Gabriel Rosario - 08/02/2024

FINAL REPORT:
STORM PREDICTION ANALYSIS

1. PROBLEM STATEMENT

With the progress of the year, every country around the world is paying more attention to the climate change and its sequels. Many scientists is warning that with climate change the recurrent of Atlantic Ocean storms are trending upwards and the severity of the storms are increasing as well, in consequence, the preparation, mitigation and restoration task of different entities (governmental, private and public) needs to account these trend increase.

The importance of Atlantic storm preparation can be understood through several short- and long-term key factors that influence economic stability and growth, here are some of the main reasons:

- Protecting lives and reducing humanitarian costs,
- Mitigation of economic losses,
- Maintaining business continuity,
- Preserving public finances,
- Risk and Insurance management,
- Long-term economic resilience.

These factors collectively contribute to a more stable and prosperous economy, demonstrating the importance of proactive measures in the face of natural disasters under the circumstance of climate change increasing frequency and severity of the storms.

2. DATA WRANGLING

This dataset was downloaded from the following Kaggle source:

https://www.kaggle.com/datasets/christinezinkand/storms?select=storms.csv

It contains a collection of records of storms name, year, month, day, hour, latitude, longitude, status, category, wind, pressure, tropical storm and hurricane force diameter.

The shape of this dataset is 13 columns and 19,066 observations (no missing values).

The goal is to clean and prepare the dataset to create a model that predicts how many storms for each category are expected for the two upcoming years (2022 and 2023).

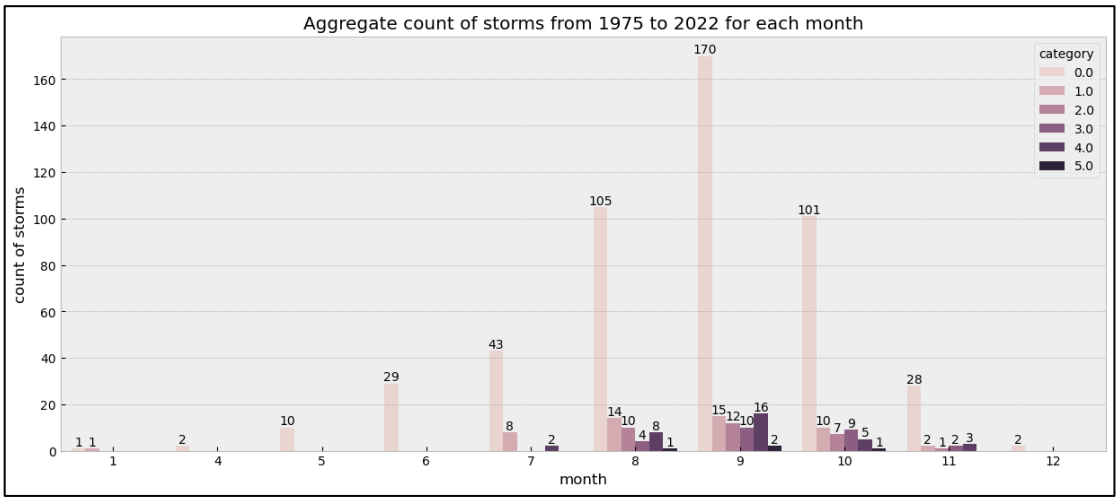
To better explore the data we perform the following steps to clean and transform the different features:

- Filling empty values (Null) for storm category according to the wind speed (which is used to determine the category of storms),
- Filling empty values (Null) for tropical storm force diameter and hurricane force diameter using the mean for each feature.
- Drop duplicated storms (using name and status features),
- Create categories for tropical depression, tropical storm and hurricane observations,
- Create cumulative features for each storm category,
- Create subset of data to compare all dataset and choose the most representative subset for model evaluation.

