Final Milestone

SUBJECT 4 - Big Data and Data Mining techniques applied to cyclones characterisation

SUMMARY :

Introduction

The mathematical modelisation of cyclone requires an understanding of the physical mechanisms of the cyclones such as viscosity, compressibility, and thermal aspects. The purpose of this project is to collect valuable data from different organization and exploit it to identify physical mechanisms according to the power of the cyclone considered.

To achieve that goal, we proceeded in four steps. We first did a bibliographic research to understand how cyclones work and determine the type of data needed to categorize cyclones. In the second step we contacted and collected data from different organizations. In the third step, we processed the data and extracted valuable data from it. And finally, we analysed the data and studied the characteristics of cyclones according to their categories.

I - Bibliographic research

A - Introduction to cyclones

A tropical cyclone is a system that forms over tropical waters in which atmospheric air is rotated around a low pressure center called ‘eye’. This cloud system is accompanied by intense rainfall and very strong winds that can reach up to hundreds of kilometers per hour.

Tropical cyclones are formed when the temperature above the surface of the seas exceed 26°C. This allows warm air to develop above the surface. This warm air become expand then become lighter and rises. Cool air comes from all directions to fill the void left by the rising warm air. This air also warms and rises. The rising air contain moister, evaporated from the ocean surface, that condense by the atmosphere and form huge cloud. The condensation release the latent heat energy stored in the water vapor, providing the cyclone more power. Combined within the Coriolis effect, the air spiral around the center of the cyclone which create a low pressure eye in the storm.

Cyclones can last up to several weeks. They dissipate when they reach land or cold water because the supply of moisture or and heat is cut off.

Tropical disturbances are classified according to the maximum sustained wind speed over 10 minutes. We talk about tropical depression if the speed is below 62km / h, tropical storm if the speed is between 63km / h and 117km / h, and finally hurricane when the speed exceeds 118km / h.

The data that has been collected consists essentially of hurricanes. Thus we naturally focused on this type of cyclone. The Saffir-Simpson scale was developed to categorize hurricanes. They are classified in 5 categories, according to the strength of the maximum winds.

Échelle de Saffir-Simpson	
Classe	**Vents maximum**
1	**118** et **153** km/h
2	**154** et **177** km/h
3	**178** et **209** km/h
4	**210** et **249** km/h
5	supérieurs à **249** km/h (catégorie des super-cyclones)

Tropical cyclones are the most devastating weather disturbances, causing hundreds (sometimes thousands) of casualties and damage every year in many countries.

Many studies are conducted to understand the functioning of this phenomenon in order to predict their formation and behavior. In recent years, there have been noticeable advances in this area that have greatly contributed to the improvement of tropical cyclone forecasts. Especially with the development of instrumental performances and numerical simulations.

B- Interesting quantities for cyclones

Thanks to the advice of our referent professor Joël Chaskalovic, we have concentrated on bibliography, in particular quantities such as Reynolds number, Mach number, shear rate, and vorticity.

1-/ Reynolds Number

The Reynolds number is a number used in fluid mechanics. It characterizes a flow, especially the nature of its regime (laminar, turbulent). This number represents the ratio between the inertial forces and the viscous forces. It appears naturally in the Navier-Stokes equations.

$Re = \frac{V*L}{\nu}$ with : V = characteristic velocity of the fluid [m/s]

L = characteristic dimension [m]

ν = kinematic viscosity of the fluid [m^2/s]

A flow is ::

- laminar when Re << 2000
- turbulent when Re >> 2000

2-/ Mach Number

The Mach number is a number that expresses the ratio between the local velocity of a fluid and the velocity of sound in that same fluid. This number measures the relationship between the forces related to the movement and the compressibility of the fluid.

$Ma = \frac{V}{a}$ with Ma = Mach number

V = velocity of the fluid

a = velocity of sound in the fliud

3-/ Shear rate

The shear rate, also called "velocity gradient", describes the variation of the flow velocity profile. The relationship between stress τ and shear rate γ is a linear relationship for viscous fluids, whose proportionality coefficient is the dynamic viscosity of the fluid η :

$$\tau = \eta\gamma$$

4-/ Vorticity

Vorticity is a measure of the rate of rotation of a parcel of fluid. Vorticity is a vectorial size that tells us about the rotation / circulation of a particle of a fluid. From a mathematical point of view, it is defined as the rotational velocity field of a fluid $\overline{V}$ at a certain point:

$$\overline{w} = rot\,\overline{V}$$

II - Data collection

Our strategy for finding consistent, high-volume data has been to learn about the countries most affected by these weather events. The 3 countries / continents most affected by tornadoes are: the United States, Canada and Europe.

