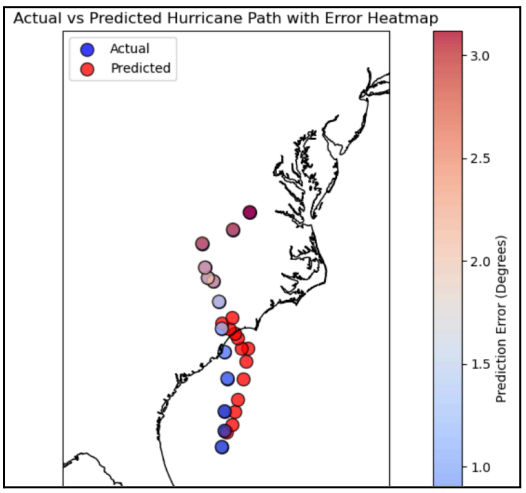
| Feature | Correlation with WMAX | Inference |
|---|---|---|
| **W** | High positive | Strong predictor for `WMAX`, as higher current wind speeds often indicate higher maximum potential. |
| **MSLP** | High negative | Lower Mean Sea Level Pressure is associated with stronger winds, making it a key factor for predicting `WMAX`. |
| **deltaLAT** | Moderate negative | North-south movement affects storm intensity, possibly due to geographical or oceanic influences. |
| **deltaLON** | Moderate positive | East-west movement also impacts intensity, indicating storm path can influence peak wind speeds. |
| **dW/dT** | Moderate positive | Rate of change in wind speed suggests the storm's intensification trend, contributing to `WMAX` prediction. |
| **TimeElapsed** | Moderate negative | Negative correlation indicates potential weakening over time after reaching peak intensity. |

## Actual vs Predicted Path

APPROACH A




APPROACH B



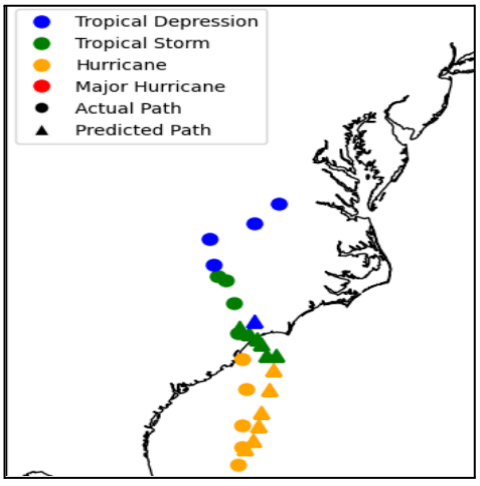

We can clearly see that by incorporating the rates of change of Latitude and Longitude, we are actually able to predict the path better and also, very accurately, the region where it will hit the land.We can assess an advisory cone of our prediction and the vulnerable counties as shown in Phase 1.

## Actual vs Predicted Intensity

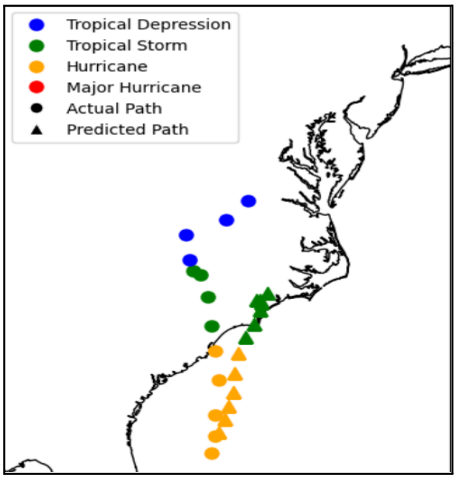
CLASSIFIER:

```python
def classify_intensity_knots(WMAX, MSLP, dW_dT):
    if WMAX >= 96 or (MSLP < 945 and dW_dT > 2.5):  # Major Hurricane conditions
        return 'Major Hurricane'
    elif WMAX >= 64 or (MSLP < 980 and dW_dT > 1.5):  # Hurricane conditions
        return 'Hurricane'
    elif WMAX >= 34:  # Tropical Storm threshold
        return 'Tropical Storm'
    else:
        return 'Tropical Depression'
```

APPROACH A




APPROACH B




We are able to predict the type of storm and its intensity when it hits land

THIS IS THE COMPLETE IMPLEMENTATION BASED ON OUR UNDERSTANDING AND INTUITION.