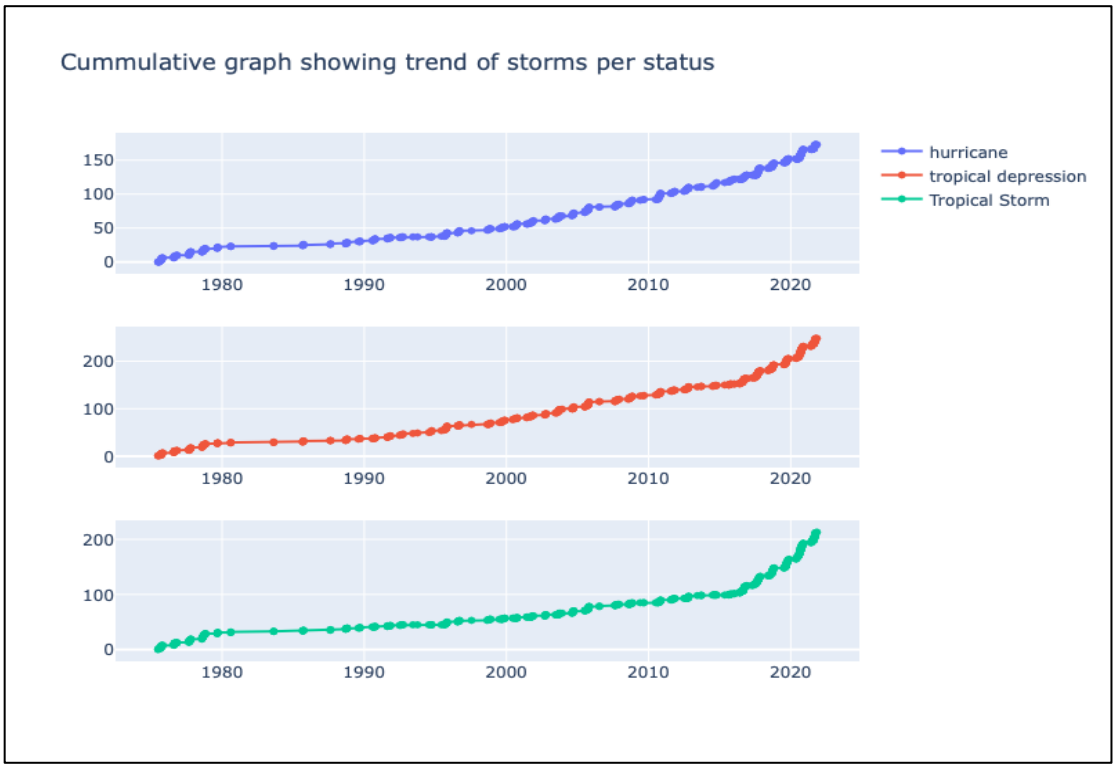
3. EXPLORATORY DATA ANALYSIS AND PREPROCESSING

In this section, we present visual representations of the data to uncover patterns, understand its structure, detect anomalies, and insights that might not be immediately apparent through summary statistics alone. We use various visualization techniques to explore the dataset comprehensively.

To understand the distribution over time of the storms per category, we show the following histograms which show the total amount of storms per month from 1975 to 2021,



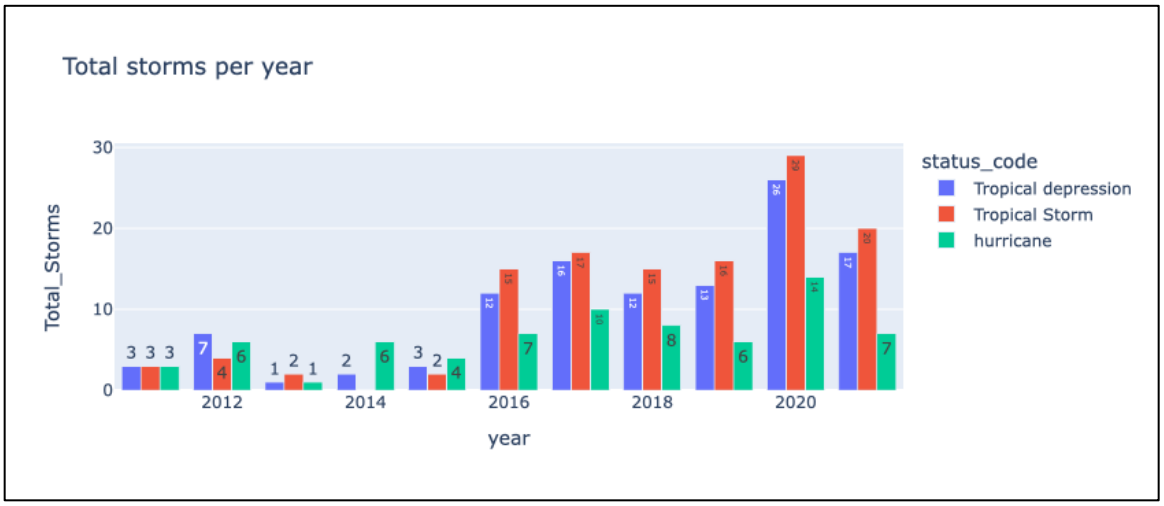
The above graph shows a well-defined seasonal behavior, with a peak of storm quantity in September. To identify what is the storm trend over the years, we use a line plot as shown below:

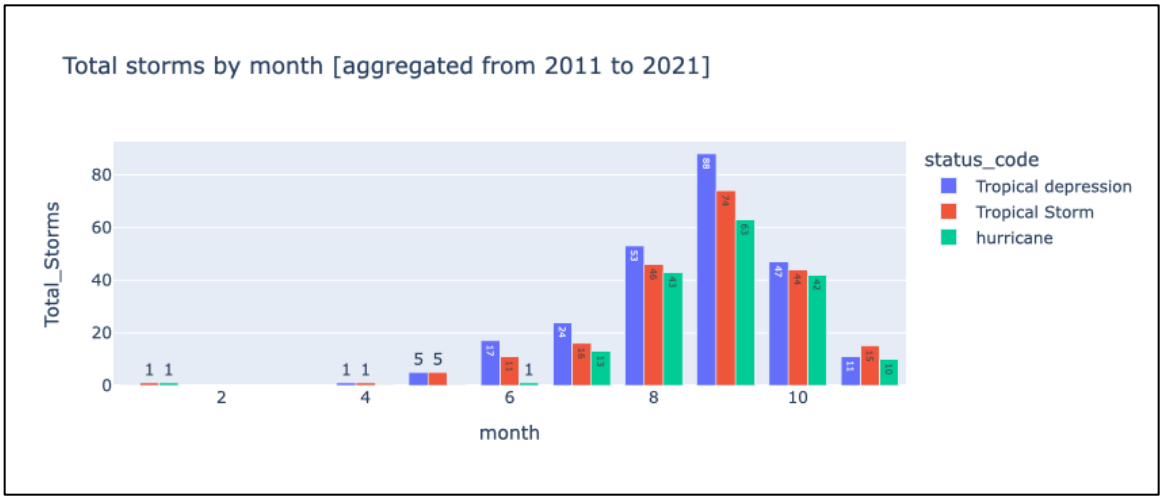


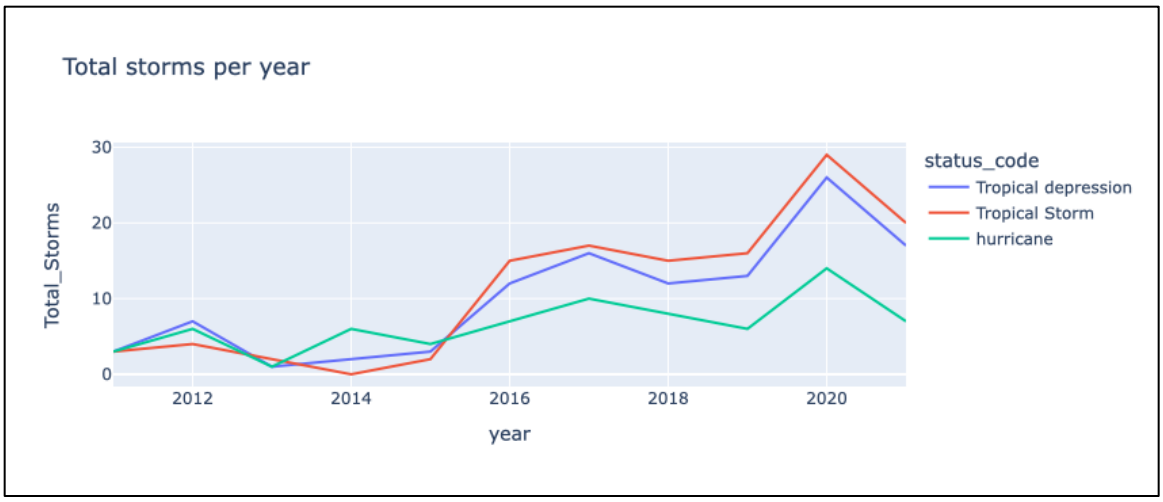
The cumulative line plot is perfect to identify trends, and the graph above show us an exponential positive trend on the last 10 years.

During the study of the graphs, we determine that using the last 10 years of data can provide to the model and predictions a better understanding of the recent trends storm behavior which in return can potentially result in better predictions of the upcoming years.

The following plots were produced with a subset of the dataset from 2011 to 2021:

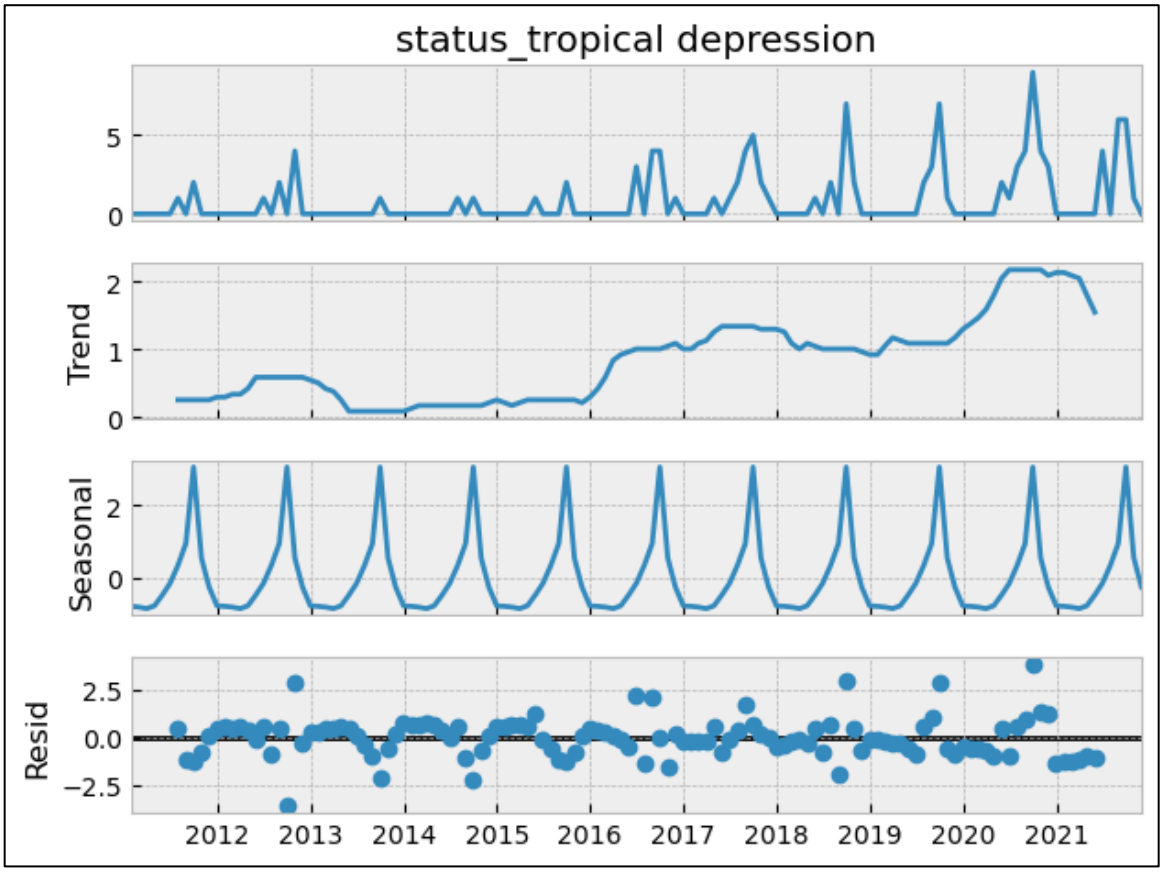






As we can see on the above graphs, the subset of the data from 2011 to 2021 maintains the general behavior of the complete dataset, with a positive trend (which increases exponentially on the recent years), and a seasonal peak in September.

The below graph shows a summary example of the timeseries representation using "Seasonal Decompose" function:



Stationarity check:

After exploring the data with different visualizations, we need to check the timeseries for stationarity using the "Adfuller" test, see below the results:

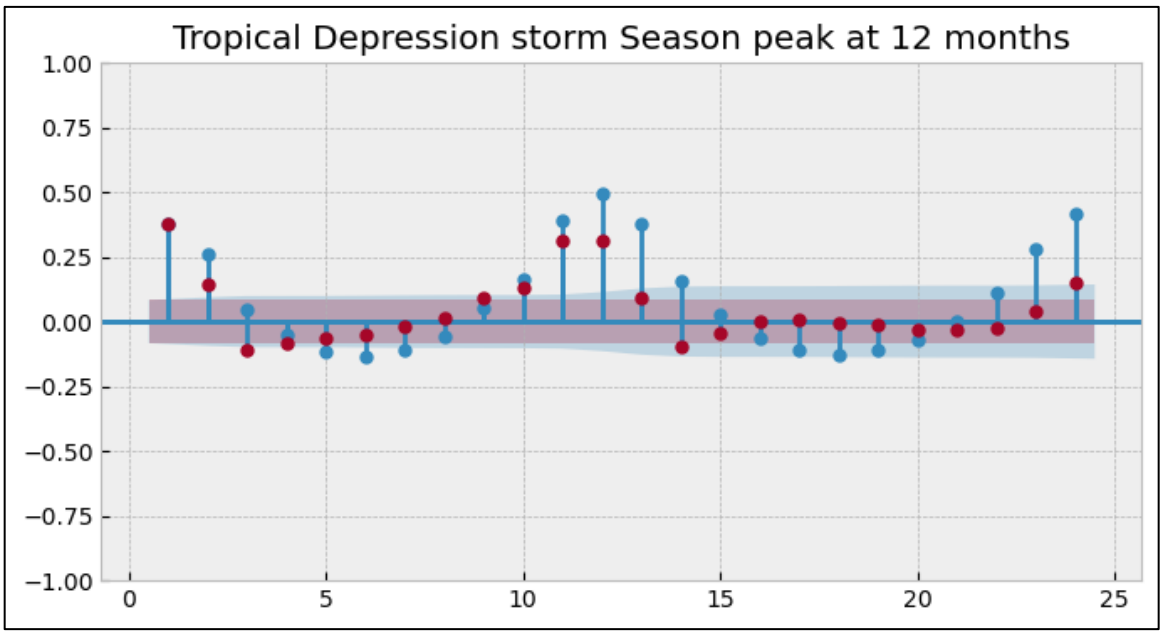
- Tropical depression time series:
    - ADF statistic: -2.63
    - P-value: 0.086
- Tropical storm time series:
    - ADF statistic: -1.71
    - P-value: 0.42
- Hurricane time series:
    - ADF statistic: -4.33
    - P-value: 0.00038

