**Mémoire présenté le : 4 mai 2021**

**pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires**

Par : Oluwaseye AGUNBIADE

Titre : Error analysis and accuracy of the Tropical Storm Risk (TSR) wind model

Confidentialité : ☐ NON ☒ (Durée : ☐ 1 an ☒ 2 ans)
*Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus*

*Membres présents du jury de l'Institut des Actuaires* — *Signature*

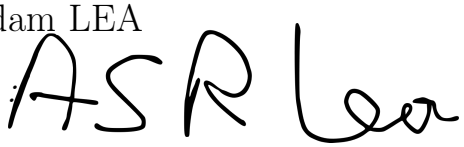
B. BALTESAR

R. NOBIS

*Entreprise :* TSR Business
*Nom :*

*Signature :*

*Directeur de mémoire en entreprise :*
*Nom :* Adam LEA
*Signature :*



*Membres présents du jury de l'ISFA*

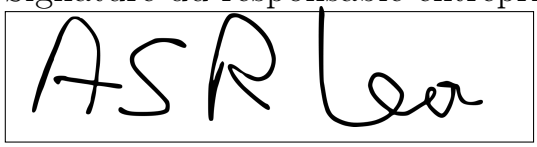
P. RIBEREAU

*Invité :*
*Nom :*

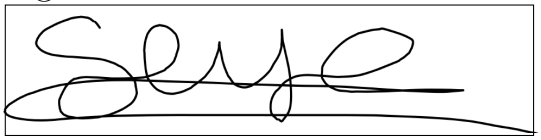
*Signature :*

***Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels*** *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise



Signature du candidat



# Acknowledgment

To my best friend I love you and thank you for cheering me and believing that I could write this paper even when I stopped believing. My family, my heroes, the **AGUNBIADES**! The greatest fans one could ever dream of having who never give up on me even during my tantrums, this joy is mine as well as yours, it was really our journey. To all my tutors: **Adam LEA, Christian ROBERT, Luis MAGNET** thank you for all the contribution and feedback. For all my unforgettable memories at CNP Assurances, I am grateful **Quentin PHUNG** and **Nicolas DELUMEAU** you were such awesome colleagues one could never forget. There are so many more that helped me throughout this process it is an endless list. So, to those not mentioned I am sincerely grateful to have all of you in my life thank you for your support, your love, for being there and your smiles.

*Each member of my family deserves a shout out: My dad,* ***Mr Oladipo****, who makes me feel like the most amazing daughter in the world. My mum* ***Mrs Adejoke****, the unshakable rock always sure somehow that everything is going to be just fine. My big sister and her family,* ***Mrs Temitope, Mr Yorick*** *and my nephew* ***Peniel****, the life of the party. My second eldest sister,* ***Mrs Oluwaseun****, we really did struggle together on this one, thanks for being my proof reader even during the night, I will forever be your agent "littéraire deux". My little brother* ***Mr Oluwafisayomi****, the adventurer, always has a joke ready for any given situation. And "la petite"* ***Mrs Ifeoluwa****, so precious in your ways and faithful in your care for me. You all really are my pride and I am so proud to be a member of this family.*

# Résumé

Au vu de l'incapacité des méthodes actuarielles à modéliser les événements extrêmes, la modélisation de Catastrophe Naturelle (Cat-Nat) est née. Le General Insurance Research Organising Committee (GIRO) (Fulcher 2006) a accentué l'importance de comprendre les principes sous-jacents des modèles Cat-Nat afin de mieux comprendre leurs limites et correctement utiliser leurs résultats (e.g. sommes assurées). L'entreprise Tropical Storm Risk (TSR) commercialise un modèle pour simuler la vitesse du vent le modèle TSR. Il permet aux risque manageurs, les différentes parties prenantes et d'autres, d'évaluer, gérer, d'allouer, mitiger les risques dûs au vent. Le modèle donne une estimation de la vitesse du vent sur une période de 3 secondes (rafale) et sur une période d'une durée d'une minute. Il est important d'évaluer l'exactitude du modèle TSR pour savoir dans quelle mesure (degré) les estimations obtenues avec le modèle sont fiables. Une analyse des erreurs a été effectuée pour déterminer la source des erreurs ayant des valeurs élevée. L'entreprise TSR a déjà calculé la performance du modèle auparavant, pour les tempêtes qui sont arrivées sur le sol américain entre 2004 et 2008. Afin de comparer la performance du modèle avec celui obtenu par l'entreprise TSR la même mesure de performance est utilisé: Absolute Percentage Error (APE) calculé avec la moyenne de l'erreur $\pm$ l'écart type en pourcentage. Dans cette étude la performance du modèle obtenu pour l'estimation des rafales de vent est de 19.6% $\pm$ 14.6% et celui de la vitesse du vent moyennant sur une période de 1 minute est de 17.8% $\pm$ 16%. Il est difficile de conclure sur l'évolution de la performance du modèle TSR pour les rafales de vent. Il y a un compromis entre le biais et la variance. En effet, la variance a diminué, mais la moyenne a augmenté par rapport à l'étude précédente. Toutefois, en ce qui concerne la performance du modèle pour les vents moyennant une période de 1 minute, la précision du modèle TSR s'est nettement

améliorée au cours du temps : 17.6% ±14.4% (dans cette étude) contre 22.3% ± 20.7% (valeur dans l'étude précédente). L'étude qui suit est conforme à la recommandation du GIRO, en évaluant la performance du modèle TSR, une explication détaillée du modèle a été effectué.

# Abstract

In light of the inability of actuarial techniques to model extreme events catastrophe models were born. The General Insurance Research Organising Committee (GIRO) (Fulcher 2006) stressed the importance of understanding the underlying principles of catastrophe models to better comprehend their limitations and how to correctly use their outputs (e.g. insured loss). The Tropical Storm Risk (TSR) Business commercialises a wind model TSR model that enables risk managers, stakeholders and others to evaluate, manage, spread, mitigate the risk due to wind hazard. The model provides estimate for the 3-sec peak gust wind speed and the 1-min average sustained wind speed for storms. Thus, it is important to evaluate the accuracy of the TSR model to know to what extent (degree) the estimates of the model are reliable. An error analysis was performed to identify the source of large errors. The TSR Business has already calculated the accuracy of the model before, for storms that made landfall in the US between 2004-2008. To be able to compare the accuracy of the model to that previously obtained by the TSR Business the accuracy measure used is the Absolute Percentage Error (APE) in terms of mean $\pm$ standard deviation. In this study, the accuracy for gust was found to be 19.6% $\pm$ 14.6% and the one found in the previous study was 17.8% $\pm$ 16%. Here, it is difficult to conclude if the accuracy of the TSR model for gust got better or not as there is a trade off between bias and variance. Indeed, the variance decreased but the mean increased compared to the previous study. However, in terms of the accuracy for the 1-min sustained wind there is a clear improvement of the accuracy of the TSR model over time: 17.6% $\pm$14.4% (in this study) vs 22.3% $\pm$ 20.7 (value from the previous study). The following study is in line with the recommendation of the GIRO, by evaluating the accuracy of the TSR model, a detail explanation of the model was done.

# Contents

# Acronyms

**APE** Absolute Percentage Error. 4, 6, 16, 76–79, 81, 88

**Cat-Nat** Catastrophe Naturelle. 4

**GIRO** General Insurance Research Organising Committee. 4–6, 20, 88

**IFoA** Institute and Faculty of Actuaries. 20

**IQR** Interquartile Range. 57, 101–103

**LOF** Local Outlier Factor. 57, 104–106

**MAD** Mean Absolute Deviation. 57, 101

**MAE** Mean Absolute Error. 17, 75, 82, 83, 89

**NHC** National Hurricane Centre. 12, 17, 22, 24, 26, 34, 35, 37, 41, 43, 45, 46, 61, 63, 65, 74, 81–83, 88, 89

**NOAA** National Oceanic and Atmospheric Association. 24

**RMSE** Root Mean Square Error. 75

**SD** standard deviation. 50–52, 57, 58, 76

**SSHWS** Saffir-Simpson Hurricane Wind Scale. 12, 14, 16, 23, 36–38, 44, 45, 77, 97

**TS** Tropical Storm. 34–39, 45, 61

**TSR** Tropical Storm Risk. 1, 4, 6, 13, 14, 17, 21, 22, 26, 28, 32, 33, 41, 43, 44, 46–48, 61, 63, 65, 66, 69, 72–74, 76, 78–85, 88, 89

# Glossary

**best track** A subjectively-smoothed representation of a tropical cyclone's location and intensity over its lifetime (Source:NHC). 26

**hurricane** A storm with a 1-min maximum sustained wind greater than or equal to 64kt. 21, 23

**outlier** An observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. 47, 57

**random errors** Uncertainty that comes from the modelling process. 47

**storm** A violent disturbance of the atmosphere with strong winds and usually rain, thunder, lightning, or snow. 21

**systematic errors** Errors with a non-random character that distort the result of a measurement. 47

**track** Path, trajectory of a storm. 27

**whisker** A range made up of values that are than the first quartile but larger than first quartile -1.5*IQR or more than the 3rd quartile but less than 3rd quartile + 1.5*IQR. 102

**wind profile** A wind profile is a set of all the different wind speed that form the tracks of a hurricane. 27, 63

# List of Figures

# List of Tables

# Introduction

## Motivation

In 1992, Hurricane Andrew rendered more than 10 insurers insolvent and cost around USD 16 billion insured losses (AIR-WorldWide 2017). This was a wake up call for the insurance industry as actuarial techniques are inefficient to model extreme events due to their rare occurrence there is a scarcity of historical loss data. Hence, the need for tools to evaluate, anticipate and estimate the potential cost of catastrophic events before they occur. This is especially relevant in 2020 as the world is shaken by the global pandemic COVID 19. The estimated insured loss due to the corona virus in 2020 is over USD 107 billion (Cohn 2020).