During our research, we quickly realized that open access data on the internet was unusable and did not have the desired variables. The only way was therefore to contact the agencies involved in the collection of tornado data.

It was in the United States that we had the most success. In total, we sent about 50 emails targeting several organizations for the majority Americans.

It was by contacting NASA that we came to the conclusion of studying hurricanes rather than tornadoes. Two NASA people told us that studying tornadoes would be difficult. Indeed, tornadoes are quite ephemeral phenomena (a few hours) and small dimensions (from 50 meters to 1 kilometer). In addition, the measuring devices have too little accuracy in dealing with tornadoes and have a sampling frequency that is too small (in the order of 6 hours at best), we had to find a solution.

Hurricanes are phenomena that can last for weeks and their dimensions are of the order of one hundred kilometers. Our referent teacher, also doing research on his own, has come to the same conclusion: studying cyclones.

	Tornado	Cyclone
Radius	0.05-1 km	1000 km
Lifetime	several minutes to some hours	**several days to some weeks**
Sampling frequencies	every 6 h	
Possible study	difficult	easy

On the few responses obtained by mail, we still had access to several databases. This is the NOAA (National Oceanic and Atmospheric Administration) database on cyclones that has interested us the most.

III - Data processing

A-Crossing the data

Once focused on the NOAA cyclone database, we had available different types of files: BestTrack, TcDiag, TcVitals, TcTrack. As explained in the 2nd Milestone, we were interested in BestTrack and TcDiag.

In summary, file types other than TcDiag and BestTrack included only redundant variables found in BestTrack, TcDiag.

First, we had to transform the text files into .csv (excel format). We had to automate this conversion because there are more than 2.2 million files. It was therefore necessary to automatically recognize the variables knowing that the rows and columns of each file changed each time. It was with TcDiag files that we took the most time. Finally, we did it after two weeks of work using the Python programming language. We chose this language and not C ++ because automation requires much less lines of code and has more efficient tools. The management of text and tables in Python is for example much simpler and intuitive. Another reason is that we were more comfortable with Python than other languages.

It was when crossing the two types of files that we encountered a difficulty. We did not know how to cross files by losing a minimum of data. Indeed, BestTrack files only record every 24 hours while TcDiag records the most important cyclones every 6 hours.

After thinking, using the date and the name of the cyclone would make cross files. It was therefore necessary to name each cyclone by its location and its annual cyclone number in the zone for both types of files. It was also necessary to determine a number to know the time of the recording of the cyclone. So we chose 1980-01-01 at 00:00 and calculate the number of hours since that date to find a reference when crossing files.

Once the two columns (cyclone name and date) were determined for both types of files, we had to find a way to cross them by losing a minimum amount of information. We had a lot of data loss because few data had the same date and the same name.

For those who had a date and name match, we had a single line in BestTrack, but several lines in TcDiag. Indeed Tcdiag performed for the same cyclone on a specific date, several records. The solution has been to make as many lines of TcDiag corresponding for a line of BestTrack. so we multiply the lines of BestTrack to get the most information.

At present, we have not yet determined the interesting quantities (Reynolds number, Mach number, etc.) for our study. We made contact with Nicolas Viltard, cyclone specialist at LATMOS. We'll see him next week.

B- Data selection

Once the data was processed and cross-referenced, the HU cyclone category combined the 5 categories 1 2 3 4 and 5. We therefore divided this category from the Saffir-Simpson scale into 5 categories (depending on the speed of maximum winds).

We then encountered a problem with rows with unknown values. If we had decided to just keep all the variables and delete all lines with unknowns, we would have had very little data.