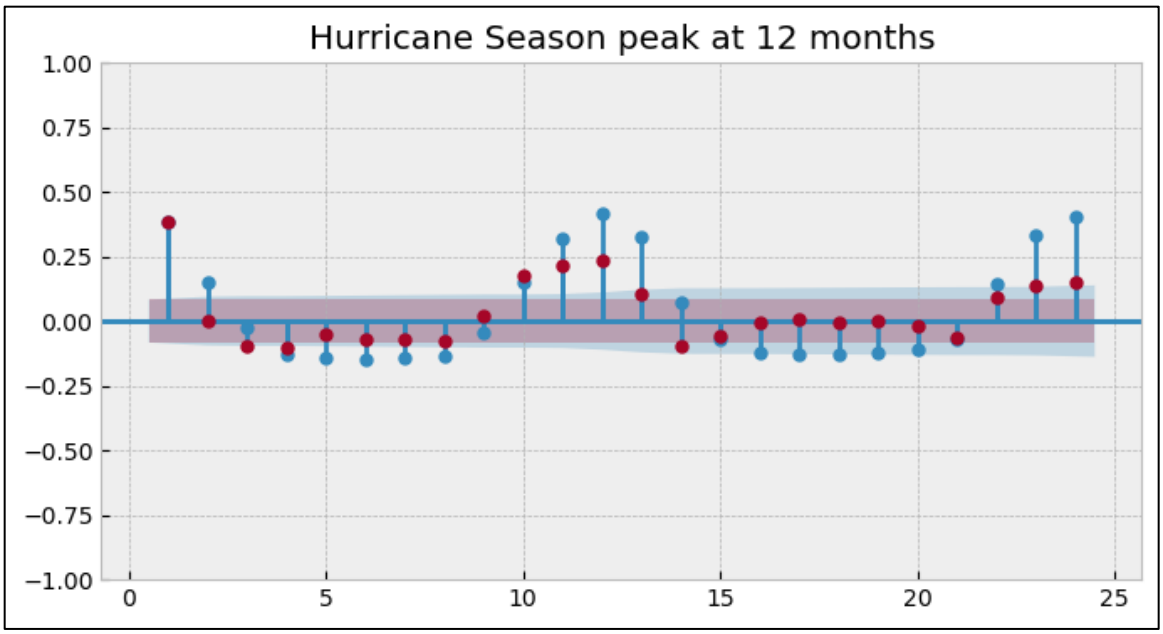
The above results shows that the Depression and Tropical storms timelines are not stationary (P value > 0.05 on Adfuller test), and that the hurricane timeline is stationary (P value < 0.05 on Adfuller test).

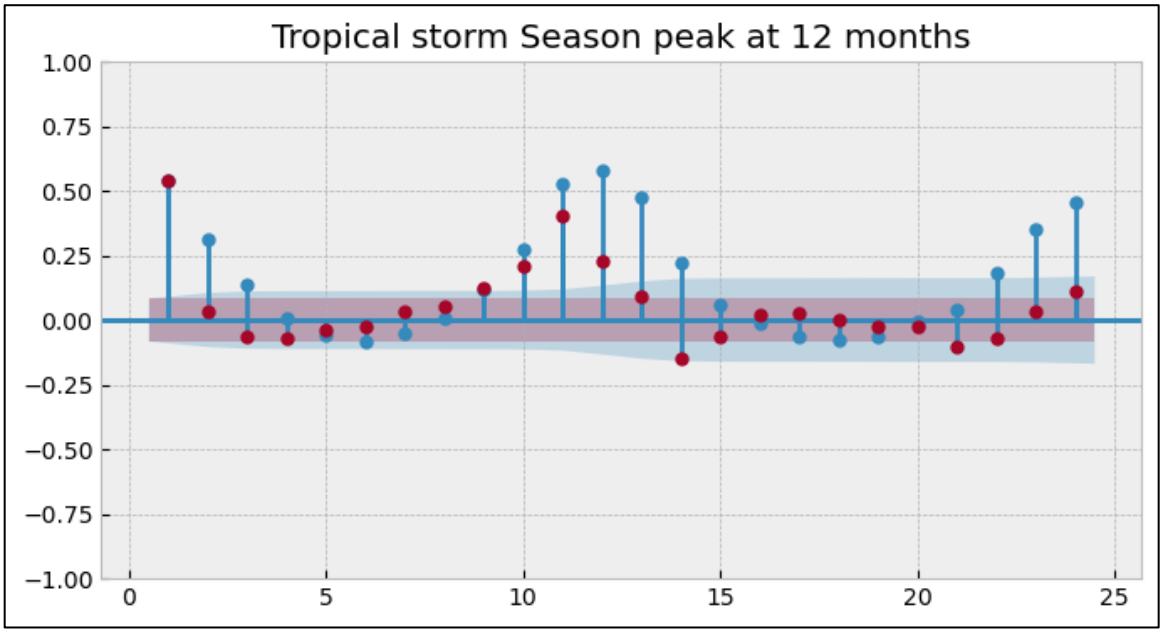
To achieve stationarity of the two non-stationary time series we are applying the difference to the respective series.