However, among everything else 2020 also stands out for the number of hurricanes that made landfall in the US during the hurricane season ($1^{st}$ of June to the $30^{th}$ of November). 2020 is a record filled year in terms of storm hazard: highest named storms with 30 named storms (previously 28 storms observed in 2005), highest US landfall on record 12 storms with 5 making landfall in Louisiana (Cappucci 2020). More than a 100 lives were lost and the insured loss due to the 2020 hurricane season is estimated to be USD 41 billion for the Atlantic basin with more than USD 37 billion in the US due to storms that made landfall (Bowen 2020; Cappucci 2020). 2020 hurricane season has also been the $5^{th}$ consecutive year with above average normal storm activity (Cappucci 2020). With this in mind the need of models to estimate the financial losses due to storms is even more relevant. Indeed, in order for insurers to remain solvent they have to have an overview of the risks they cover by estimating their potential losses in the event the hazard occurs. This is also of interest for governments and non-profit organisations to

alert population, put in place mitigation strategies and build up resilience.

## Objective

A catastrophe model can be divided into 3 parts: the hazard component, the engineering component and the financial component an illustration is given in figure 1.

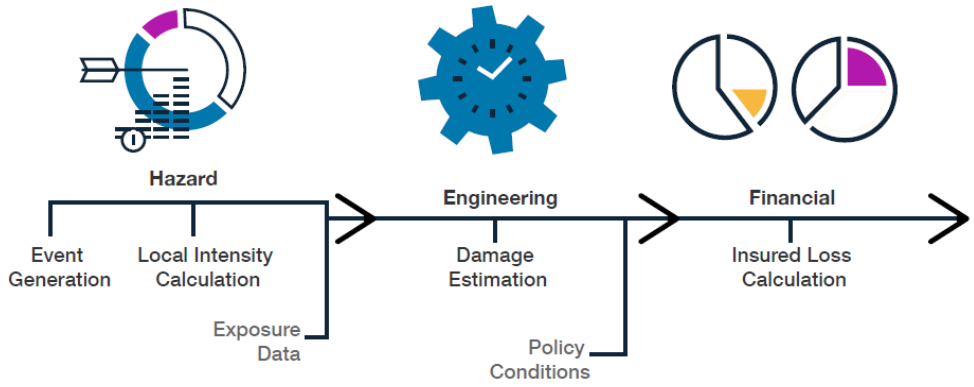



Figure 1: Schema: How Catastrophe Models Are Constructed, Source: AIR-WorldWide (2017)

*The following paragraphs are a simplified overview of each component.*

The hazard component focuses on building the framework to model the risk at hand by generating several scenarios of events with varying intensities while taking into account the exposure data of the underlying portfolio of interest.

The engineering component links the intensities obtained in the hazard component to corresponding level of damage. This is done using damage functions. Damage functions produce damage ratios that are values between 0 and 1. Multiplying the damage ratio with the insured value gives the loss due to the intensity. An illustration is given in figure 2.

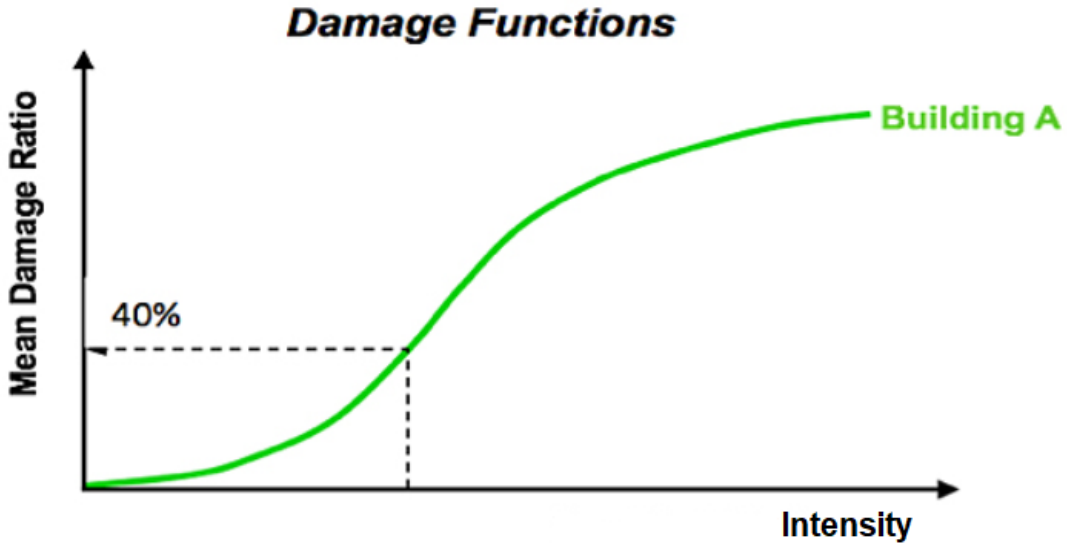



Figure 2: An illustration of damage functions with exposure in regards to building A. As the intensity increases so does the (mean) damage ratio. Source: Latchman (2009), figure modified

The financial component uses the loss obtained in the engineering component to calculate the insured loss in regards to the policy terms (e.g. apply deductible).

The General Insurance Research Organising Committee (GIRO) "develops and builds on the strong research output carried out in the general insurance practice area [...] and supports the Institute and Faculty of Actuaries (IFoA)'s Council in the delivery of it's corporate strategy"(Source: site IFoA). In 2006 the GIRO produced the document *Catastrophe Modelling Working Party* (Fulcher 2006) that highlighted the use of catastrophe model by actuaries. In that paper the GIRO emphasised that catastrophe models shouldn't be a "black box" for actuaries as it is important to have an understanding of the different steps between the inputs and outputs of a catastrophe model. Thus, to be able to have more insight on the evaluation of the risk at hand as the insured loss obtained with the catastrophe model are just estimates and contain a certain degree of error. It is then the responsibility of the actuary, risk manager to know if the results obtained are in line with the companies risk appetite so as to be selective of the risk they undertake. For an actuary, understanding the engineering and financial components of a catastrophe model isn't a daunting task as they are based on familiar concepts such as probability

distributions, policy conditions e.t.c. The hazard component can be more challenging as it is based on more foreign notions.

In this study the focus will be on the hazard component and more specifically on wind hazard. Indeed, storms are "violent disturbance of the atmosphere with strong winds and usually rain, thunder, lightning, or snow "(Lexico 2020), hurricanes are storms with wind speed greater than or equal to 64kt. The Tropical Storm Risk (TSR) Business commercialises a wind model Tropical Storm Risk (TSR) model for risk management purpose which will correspond to the hazard component in a catastrophe model. In view of what was discussed in the previous section it is important for the TSR Business to verify that the TSR model gives reliable estimates of the damaged caused by storms. The Tropical Storm Risk (TSR) model provides approximations of the wind speed of storms (i.e intensity of storms); given the wind speed of a storm it is then possible to evaluate the damage the storm could cause.

The aim of this study is two folds: to determine the accuracy of the TSR wind model for storm intensity in land as that is where most of the damage occurs. By doing so it also provides a global understanding of the hazard component of a catastrophe model for wind hazard. The wind data used is for storms that made landfall in the US between 2015 and 2019. Accuracy in this sense, measures the degree to which the values calculated with the model differs from the corresponding observations. The gap between these two values is called the error term. Ultimately, the ideal situation will be to have the errors equal to 0 such that the accuracy of the model will be 100%. However, observations themselves contain a certain level of uncertainty. The same instrument can be used to measure the same experiment or phenomenon multiple times and yet different values can be obtained every time. Observations can be viewed as a random sample from a probability distribution where the true value of the measured process is actually the mean of the probability distribution (Berendsen 2011). Yet, most of the time the distribution of the observations is unknown. Thus, the importance of error analysis to give the level of uncertainty associated with any measurement and/or experiment (Hase 2010).

Before performing the error analysis of the intensity estimates from the TSR model, different storm model framework are discussed then a description of the TSR wind model is given in chapter 1. Next, the wind data is presented, homogenised, cleansed chapter 2 and the error analysis is done chapter 3. In the following chapter the accuracy of the TSR model is calculated then compared to the previous study of accuracy of the TSR model for storms that made landfall in the US between 2004 and 2008; then the accuracy of the TSR model is compared to the accuracy of the model used by the National Hurricane Centre (NHC) chapter 4. Finally, the last section of chapter 4 is on the importance of understanding the catastrophe model used to correctly evaluate the insured loss obtained.

# Chapter 1

# Storm model framework

## 1.1 Tropical storm catastrophe models

### 1.1.1 The importance of wind models

*This section is based on the paper "Wind profiles in Parametric Hurricane Models" by Desflots (2010).*

Hurricanes are storms with a 1-min maximum sustained wind greater than or equal to 64kt.

Tropical storm catastrophe models are generally centered around wind speed estimation. This can be justified as the severity of a storm is actually determined by the maximum 1-min sustained wind speed (i.e intensity) of the storm according to the Saffir-Simpson Hurricane Wind Scale (SSHWS). The higher the intensity of the storm the more damage the storm will cause in-land. A table presenting the different level of damage depending on the wind speed/ category of a storm according to the SSHWS is available in appendix A.

The goal of a wind model is to compute the spatial distribution of the tropical storm winds. Two methods can be used to do so: the parametric numerical approach and the dynamical numerical approach.

- **Dynamical models** are models that mimic reality as closely as possible by incorporating all the physical laws and principles that reflect the atmospheric and oceanic conditions of the storm and run on a fine grid resolution. Despite it being highly efficient only hurricanes advisory forecaster agencies such as the National Hurricane Centre (NHC) use them as it is computationally expensive (consuming) in terms of time and computer memory. Thus, catastrophe modellers prefer parametric models.

- **Parametric models** generate an approximation of the wind field of the storm by focusing on the estimation of the wind profile; the distribution of the wind speed from the core of the storm see figure 1.3. This method is a low-cost alternative to the dynamical model and renders satisfying results.