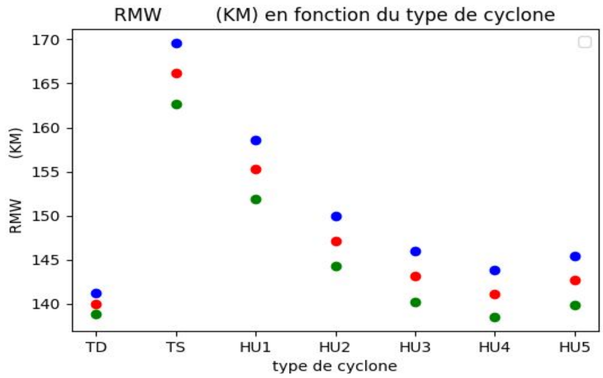
Our solution was therefore to eliminate some variables (not very relevant for our study) with a majority of unknown values (around 80% of unknowns). It is then and to the detriment of the quantity of data that we kept the lines containing 100% of known variables. This method allowed us to keep a maximum of lines for a maximum of relevant information.

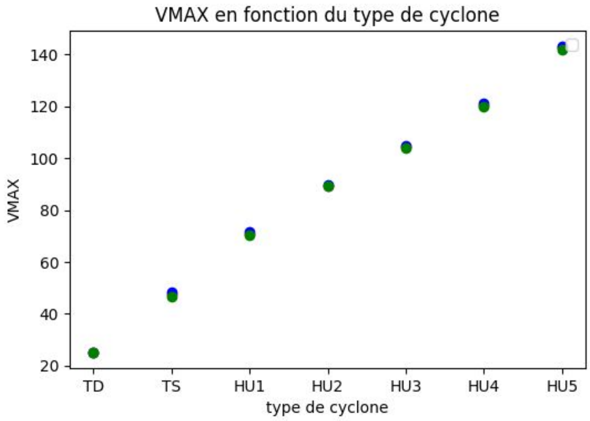
Our cyclone categories are (in increasing order of power): TD (Tropical Depression), TS (Tropical Storm), HU 1 (Hurricane class 1), HU 2 (Hurricane class 2), HU 3 (Hurricane class 3), HU 4 (Hurricane class 4), HU 5 (Hurricane class 5)

The relevant variables are: maximum speed of sales recorded over 10 minutes, cyclone radius, cyclone category, minimum sea surface pressure, shear rate, cyclone velocity, surface temperature of the sea, the average vorticity between 0 and 1000 kilometers, the precipitation, the pressure at the level of the maximum radius of the cyclone, pressures at different altitudes, temperatures at different altitudes

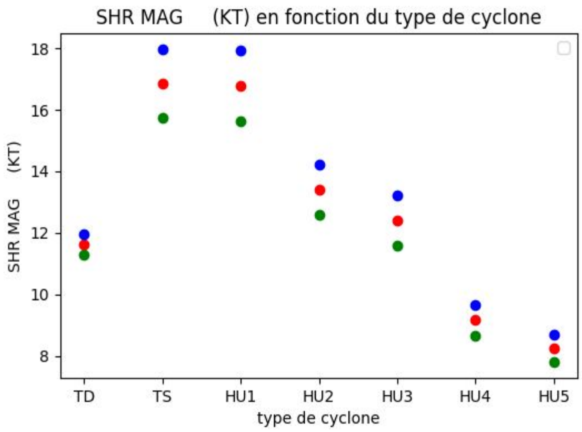
IV - Elementary statistical study of cyclones according to their power

- On our graphical representations we have:
- the red dots are the averages
- the green dots are the averages minus the standard error
- the blue dots are the averages plus the standard error