Uncovering seasonality using Autocorrelation (ACF) and Partial Autocorrelation on monthly data

To determine the appropriate lags for our time series model, we analyze the autocorrelation (ACF) and partial autocorrelation (PACF) plots.







Insights:

- The ACF plot shows significant correlations at lags 1 and 12, suggesting potential AR terms.
- The PACF plot shows a sharp cutoff after lag 1, indicating potential MA terms.

The exploratory data analysis revealed significant seasonal patterns and trends in storm occurrences. The data was made stationary through differencing techniques, facilitating accurate model training. The ACF and PACF plots suggest potential parameters for ARIMA modeling. The next steps will involve building and validating time series forecasting models to predict future storms occurrences.

4. MODEL SELECTION

In this section, we describe the process of selecting an appropriate time series forecasting model for our storms dataset. We have previously explored and preprocessed the data, making it ready for modeling. The goal is to identify the best model that accurately captures the underlying patterns and provides reliable forecasts.

Model Candidates

We consider several common time series models for our analysis:

- ARIMA (AutoRegressive Integrated Moving Average)
- SARIMA (Seasonal ARIMA)

To evaluate the performance of each model, the following library was used to iterate over different ARIMA and SARIMA components:

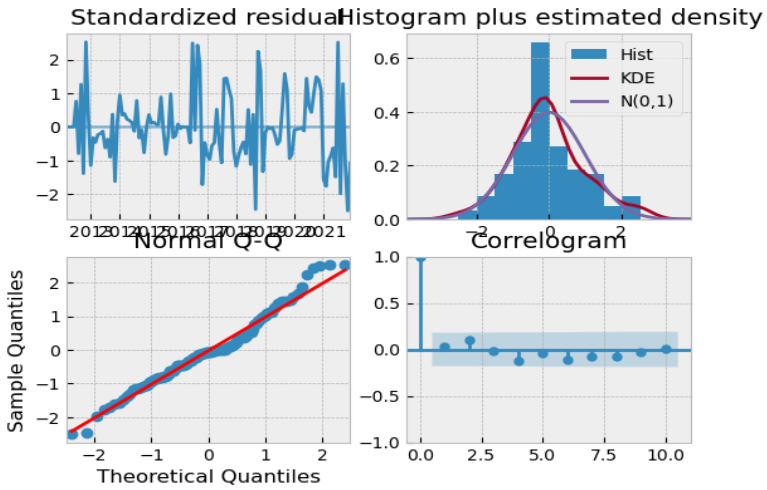
- Pm.Auto_arima (former named, Pyramid)

"pmdarima" is essentially a Python & Cython wrapper of several different statistical and machine learning libraries (statsmodels and scikit-learn), and operates by generalizing all ARIMA models into a single class (unlike statsmodels).

The auto_arima function itself operates a bit like a grid search, in that it tries various sets of p and q (also P and Q for seasonal models) parameters, selecting the model that minimizes the AIC (model criteria used for this Storm dataset). To select the differencing terms, auto_arima uses a test of stationarity (such as an augmented Dickey-Fuller test) and seasonality (such as the Canova-Hansen test) for seasonal models.

For this project, the focus was on finding the best tuning parameter that results on the best AIC scores.