In the following paragraph parametric wind models are presented.

### 1.1.2 The 3 well-known wind parametric models

#### 1.1.2.1 NWS-23 (1979)

The NWS-23 wind profile was first presented in the report *Meteorological criteria for standard project hurricane and probable maximum hurricane windfields, Gulf and East coasts of the United States* published by NOAA (1979). Aircrafts reconnaissance flights in the North Atlantic collected information on storms such as wind data between 1957 and 1969. Using this data the National Oceanic and Atmospheric Association (NOAA) came up with a statistical representation of a storm's wind profile. The NWS-23 model is made up of two distinct parts (i.e equations). The first component accounts for the winds speed rate of decay from Rmax (radius, distance between the eye of the storm and the location where the maximum wind speed is observed) to the center of storm. This is done with an exponential rate of decay to portraiy the steep increase of wind speed from the eye of the storm to Rmax. On the other hand, the second component translates the rate of wind speed decay from Rmax to the outer bound of the storm. A logarithmic rate of decay is used as this translates the more moderate and slow decrease of wind speed

between Rmax and the outer bound of the storm.

#### 1.1.2.2 Holland (1980)

The Holland model differs from the NWS-23 model as it incorporates a shape parameter B. This parameter controls the size of the region around the highest wind Rmax. When B increases the region around Rmax becomes narrower and when B decreases the region around Rmax is broader, an illustration of the Holland wind profile for varying B shape parameters is shown in figure 1.1.

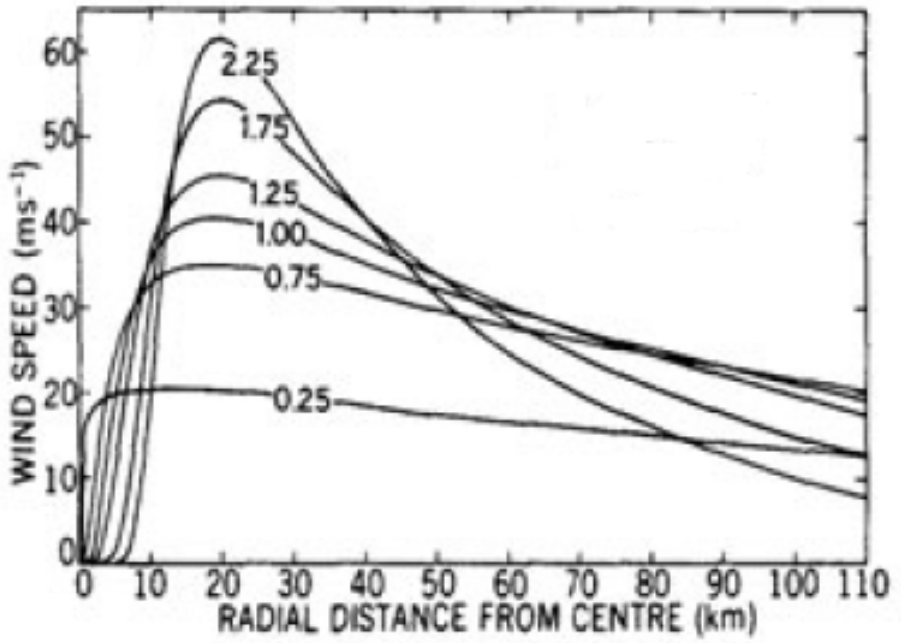



Figure 1.1: Holland's wind profile varies as the shape profile B is modified. High values of B produces a wind profile with a narrow peak and low values of B produces a wind profile with a broad peak. Source: Holland (1980)

This model was tested on 12 hurricanes from the Atlantic and Australian basins and it yielded acceptable results.

#### 1.1.2.3 Willoughby and co-authors (2004, 2006)

Though, the Willoughby (2006) wind profile also emerged from a statistical fit it defers from the two previous wind profiles as it makes use of three equations to compute the distribution of the wind speed of a storm. The three equations describe sections of the storm: the eye of the storm to Rmax, the transitional region, outer bound region (i.e

periphery area of the storm). A power law is used to describe the wind speed between the core of the storm and Rmax. Then a double exponential is chosen for the transitional region and the outer bound region, to translate a more rapid decay rate in the transitional region compared to a much slower decay rate in the outer bound region. This method was found to capture the storm's wind profile complexity.

The three wind profiles presented above all have the potential to capture the wind profile of storms. The choice of the appropriate model is done by finding the right fit for the wind data at hand.

In the following paragraph the wind model used in TSR model is given and the wind model is an hybrid between the NWS-23 wind profile and the Holland wind profile.

## 1.2 TSR wind model

The Tropical Storm Risk (TSR) model is a parametric wind model that is used to construct historical and forecast the wind map of a storm. It is based on 4 major components: the modified Rankine vortex model, the adjustment for translational velocity, the roughness correction factor and the gust factor. These elements are presented in the following paragraphs.

The input of the TSR model is given by the National Hurricane Centre (NHC) that provides forecast advisory for storms: best track for real time storm tracking. However, this study is based on past storms so the input from the TSR model comes from the best track data from the HURDAT 2 database (NOAA 2020). It is made up of different elements such as the storm position (longitude and latitude), radius from the hurricane eye (center of the hurricane) for different wind speed (34kt, 50kt and 64kt) in the quadrant (NW, SE, NW and SW) (Lea 2020) . These storm data are given at a 6 hour interval: 0h, 0h+6h, 6h+6h, 12h+6h according to the Universal Time Coordinate; at times storm data in between these time periods are provided if the storm makes landfall in between the 6 hour interval; in that case these points are used as well as input of the model if available.

The wind profile used to obtain the output data is presented in the following subsection.

### 1.2.1 Modified Rankine vortex wind profile

A wind profile is a set of all the different wind speeds that form the track of a storm. To generate the wind profile the modified Rankine vortex model is used, its equation is given below (Macfee 2006):

$$V = \begin{cases} V_{max} \times \left(\frac{r}{R_{max}}\right)^B & \text{if } r < R_{max} \\ \\ V_{max} \times \left(\frac{R_{max}}{r}\right)^B & \text{if } r \geq R_{max} \end{cases} \tag{1.1}$$

, where V is the calculated wind speed, $V_{max}$ is the maximum sustained wind observed, $R_{max}$ corresponds to the radius of $V_{max}$ , r the radius: the distance from the hurricane eye and B the shape parameter. Phadke (2003), also describes B as the parameter that incorporates the radial direction in the wind speed distribution. Thus, translating the decay of the wind speed from $R_{max}$ to the extremity of the storm as the radius increase with the distance from the eye of the storm.

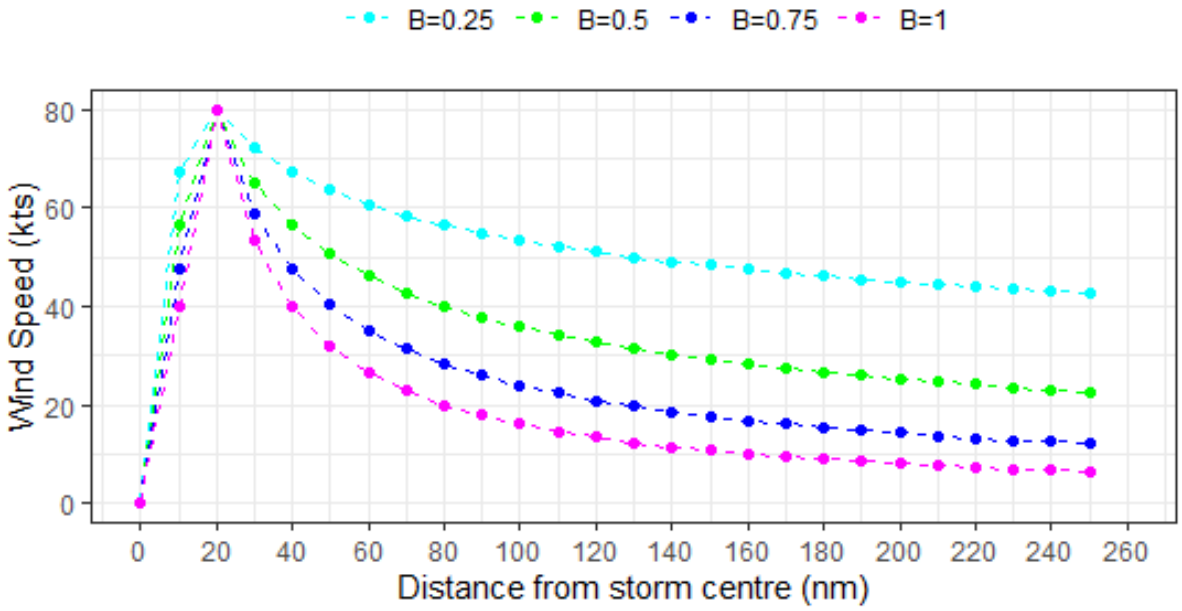



Figure 1.2: Rankine vortex wind profile examples for $V_{max} = 80kts$ and $R_{max} = 20nm$

A possible optimisation constraint to obtain the parameters is to pick $R_{max}$ and B such as to minimise the error between the Hurdat 2 best track wind radius and the TSR model's wind radius (Lea 2020). In figure 1.2 there are different wind profiles but they show that the maximum radius $R_{max} = 20$ is associated with the maximum wind speed $V_{max} = 80$ as expected. Furthermore, figure 1.2 illustrates that the Rankine model incorporates one of the fundamental characteristics of storms. At the center of the storm (distance from the eye of the storm equal 0), the wind speed is equal to 0, the wind speed increases rapidly as the distance from the hurricane eye increases until it reaches the maximum radius $R_{max}$ corresponding to the distance associated with the maximum wind speed $V_{max}$. The wind speed then decreases more slowly (compared to the previous increase to reach the maximum wind speed), as the distance from the hurricane eye increases beyond the maximum radius $R_{max}$.