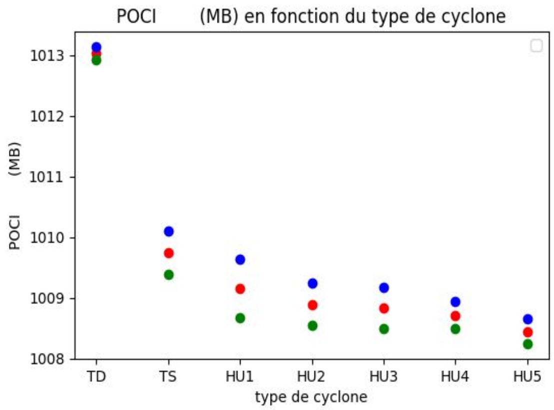


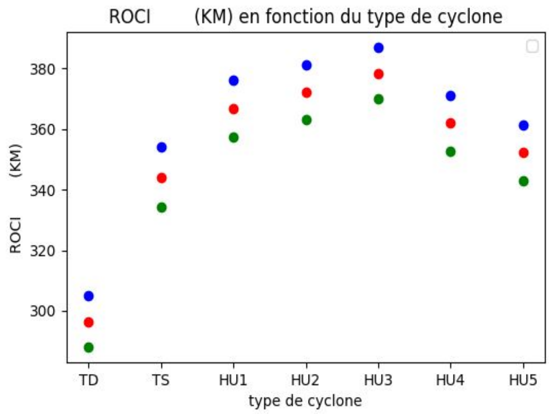


Here we have a representation of the radius of maximum wind speeds depending on the type of cyclone. The trend is downwards, that is, the higher the cyclone's power, the faster the maximum speeds are closer to the eye. According to the second figure, the maximum speed that winds can reach increases proportionally with the type of cyclone, which is quite consistent because the type of cyclone is defined with the max.



We were also interested in the shear rate of cyclone winds, we note that it seems to decrease drastically with the power of the cyclone. The wind speed profile seems to be homogenizing. It can be seen that the average shear rate between 0 and 500 km decreases with the power of the cyclone category. It is with the exception of TD that we can certainly explain this decrease according to the category of cyclones. The reason for this decay can be explained by the increase of small additional velocity values because of the spread over longer distances of the cyclones (the stronger the cyclone, the larger the radius)

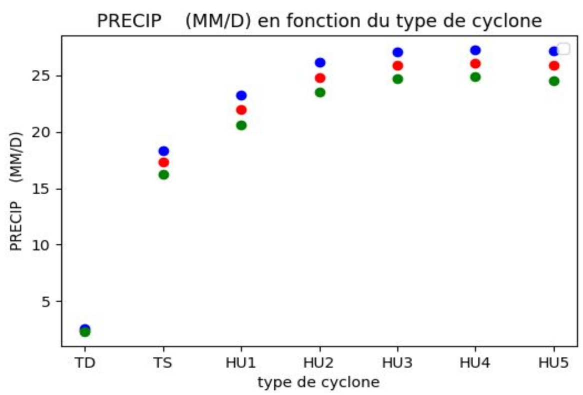


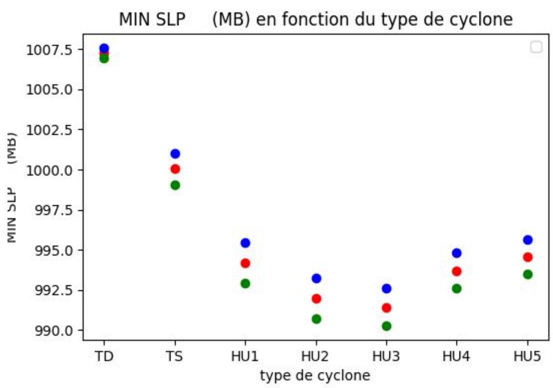


The last closed isobar is the last isobar circle of the cyclone. We can consider that this ray is the radius of the cyclone.

We observe if we focus on the graph on the left that the pressure of the last closed isobar (POCI pressure of the Outermost Closed Isobar) sees its value decrease very slightly depending on the type of cyclone. The graph on the right explains that the radius of the cyclone increases according to its power. On average, category 3 is considered the largest.

This may be due to the fact that the higher the speeds, the more isobars are broken and therefore the ROCI of category 3 is larger than that of categories 4 and 5. The approximation that the ROCI is the radius of the cyclone is not exact but its order of magnitude is compatible with this approximation. These two graphs represent quite intuitive and understandable phenomena.





The graph on the left shows the precipitation in mm / day depending on the type of cyclone. We find that it increases. When the graph on the right is a graphical representation of the minimum pressure in milliBar depending on the type of cyclone. These two graphs represent quite intuitive and understandable phenomena. In addition, the minimum pressure decreases with the pressure of the last closed isobar.

Conclusion

During this project, the search for the data took us a lot of time as well as the data processing. However, we have not yet completed the study of cyclones. Indeed, we have only done a basic statistical study and we have not yet had the time to calculate the variables relevant for cyclone studies according to their category. Next week we will see our referent professor to continue the project as well as Nicolas Viltard, cyclones specialist working at LATMOS, to finish the project.