## APPLICATION

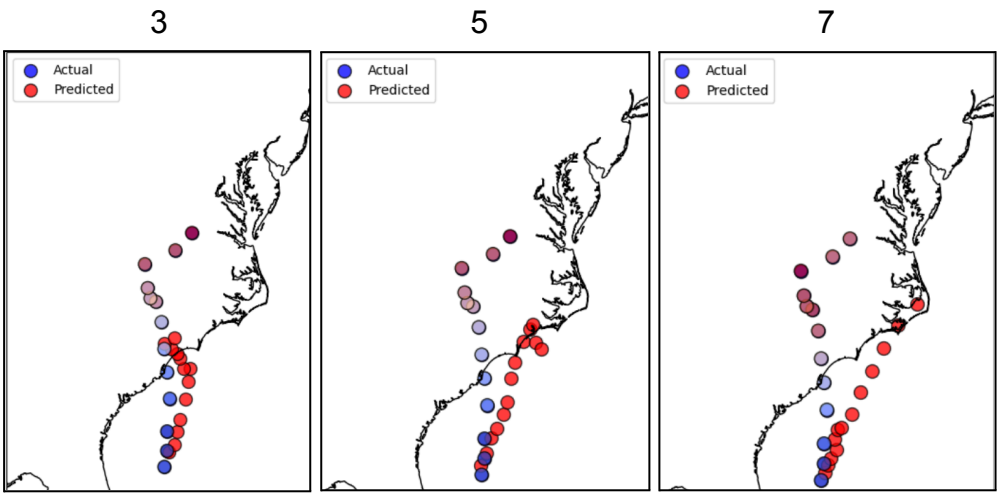
These forecasted graphs and the vulnerable county extraction script will be able to provide crucial information to the concerned authorities ,in time, to help provide preventive aid.

## MAJOR CHALLENGES

1. Firstly the major issue is the current used data set . The size of the final data set that we used to train has just 68 rows which we actually got from a total of 495 rows that too only from one hurricane data.
2. Secondly the issue we faced is the choice of window length. Intuitively the prediction accuracy should improve if we increase the window size but we are getting some conflicting results.

We used window size 3,5,7 for the model and compared their results-

| Feature | Window Size 3 MAE | Window Size 5 MAE | Window Size 7 MAE |
|---|---|---|---|
| LAT | 1.4213 | 1.2569 | 1.6972 |
| LON | 0.6780 | 1.0879 | 1.1966 |
| deltaLAT | 0.0479 | 0.0546 | 0.0489 |
| deltaLON | 0.0473 | 0.0649 | 0.0508 |
| MSLP | 7.2242 | 4.0465 | 5.1440 |
| W | 13.4967 | 11.7775 | 22.3389 |
| dW/dT | 0.4956 | 0.4481 | 0.6257 |
| WMAX | 7.5197 | 5.1854 | 12.0698 |
| TimeElapsed | 4.9135 | 3.7640 | 5.3650 |



We can see that for-

- Path prediction -window size 3 is best
- Max Wind and intensity prediction - window size 5 is best
- Window size 7 performs bad for both cases possibly due to lesser data points to train from
- Overall, we can infer that there will be an optimised window length

## FUTURE WORK

To further enhance the **accuracy and robustness** of the hurricane intensity forecasting model, several extensions and improvements are planned. A key direction involves incorporating **historical hurricane data** from the Atlantic Ocean basin, such as hurricanes **Fiona**, **Nicole**, and others, which can be obtained from the **NHC hurricane advisory dataset**. This will allow the model to leverage a **larger dataset**, increasing the reliability of predictions by learning patterns not only from individual hurricanes but also from a broader range of past events.

Another improvement will involve expanding the **sliding window** approach. While the current model uses three previous advisories to predict the next, the incorporation of historical data will enable the use of **larger input sequences**, which could significantly enhance the model's performance in capturing **long-term trends** in hurricane behaviour. This will allow for more accurate forecasts, particularly for **long-duration storms**, as the model will have access to a wealth of diverse hurricane patterns.

The evaluation of the model's forecast accuracy will continue to focus on **Mean Absolute Error (MAE)**, comparing the predicted values with the actual advisories. The aim is to assess improvements in the short-term (next advisory) as well as over extended time horizons (48-72 hours), which could prove valuable for **disaster preparedness and response**.

Additionally, we plan to explore the integration of **external environmental factors**, such as **sea surface temperature maps** and **atmospheric pressure distributions**. **Convolutional Neural Networks (CNNs)** will be employed to convert these temperature and pressure maps into **feature vectors**, which can then be merged with existing hurricane data in the **LSTM model**. This **multimodal approach** could enable the model to better understand the **dynamic atmospheric conditions** influencing hurricane intensification.

Finally, future work will also focus on improving the model's ability to handle **uncertainty**. We aim to implement **probabilistic forecasting methods**, which would provide a range of possible outcomes for hurricane intensity rather than a single deterministic forecast. This approach will help account for the inherent

**unpredictability** in weather systems, offering a more comprehensive understanding of potential hurricane scenarios and improving decision-making for **disaster management**.