### 1.2.2 Other calibration parameters: translational velocity, roughness correction and gust factor

#### 1.2.2.1 Storm translational velocity adjustment

The Rankine vortex generates a wind profile depending solely on the distance from the hurricane eye. However, the storm is not a perfect circle as it is in motion. Thus, there are some points (i.e. location), where the distance from the hurricane eye is larger or smaller than others, an illustration is provided in figure 1.3

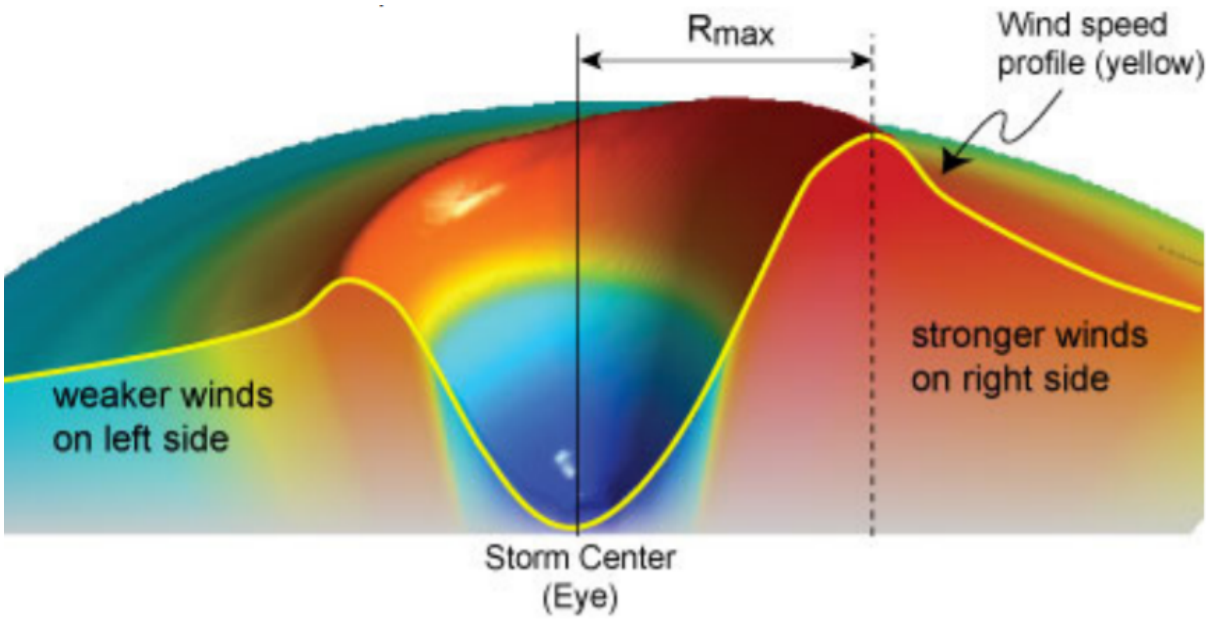



Figure 1.3: "Schematic cross section of a hurricane wind field. Rmax is the distance from the center of the storm to the location of the maximum winds. Note how winds are strongest on the right-hand side of the hurricane (assuming a northward direction) due to the additive effect of the hurricane's winds and the storm's forward motion " Source: Desflots (2010)

To account for this difficulty, an adjustment for translational speed is introduced. The storm translational velocity is obtained by evaluating the progression (distance) of the hurricane between two best track coordinates.

#### 1.2.2.2 Roughness correction factor

Another new parameter is added to the wind model : *the roughness correction factor*. A change in surface implies a change in surface friction, so as the storm moves from sea to land the storm's friction with the surface changes. The land surface is rougher

(more friction) than the sea surface such that as the storm makes landfall it loses speed, the average wind of the storm decreases. The roughness correction factor is based on Wieringa (1973) formula. It makes use of the roughness length over water while keeping the roughness length over land fixed at 0.04m as well as incorporating the reduction from the difference of the two roughness lengths for the sustained wind and peak gust. Then, the 1-min sustained wind is obtained after parameterisation of the wind model and accounting for the different aspects stated above.

#### 1.2.2.3 Gust factor

Finally, the value of interest is calculated : the maximum wind speed over a 3 second interval called the 3 second peak gust wind. The 3 second gust peak velocity is obtained by applying a gust factor to the maximum 1-min sustained wind. In the TSR model the analytical model by Wieringa (ibid.) is used to obtain the gust factor.

Given the previous elements described for the model the simulation for a given location proceeds as follows:
The data from the HURDAT 2 database (NOAA 2020) database is used to calibrate the model, a schematic representation of the data in the HURDAT 2 database (ibid.) is given in figure 1.4

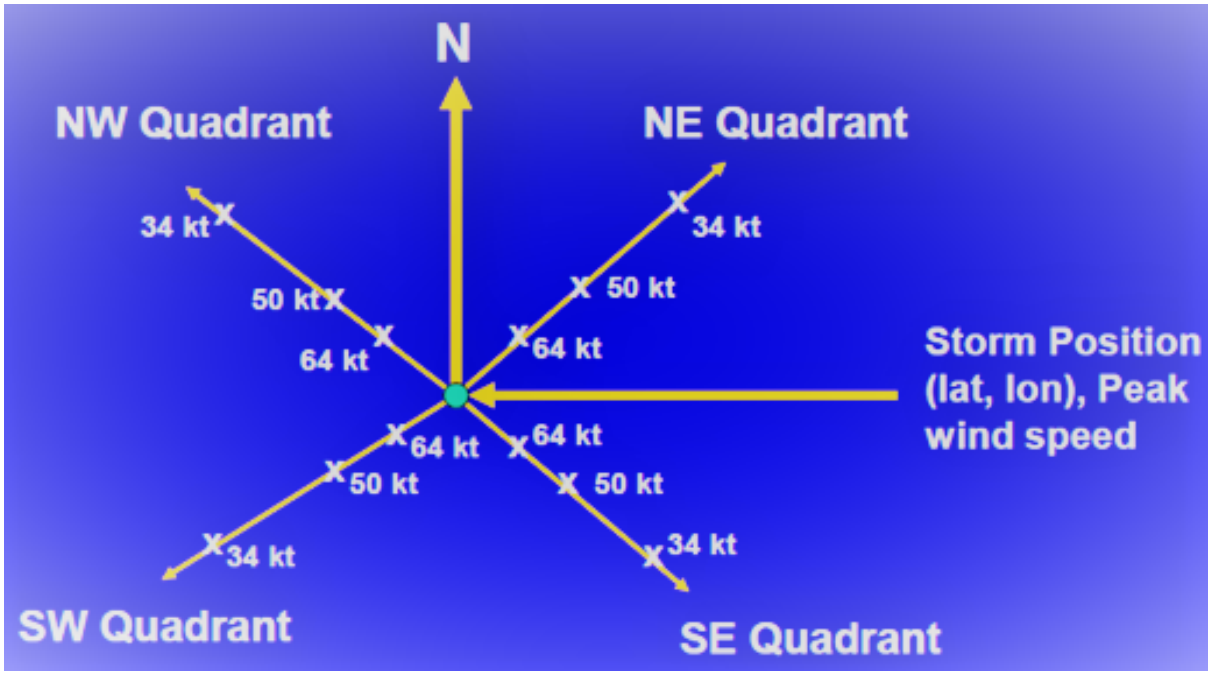



Figure 1.4: " Diagram of information contained in advisory for initial position and each forecast lead time. " Source: Lea (2020)

After calibrating the model to incorporate the wind speed available in the advisory HURDAT 2 database (NOAA 2020) the next step is to estimate the wind speed for radius that are not provided in the advisory HURDAT 2 database (ibid.). The parameters having been already obtained previously, only the radius r is needed to obtain these wind speed (see figure 1.5). Ultimately, this can be used to generate wind maps for the considered storm position and/or each forecast lead time. Wind maps can also be obtained for locations in between two lead times by interpolation.