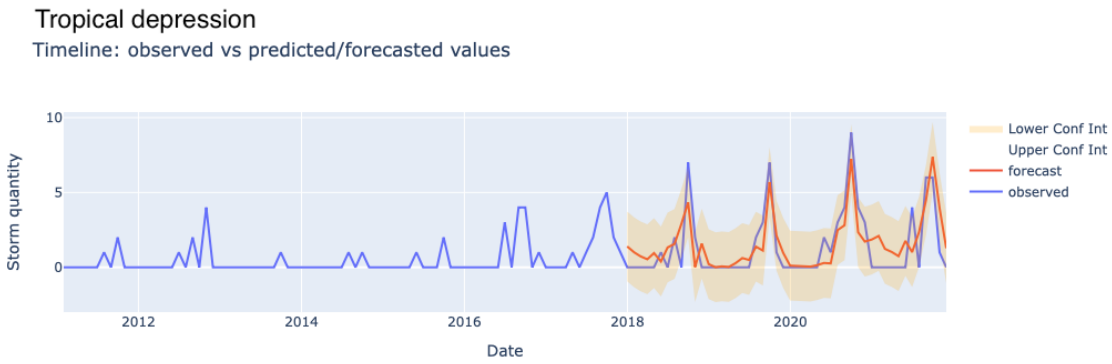
Data used	AIC	BIC	Best SARIMA parameters
All data (1975 to 2021)	1365.31	1395.42	(3,1,2)(0,1,1)[12]
Subset of data (2011 to 2021)	389.41	403.26	(1,1,2)(0,1,1)[12]

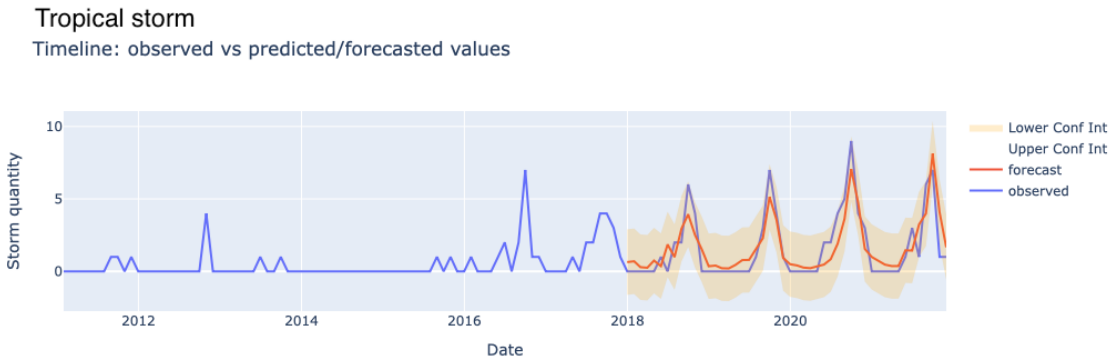


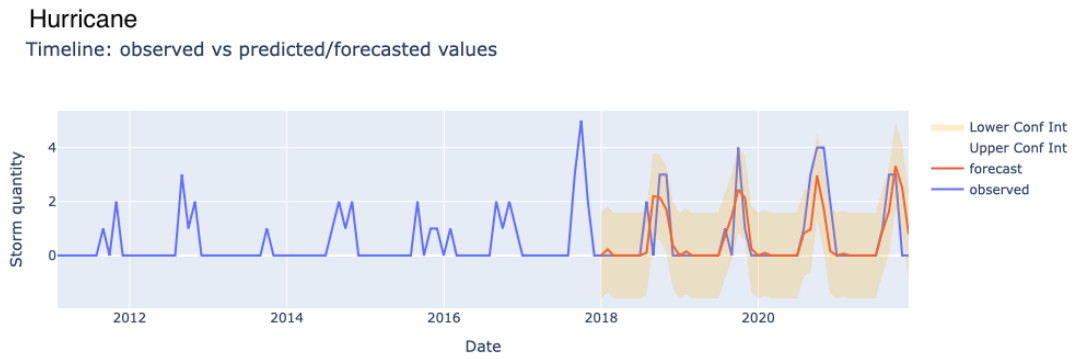
Based on the evaluation criteria, the SARIMA model with best parameters (1,1,2)(0,1,1)[12] and subset of data for the last 10 years (from 2011 to 2021) is selected as the best-performing model for forecasting our storms data. It provides the most accurate predictions with the error metrics. In the next steps, we will use the model to predict the trend, median and confidence interval of storms for the upcoming 2 years.