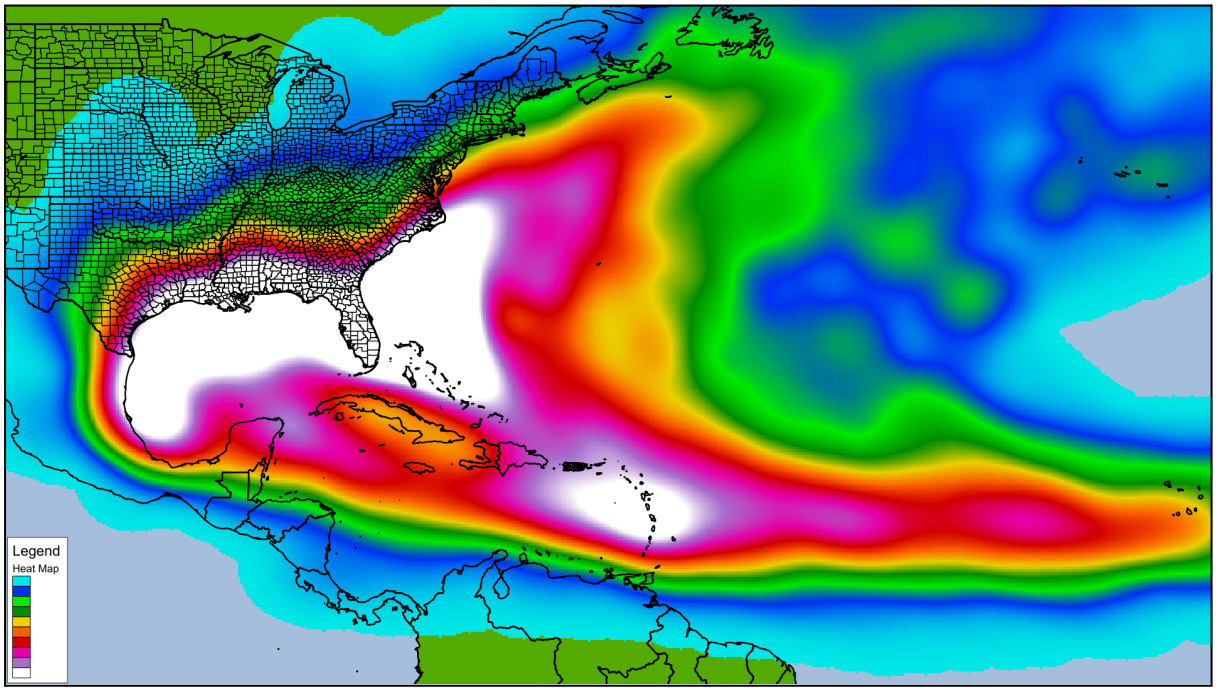



Figure: An example heat map of wind speed data, which can be converted using CNNs into meaningful information for hurricane intensity prediction, Ref: Texas Hurricane

# SUMMARY

This project aimed to forecast Hurricane Ian's path and intensity using an integrated machine-learning approach that combines socio-demographic vulnerability assessment with advanced time-series modelling. The project was conducted in two main phases:

1. **Vulnerability Analysis**:
    - We collected and analysed demographic data (age, income, vehicle ownership, etc.) from sources such as the US Census Bureau and population shape files to assess socio-economic vulnerabilities across impacted counties in Florida. Using QGIS, we intersected hurricane cone data with demographic layers to create a cohesive dataset that maps areas affected by each advisory to corresponding demographic information. Key metrics, such as the Vehicle Availability Index, Income-Based Vulnerability Score, and Aid Need Score, were developed to rank counties based on their evacuation needs and socio-economic resilience. The top five most vulnerable counties were identified and visualised with hurricane advisory overlays, providing critical insights for targeted disaster response.
2. **Hurricane Path and Intensity Forecasting**:
    - In this phase, we developed a Long Short-Term Memory (LSTM) model to predict the path and intensity of Hurricane Ian by analysing sequential data. Feature engineering involved extracting and transforming temporal features from advisory data, such as changes in wind speed, pressure, latitude, and longitude, to capture storm dynamics. A sliding window approach was implemented to train the model on historical advisories, achieving a balance between short- and long-term prediction accuracy. Our model demonstrated significant accuracy in predicting both the path and the intensity of the hurricane under different window lengths, providing advisories that closely matched actual outcomes in terms of intensity and trajectory.
3. **Challenges and Future Improvements**:
    - The project faced limitations, such as the dataset's relatively small size, which constrained prediction accuracy, especially with longer window sizes. We plan to enhance the model by incorporating additional historical data from other hurricanes and expanding the feature set with environmental factors, such as sea surface temperature and atmospheric pressure. Further improvements include applying probabilistic forecasting to better handle uncertainties in storm behaviour, offering a range of potential outcomes rather than a single deterministic forecast.