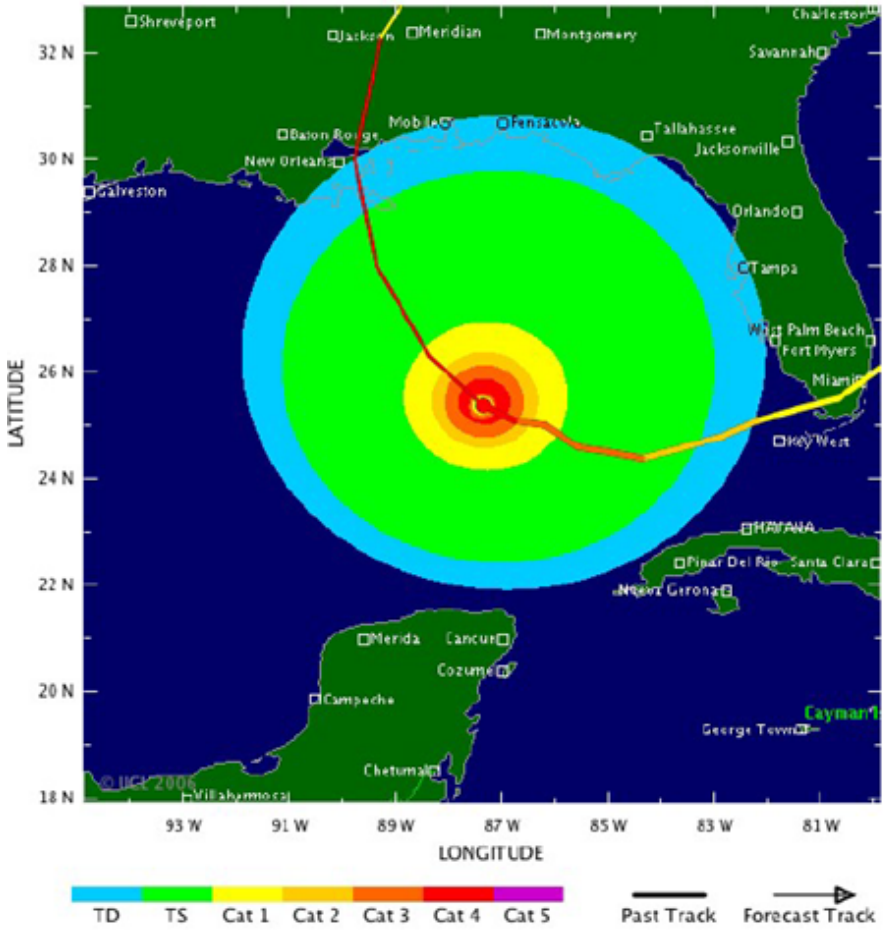
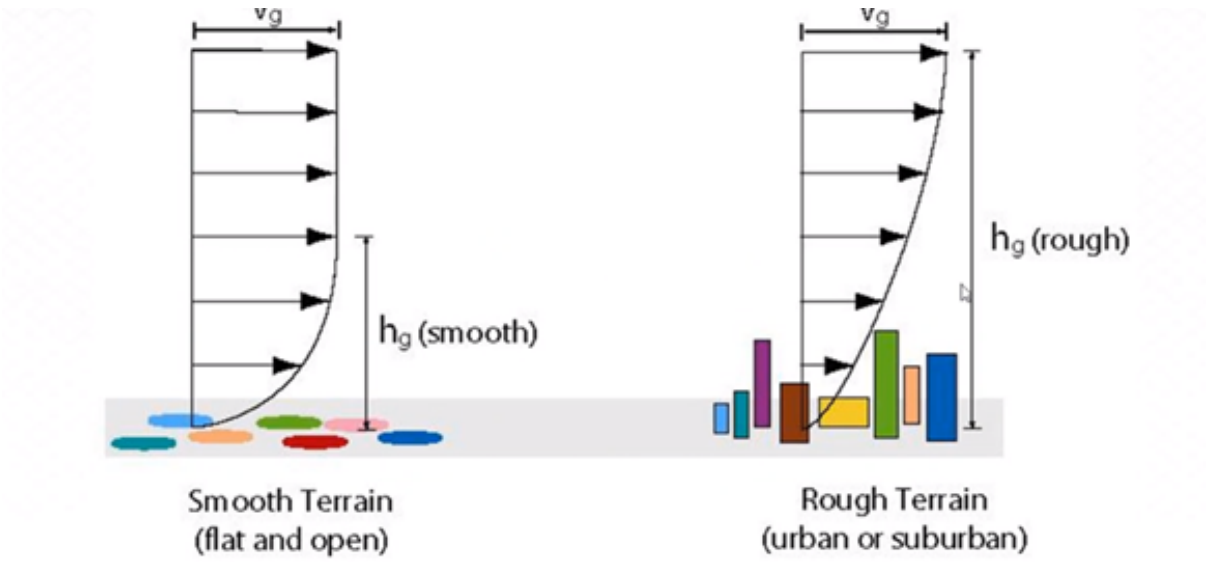



Figure 1.5: Wind map: wind speed estimation given the distance from the location with the highest wind speed in red Source: Lea (2020)

A major limit of the TSR model that will be discussed in a later chapter is the hypothesis overland of the friction of the wind with the surface being constant. It is false, as the roughness of an area changes with the topography. This has a direct impact on the estimated wind speed. An area with a smooth terrain will have a higher wind speed than the same area if it had a rough terrain. Indeed, intuitively the area with a rough terrain has obstacles in the wind path which will then reduce the wind speed (see figure 1.6).




***Sources: Simiu and Scanlan "Wind Effects on Structures", N. Cook "The Designers Guide to Wind Loading of Building Structures. Part 1.", ESDU Engineering Sciences Data "Wind Engineering.", etc.***

Figure 1.6: Wind speed: Smooth terrain vs Rough terrain

Now that the TSR model has been described the weather station data used to evaluate the performance of the TSR model is presented in the following chapter.

# Chapter 2

# Data

The data used in this study is the wind data (i.e weather station data) from storms that made landfall in the US between 2015-2019 extracted from the National Hurricane Centre reports. A quick overview of each of these storms is given. Afterwards the wind data is reviewed so it can be used for error analysis chapter 3 and accuracy purposes chapter 4.

## 2.1 Overview of storms that made landfall in the US between 2015-2019

### 2.1.1 Tropical storms that made landfall in the US between 2015-2019

Tropical Storm (TS) that made landfall in the US in 2015:

- Tropical storm Ana originated on the $3^{rd}$ of May and reached Tropical Storm (TS) on the $8^{th}$ of May and dissipated on the $11^{th}$ of May. It made landfall on the northeastern coast of South Carolina and made minor damage and was linked to one death (Stewart 2015).

- Tropical storm Bill originated in the Gulf of Mexico and made landfall on the coast of Texas. Tropical storm Bill had Tropical Storm (TS) status from the $16^{th}$ of June to the $18^{th}$ of June (Berg 2015).

  Darbinyan (2015) estimated that the economic damage due to storms in the US

in 2015 was USD 500 million and the total insured loss was USD $ 260 million and 3 people died as a result of these storms.

TS that made landfall in the US in 2016:

- Tropical storm Bonnie formed northeast of the Bahamas having TS status between the $27^{th}$ of May and the $4^{th}$ of June (Brennan 2016). The tropical storm made landfall on the $29^{th}$ of May in South Carolina on the east of Charleston (ibid.). No death was linked to storm Bonnie, the economic damage was in the millions but for claims in the hundreds (Darbinyan 2016).

- Tropical storm Colin originated from the west coast of Africa on the $27^{th}$ of May and reached TS status $5^{th}$ of June which was shortly lived as it lost TS status on the $7^{th}$ of June. No death was linked to storm Bonnie, the economic damage was in the millions but for claims in the hundreds (ibid.). Tropical storm Colin made landfall in Florida and Carolinas (ibid.).

- Tropical storm Julia had TS intensity from the $13^{th}$ to the $18^{th}$ of September. It made landfall in Florida near Jensen Beach on the $13^{th}$ of September. No information was found concerning the economic loss or insured loss due to tropical storm Julia. The NHC report states for storm Julia "It produced relatively minor impacts over land" (Blake 2017).

TS that made landfall in the US in 2017:

- Tropical storm Cindy made landfall on the $21^{st}$ of June in southern Louisiana (Berg 2018b). Podlaha (2017a) states that the "damage costs [...] were largely negligible". Tropical storm Cindy maintained a TS wind speed from the $20^{st}$ of June to the $23^{rd}$ of June.
- Tropical storm Emily had the same impact in terms of damage as storm Cindy (Podlaha 2017b). Storm Emily was a TS from $30^{th}$ of July to the $1^{st}$ of August and made landfall on the $31^{st}$ of July in Florida south of Tampa Bay (Cangialosi 2019).

TS that made landfall in the US in 2018:

- Tropical storm Alberto originated from the northwestern Caribbean sea on the $12^{th}$ of May (Berg 2018a). It made landfall on the $28^{th}$ of May Panhandle near Panama City,Florida (Bhattacharya 2018a). Tropical storm Alberto conserved a level of wind speed of at least TS status from the $25^{th}$ of May to the $31^{st}$ of May (Berg 2018a). 5 deaths were linked to tropical storm Alberto, it also caused millions in economic losses and thousands in claims (Bhattacharya 2018a).

- Tropical storm Gordon made landfall on the coast of Florida and at the Mississippi-Alabama border (Berg 2019). Bhattacharya (2018b) states that 2 deaths were due to tropical storm Gordon and an economic loss of more than USD 250 million including claims of more than USD 15 thousand (ibid.). Tropical storm Gordon had a wind speed greater than or equal to that of a TS between $3^{rd}$ and $6^{th}$ of September.

### 2.1.2 Hurricanes that made landfall in the US between 2015-2019

Hurricanes that made landfall in the US in 2016

- Hurricane Hermine made landfall in Florida on the $2^{nd}$ of September as a category 1 hurricane in regards to the Saffir-Simpson Hurricane Wind Scale (SSHWS). The first Hurricane to make landfall in a decade in Florida; hurricane Wilma was the last one since October 2005 (Darbinyan 2016; Berg 2017). Hurricane Hermine was at least TS between $28^{th}$ of August and $3^{rd}$ of September (Berg 2017). The hurricane caused 3 deaths, over USD 800 million in economic losses and over USD 400 million in insured losses(Darbinyan 2016).

- Hurricane Matthew originated on the $23^{rd}$ of September 2016 as a tropical wave situated on the west coast of Africa and dissipated on the $9^{th}$ of October 2016 near Atlantic Canada. On the $1^{st}$ of October 2016 the storm reached its maximum 1-min sustained average wind speed of 145kt and thus was classified as

a category 5 hurricane in accordance with the Saffir-Simpson Hurricane Wind Scale (SSHWS). The areas impacted by this storm were: "coasts of southwestern Haiti, extreme eastern Cuba, western Grand Bahama Island, and along the central coast of South Carolina" (Stewart 2017). Being a category 5 hurricane (highest category on the SSHWS scale), the economic loss was considerable and was evaluated to be USD 15 billion, the insured loss USD 4.5 billion in 2017 and nearly 600 deaths were directly linked to this hurricane Adam (2017). Causing more than 500 deaths in Haiti it became the "deadliest Atlantic hurricane since Hurricane Stan in 2005".