5. DOCUMENTATION – RESULT AND RECOMENDATIONS

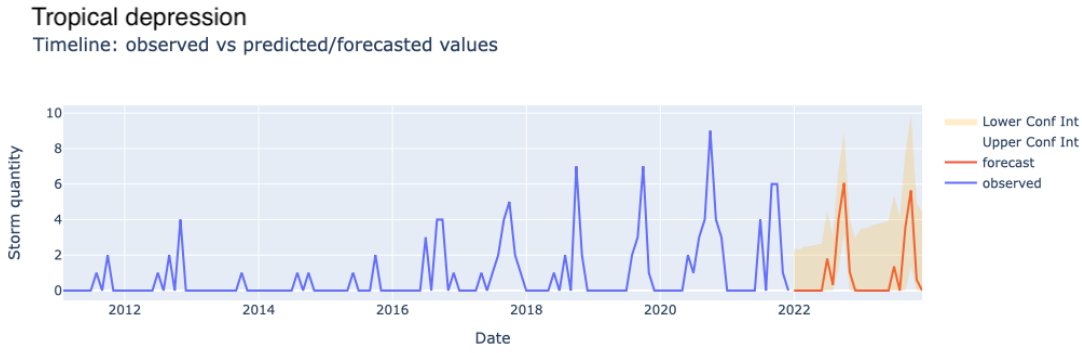
The objective of this project was to develop a time series forecasting model for predicting storms for each status on the upcoming two years. After performing data cleaning, transformation and model fitting and tuning, we can proceed with performing visual comparison of the true storm data against the predicted values from the model over the test period (within the data),

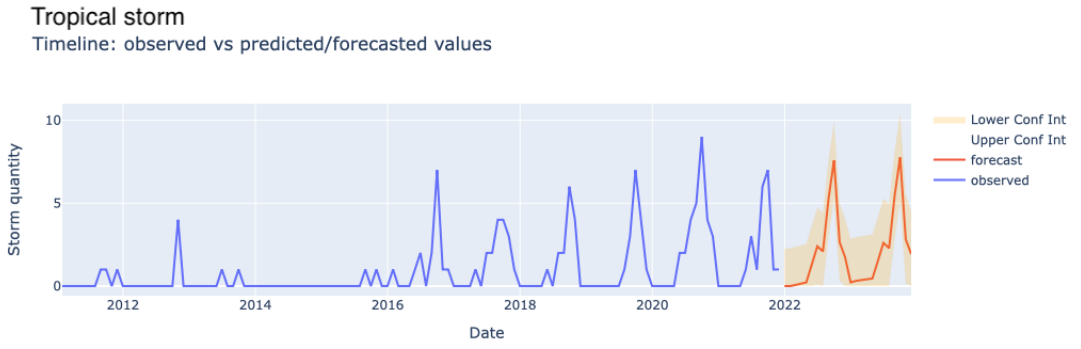


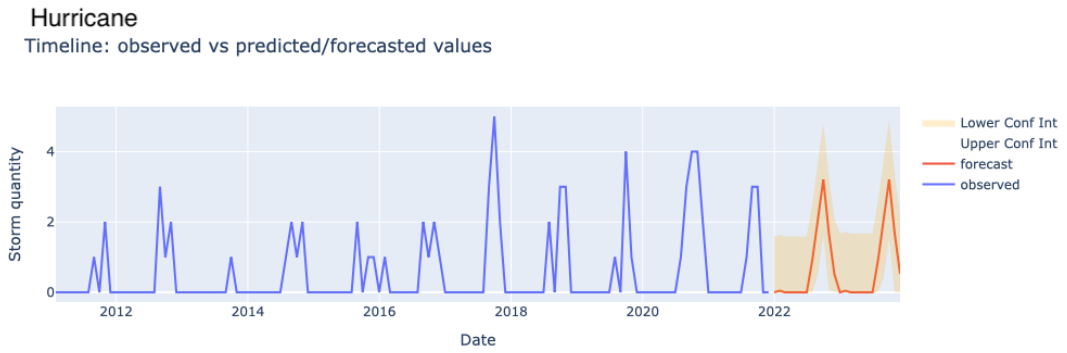




The visual inspection of the predicted values against the real data demonstrate that the fitted model generalized very well to our test data, also we can see that the confidence interval covered the most extreme values of the storm data at each peak season period, therefore we can feel confident to proceed with forecasting the storm quantity of the next 2 years, see below:







Based on the results of our analysis, we provide the following recommendations:

- Adopt the SARIMA model for forecasting with parameters (1,1,2)(0,1,1)[12]:

The SARIMA model with the specify parameters demonstrated good performance in terms of accuracy and error metrics. It is recommended to use the model for future storm forecasting. Fine-tuning the model by adding an exogenous variable could further improve its performance.

- Implement a Comprehensive Model Monitoring System:

Regularly monitor the performance of the deployed forecasting model. Implement a feedback loop to update the model with new data and retrain it periodically to maintain accuracy.

- User Training and Documentation:

Provide comprehensive documentation and training for the end-users to understand the model outputs and make informed decisions based on the forecasts.

This analysis provides a good foundation for storm forecasting, helping the end-user to make data-driven decisions and prepare for future storms impact to our society and economic partners. Accurate storm prediction is crucial for timely preparedness and response, which can significantly reduce the impact on human life, infrastructure, and economic activities.

Implementing these model and recommendations will enhance preparedness efforts, mitigate risks, and protect lives and property from the adverse effects of storms. This comprehensive approach to storm prediction marks a significant step forward in leveraging data science for natural disaster management.