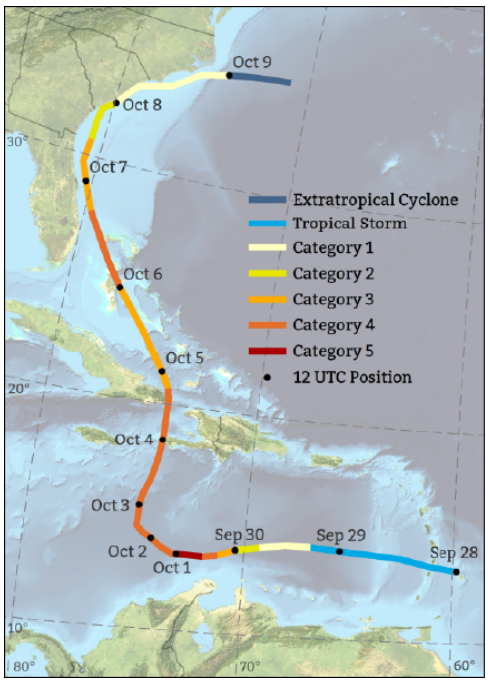



Figure 2.1: Map Matthew best track data from hurdata2 (NHC) and wind speed intensity according to the SSHWS classification and the US National Weather Service tropical definitions Source: Adam (2017)

Hurricanes that made landfall in the US in 2017

– Hurricane Harvey made landfall in the US along the coast of Texas as a category 4 hurricane (Zelinsky 2018). With more than USD 30 billion in insured losses it is one of the most costly storms to make landfall in the US; the number of deaths was estimated to be 90 and the economic loss around USD 100 billion (Bogg 2018). It had the wind speed at least equivalent to TS from $17^{th}$ of August to $1^{st}$ of September.

– Hurricane Irma made 7 landfalls and 4 of them as a category 5 hurricane in the northern Caribbean Islands, highest category on the SSHWS scale (John P. Cangialosi 2018). These events took place between the $30^{th}$ of August and the $12^{th}$ of September. It is also one of the most costly storm that ever occurred in the Atlantic basin with a total economic loss of USD 55 billion, USD 23 billion and 134 deaths (Bogg 2018).

– Hurricane Nate maintained at least TS wind speed between the $4^{th}$ and $8^{th}$ of October. It made landfall on the Gulf Coast as a Category 1 hurricane (Beven 2018). No casualties were associated with storm Nate; the economic damage was estimated to be around USD 250 million and the claims USD 20.000 (Claire 2017).

Hurricanes that made landfall in the US in 2018

– Hurricane Florence a category 4 hurricane made land fall as a category 1 hurricane on the coast of North Carolina (Stacy 2019). 53 lives were lost, the claims amounted around USD 350.000 with an economic loss of USD 10 billion (Bhattacharya 2018b). These events took place from the $31^{st}$ of August to the $17^{th}$ of September.

– Hurricane Michael is a category 5 hurricane that made multiple landfalls: near Mexico Beach, Tyndall Air Force Base, Florida and others (J.Beven 2019). The economic loss due to the hurricane was USD 17.0 billion with insured losses at USD 10.0 billion the number of death due to this hurricane is 32 (Bhattacharya 2018c). The period where hurricane Michael was at least considered of TS level was from $7^{th}$ to $11^{th}$ of October.

Hurricanes that made landfall in the US in 2019

– Hurricane Barry originated from southwestern Kansas on the $2^{nd}$ of July making landfall as a category 1 hurricane in Louisiana on the $13^{th}$ of July (R.Berg

2019). No casualties were associated with hurricane Barry; nevertheless the storm caused more than USD 630 million in economic losses with more than 50 thousand in insurance losses (S.Bowen 2019). Hurricane Barry had a wind speed of a least TS level between the $11^{th}$ and $15^{th}$ of July (R.Berg 2019).

– Hurricane Dorian is a category 5 hurricane, the strongest hurricane to ever make landfall in the Bahamas (Hagen 2020). 73 deaths were directly cause by hurricane Dorian it also generated an economic loss of USD 8.0 billion and an insured loss of USD 2.5 billion (S.Bowen 2019). These events took place between the $24^{th}$ of August and the $7^{th}$ of September (Hagen 2020).

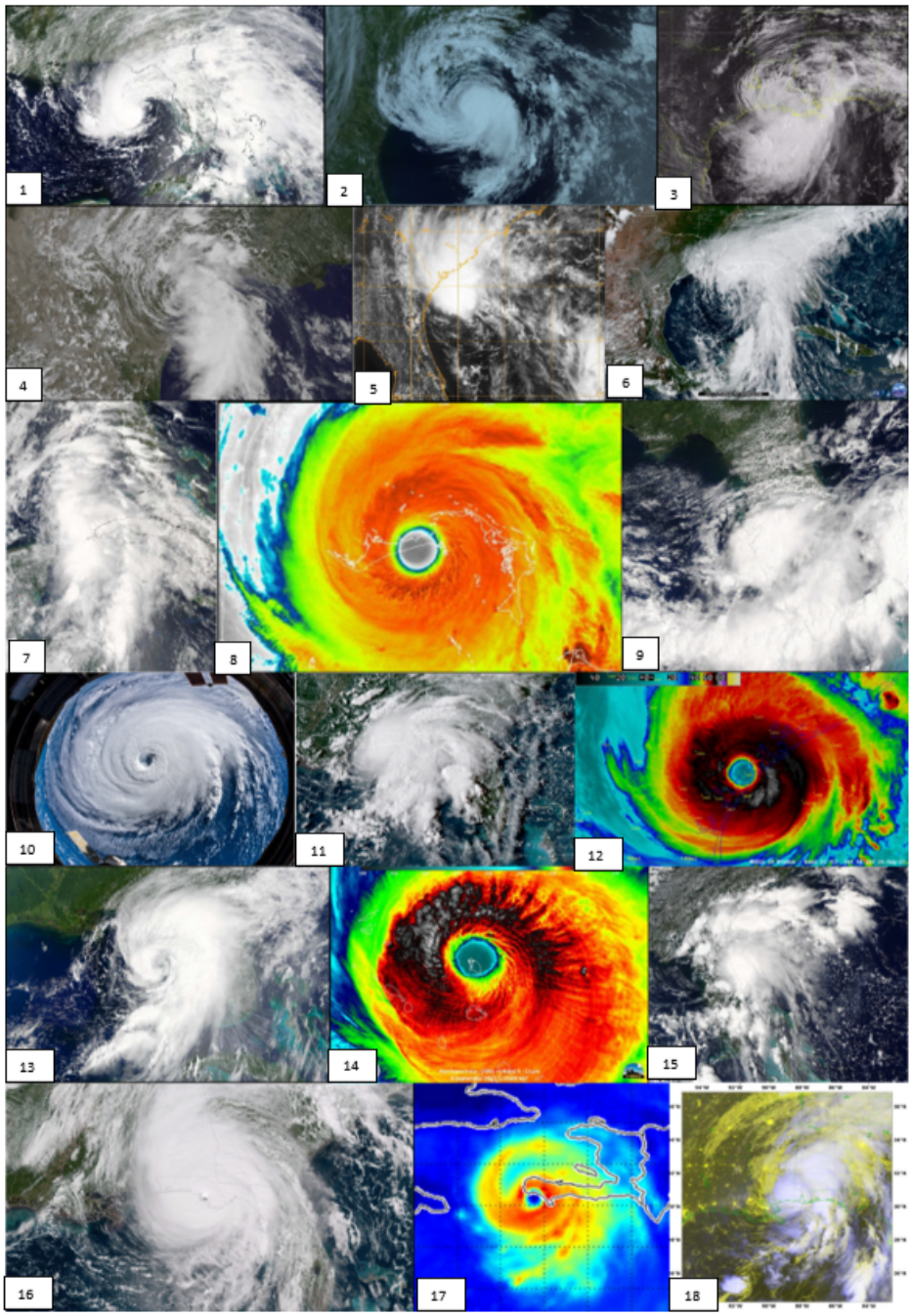



Figure 2.2: Snapshots during the life cycle of the 18 storms in the weather station data. 1. ALBERTO (Berg 2018a), 2. ANA (Stewart 2015), 3. BARRY (R.Berg 2019), 4. BILL (Berg 2015), 5. BONNIE (Brennan 2016), 6. CINDY (Berg 2018b), 7. COLIN (Penny 2017), 8. DORIAN (Hagen 2020), 9. EMILY (Cangialosi 2019), 10. FLORENCE (Stacy 2019), 11. GORDON (Berg 2019), 12. HARVEY(Zelinsky 2018), 13. HERMINE (Berg 2017), 14. IRMA (John P. Cangialosi 2018), 15. JULIA (Blake 2017), 16. MICHAEL (J.Beven 2019), 17. MATTHEW (Stewart 2017), 18. NATE (Beven 2018). A more in depth legend can be found in appendix B

*This section has shown how each individual storm has its own pattern of behaviour (see figure 2.2). It is then justified to assume that the simulation of the wind profile of a storm is independent from the simulation of another storm. This hypothesis will be used*

*all along in this study.*

## 2.2 Referential data

The referential data, is the data that is used for verification purpose in the accuracy evaluation done in chapter 4.

The National Hurricane Centre (NHC) publishes reports on storms that occur all around the globe. Each report has the same basic structure: name of the storm, description of the evaluation of the storm from the genesis of the storm to its dissipation, records of 6 hourly-interval (i.e lead time) of the location of the storm (i.e longitude and latitude points), the observed gust, the observed sustained winds and others. These reports can be found on the NHC websites and in the bibliography.

The data used as referential in this study are the location, observed gust and observed sustained wind present in these reports for the storms that made landfall in the US from 2015-2019. The reports being in pdf, the tables containing the information previously discussed were extracted and converted into a CSV format to be able to be used in ACCESS, Excel, R and in the Tropical Storm Risk (TSR) model.

Hereinafter, the data extracted from the NHC reports are referred to as the weather station data.

### 2.2.1 Data homogeneity

Another aspect that needs to be accounted for is the lack of homogeneity in the way the weather stations data were collected. In order to measure the sustained wind and gust instruments called anemometers are used in weather stations. The general consensus in the US (Ginger 2008) as well as an underline assumption in the TSR model is that the anemometer used at a given weather station is situated at a height of 10 meters and the averaging period of the sustained wind is 1 minute. However, the NHC reports show that in reality this is not actually the case, different height and averaging periods are present for the anemometers that compose the weather station data. It is thus necessary to clean, homogenise the weather station data before it can be used in this study.

The fact that the height of the anemometers are uneven cannot be overlooked as this has an effect on the wind speed. Given a fixed location, the value of the wind speed changes with altitude. More specifically, wind speed increases as height increases (Mohandes 2011; Haby 2020). Haby (2020) gives 3 reasons to explain this phenomenon:

- Wind gradient

- Surface friction

- Air density

In regards to the sustained wind averaging period the effect of different averaging period on the wind speed isn't as straight forward. Ginger (2010) argues that longer averaging periods (in respect to the 1-min, 2-min, 3-min averaging periods) tend to result to smoother wind speed and are most likely to be closer to the real wind speed mean. However, the authors also specify that shorter averaging periods are still statistically sound and efficient as estimate of the real sustained wind mean in the absence of turbulence in the wind. They conclude by stating different averaging period are not equivalent as the more the averaging period diverge from another so does their variance. Thus in this study the threshold of 5-min averaging period was chosen such that the weather station data used to conduct the analysis have anemometers with averaging period less than or equal to 5-min.

The different source of discrepancy in the weather station data are due to:

1. Anemometer height different from 10m.

   After reviewing the weather station data it was found that height ranges from 1 meter to 160 meters.

2. Averaging period for the 1-min sustained wind greater than 5 minutes.

   After reviewing the weather station data it was found that the wind averaging period ranges from 1 minute to 60 minutes.

3. Inconsistency of the weather station data (i.e the data is recorded incorrectly)

   The notion of consistency of the weather station data is important. In the NHC reports, there are annotations that reference "Incomplete data" which mentions the failure of anemometers at some weather stations and thus they record the wind speed values inaccurately. More specifically there are some weather stations where the anemometers just stopped recording.

   Example of incomplete weather station data are given below:

   - The NHC report for hurricane Matthew mentions "Wind speed data missing 0510-0650 UTC 3 October 2016", "All wind data missing 0800-1000 UTC 6 October 2016 " and "All wind data missing 1300 UTC 9 October – 0200 10 October 2016." (Stewart 2017)
   - The NHC report for hurricane Harvey mentions "Anemometer damaged or data recording otherwise interrupted and likely did not record maximum winds" and "Station 42020 went adrift on 8/25 around 1800 UTC and stopped reporting at that time." (Zelinsky 2018)

   Thus when there was any doubt or mention that the value of the wind speed might be incorrect the corresponding weather station data was removed from the database.

   5% of the data was found to be recorded incorrectly for the sustained wind and 4% for gust.

4. The data was collected over sea (i.e the weather station data was collected when the storm was still at sea).

   The notion of weather station data collected over sea was added as source of discrepancy in the database because it is more relevant to focus on the accuracy of the TSR model over land rather than sea as that is where the storms has more impact on human lives and the damage occurs. The site

https://www.coordonnees-gps.fr/ was used to verify the location of the data point by filling in the longitude and latitude points to know if the data point is on land or sea.

Around 60% of the data was found to be recorded over land for the sustained wind and for gust data.

Any weather station data that has any of the 4 source of discrepancy mentioned above is dropped from the database.

Every weather station data present in the homogenised database has an anemometer at a height of 10m, with an averaging period of less than or equal to 5-min and that the data was recorded correctly over land. This is the referential data to evaluate the accuracy of the TSR model.

Initially, for the wind data there were 1437 records after removing the data as explained above only 554 records were kept. As well in the gust data, at first there were 2025 records but finally only 825 records were kept. In both cases around 60% of the weather station data were dropped.

Now that the wind and gust data have been homogenised and cleaned the error data can be constructed.
In the next section, an in depth overview of the error database used in this study is given.

### 2.2.2 Description of the error database

This section is an overview of the datasets that are used in this study.

The database that is used in this study to evaluate the accuracy of the TSR model contains the characteristics of the 18 storms that made landfall in the US from 2015 to 2019 (i.e the weather station data). The names of these storms are given in table 2.1 with their category according to the Saffir-Simpson Hurricane Wind Scale (SSHWS).

*The detailed table of Saffir-Simpson Hurricane Wind Scale provided by NHC is shown in annex A; it contains examples and explanations of the damage and impact associated with the intensity (wind speed) of a storm (i.e for the 5 categories).*

| Name | Year | Category | NB Record gust | NB Record wind |
|---|---|---|---|---|
| ALBERTO | 2018 | TS | 39 | 22 |
| ANA | 2015 | TS | 1 | 1 |
| BARRY | 2019 | 1 | 1 | 1 |
| BILL | 2015 | TS | 32 | 7 |
| BONNIE | 2016 | TS | 5 | 0 |
| CINDY | 2017 | TS | 7 | 3 |
| COLIN | 2016 | TS | 22 | 5 |
| DORIAN | 2019 | 5 | 29 | 22 |
| EMILY | 2017 | TS | 6 | 3 |
| FLORENCE | 2018 | 4 | 30 | 29 |
| GORDON | 2018 | TS | 35 | 18 |
| HARVEY | 2017 | 4 | 78 | 61 |
| HERMINE | 2016 | 1 | 79 | 53 |
| IRMA | 2017 | 5 | 128 | 126 |
| JULIA | 2016 | TS | 8 | 8 |
| MATTHEW | 2016 | 5 | 98 | 91 |
| MICHAEL | 2018 | 5 | 187 | 82 |
| NATE | 2017 | 1 | 40 | 22 |
| | | Total | 825 | 554 |

Table 2.1: For each storm in the database : name, year, number of recordings present in the database (wind data and gust data) and category of the storm according to SSHWS

Out of the 18 storms in the database 9 are Tropical Storm (TS), 3 are of category 1, 2 of category 4 and 4 of category 5.

The database is made up of two datasets. Each dataset has 7 variables. The first dataset contains the 3 second peak gust wind of the hurricanes and the second dataset contains the 1-minute sustained average wind speed of the hurricanes.
Hereinafter, the $1^{st}$ dataset is referred to as the gust data and the $2^{nd}$ dataset as the wind data. The only difference between the gust and wind data is that the last two variables:

observed gust and TSR gust in the $1^{st}$ dataset become observed wind and TSR wind.

There are 7 variables in each dataset:

- Name: the name of the storm
- IdRecord: Rank, order that the record for a given storm appears in the data.
- Year: the year the hurricane occurred
- Longitude: longitude recorded for a given IdRecord of a storm
- Latitude: latitude recorded for a given IdRecord of a storm
- Observed gust/wind : observed gust/wind for a given IdRecord of a storm extracted from the NHC report
- TSR gust/wind : value obtained with the TSR model for a given IdRecord of a storm

The data for all the variables apart from the data for the TSR model output variables were extracted from NHC reports.

Given the longitude and latitude values extracted from NHC reports the gust and wind values of the TSR model are calculated. Thus for every location present in the NHC reports the equivalent gust and wind values are calculated with the TSR model. This also brings a certain level of uncertainty as locations present in the NHC reports don't always match-up with the location in the TSR model. The observed weather station data has a two digit precision meanwhile the TSR model has a four digit precision so this creates a spatial uncertainty and can thus be a source of discrepancy between the observed data and the estimated data by the TSR model. An example of spatial uncertainty as a cause of large error is shown in section 3.2.1.

The variables that are going to be used the most in this study are the observed gust/wind and the estimated gust/wind by the TSR model. A more in depth overview of the value of these variables are given in the next subsection.

#### 2.2.2.1 Descriptive statistics: gust and wind variables in the database

A visual representation of the data for the variables observed gust, observed wind, TSR gust and TSR wind, is given in figure 2.3.
In figure 2.3. the bisector is poltted as a reference. Indeed, if there was a 100% accuracy between the observed data and the TSR model, the points in the two scatter plots should all be aligned on the bisector. In reality, 100% accuracy is impossible as a model is a simplified representation of reality. This implies the existence of inherent errors from the modelisation process: irreducible errors.

However, not all errors are due to the modelisation process some are linked to the quality of the data collected (or observed). Generally, errors are commonly grouped into two classes : systematic errors and random errors (Taylor 1982). This is where the importance of error analysis comes to play in order to identify and separate the errors that are inherent to the model: random errors from systematic errors which are often linked to data inconsistency : outliers, human error, inadequacy of the instrument used to measure the data (Regler 1993). Thus, systematic errors displace the mean of error in a given direction meanwhile random errors have an arbitrary effect on the mean. Indeed, the underlying hypothesis is that random errors are associated with a probability distribution with mean 0 (Taylor 1982).

Thus in order to have a reliable measure of the accuracy of the TSR model, systematic error needs to be reduced even more eliminated if possible. A first step has already been taken to do so in section 2.2.1 by homogenising the data.

Eliminating systematic error requires identifying the pairs of $(x_{obs}, x_{TSR})$ that produce "large errors", where $x_{obs}$ is an observed data point of gust or wind and $x_{TSR}$ is a data point from the estimated gust or wind by the TSR model. These pairs of data points will then be analysed to know if the error is due to a systematic errors that can be dealt with. The term "large errors" still needs to be defined.

Before tackling the issue of systematic errors, a graphical representation of the empir-

ical distribution of errors for gust and wind is shown in figure 2.4 and 2.5 and general statistics are performed on the error data.

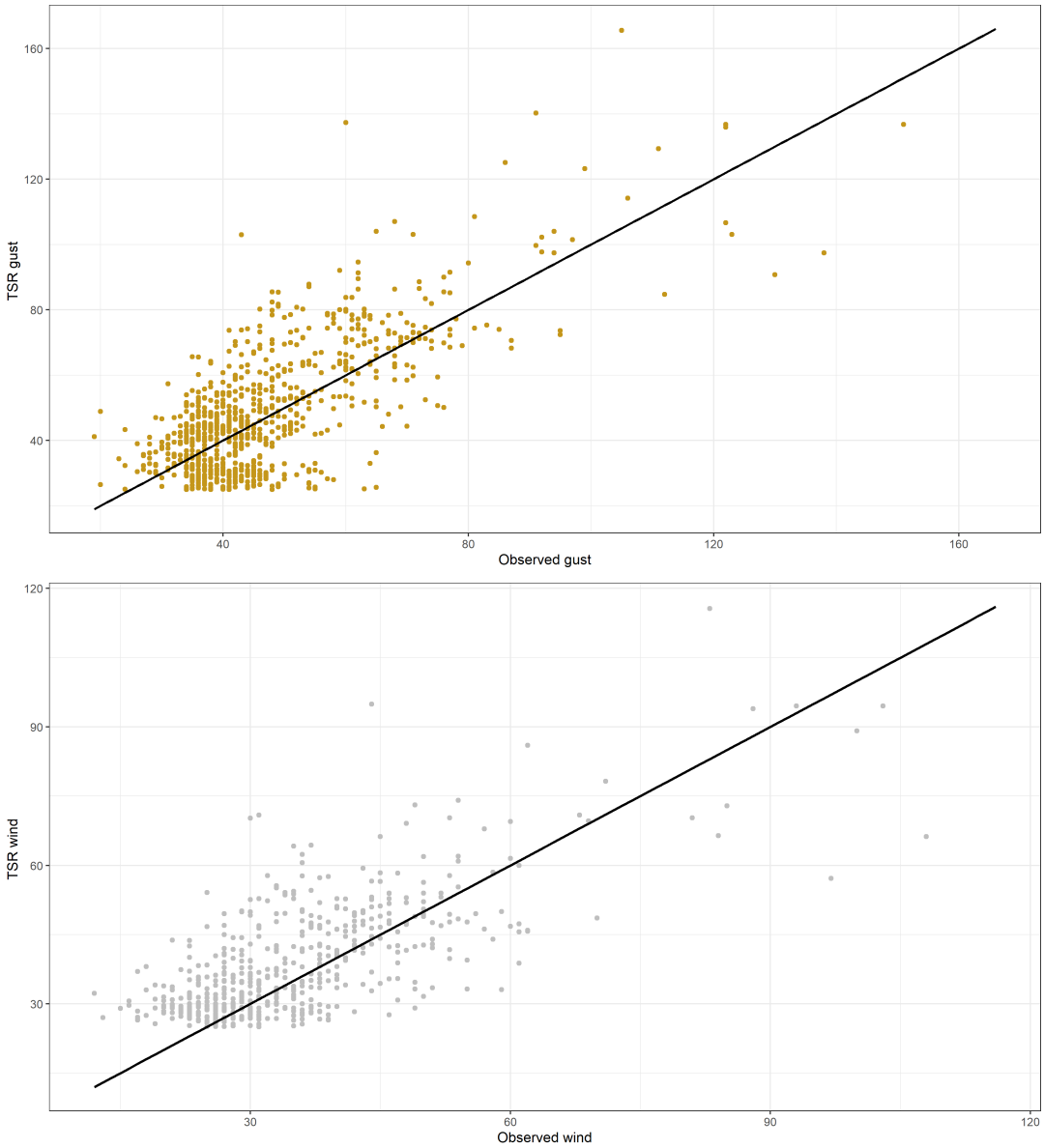



Figure 2.3: The top (bottom) graph is the observed gust (wind) data against TSR gust(wind) data, with the bisector. This shows where the points on the scatterplot would have aligned if the observed data and TSR data were equal, a case of 100% accuracy.

In this study the error for a pair of data points $\left(x_{obs_i}.x_{TSR_i}\right)$ is given by

$$\varepsilon_i = x_{obs_i} - x_{TSR_i} \tag{2.1}$$

, where i is the i-ième data point from either the gust or wind dataset.

Figures 2.4 and 2.5 suggest that the empirical gust and wind error distributions have a bell shaped distribution. Based on the classification of errors by Taylor (1982) this means that the systematic error in the database is less predominate than the random errors as the distribution of errors is centered around the mean of errors. However, the presence of systematic errors can be deduced by the fact that the empirical distribution of errors are not centered around 0. This implies that the systematic errors are pulling the mean away from 0.

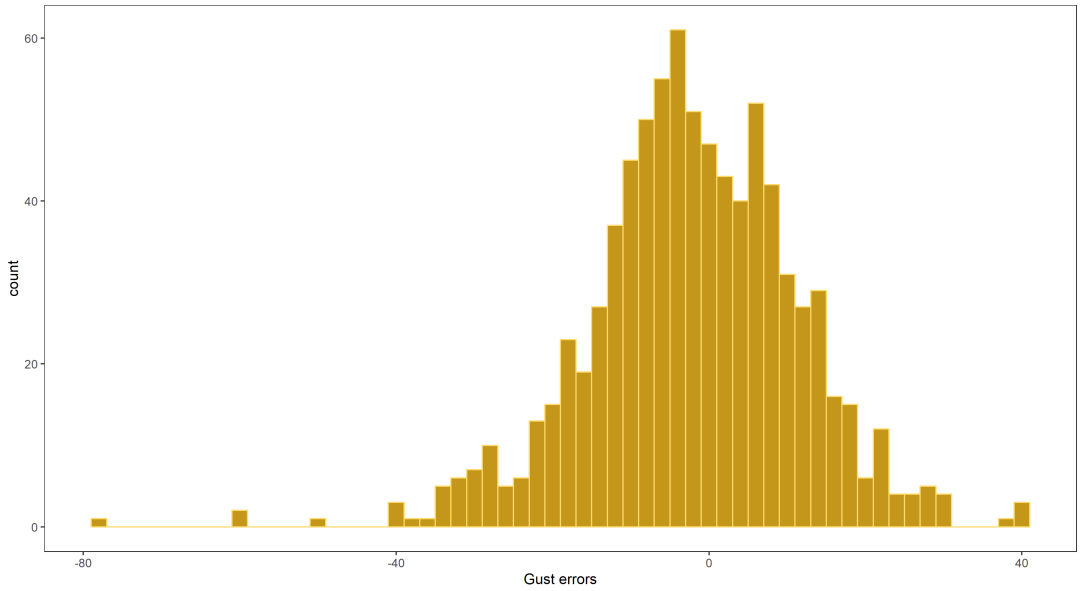



Figure 2.4: Histogram of gust error : shows the empirical distribution of gust error has a bell shaped distribution, that resembles the normal distribution.

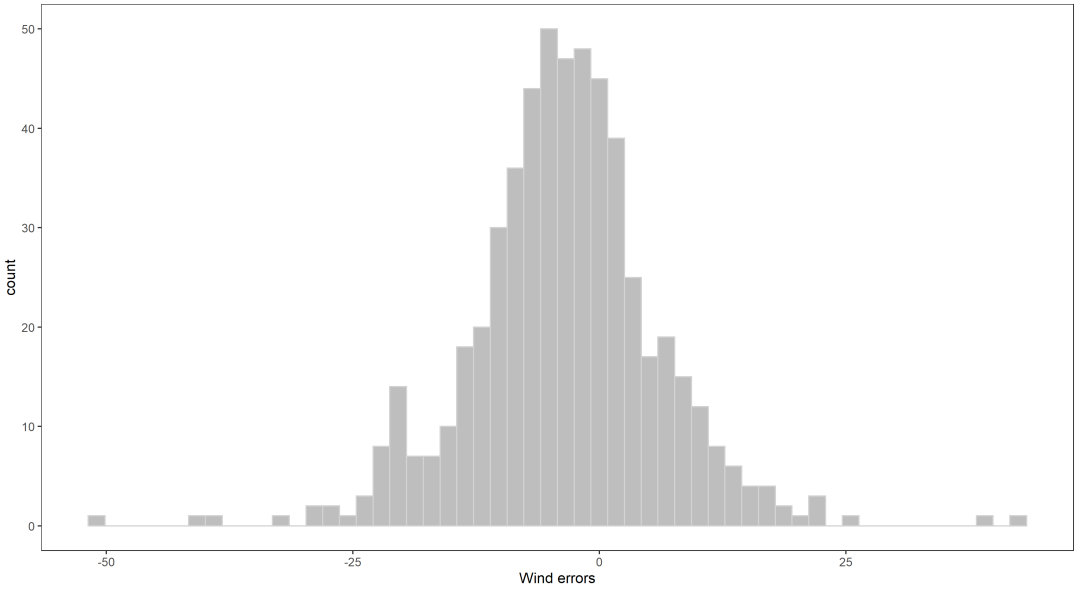



Figure 2.5: Histogram of wind error : shows the empirical distribution of wind error has a bell shaped distribution, that resembles the normal distribution.

The gust distribution of error figure 2.4 and the wind distribution of error figure 2.5 appear to be centered with a few extreme values. They both have a bell shaped distribution that looks similar to that of the normal distribution. However, both distributions are not symmetric (centered) as the mean, median and mode are not equal (Manikandan 2011). These 3 values are given in table 2.4 for gust and wind.

Next, the first descriptive statistics values are calculated : mean and standard deviation (SD).
Given that the two datasets (gust data and wind data) are made up of 18 storms the overall mean of error for each dataset is calculated as follows:

$$\begin{aligned} \bar{\varepsilon} &= \frac{1}{m}\sum_{j=1}^{18}\sum_{i=1}^{n}\frac{1}{n}(x^j_{obs_i} - x^j_{TSR_i}) \\ &= \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{\varepsilon^j_i}{n} \\ &= \frac{1}{m}\sum_{j=1}^{m}\bar{\varepsilon}^j \end{aligned} \tag{2.2}$$

. where n is the number of records in the dataset for a given storm and m the number

of storms in the dataset (m=18 for gust data and m=17 wind data, cf: table 2.1) .

The mean error gust $\bar{\varepsilon}_{gust} = -2.09$ and the mean error wind $\bar{\varepsilon}_{wind} = -2.35$.

The standard deviation (SD) of error for the gust and wind can not be calculated the same way as the sum of the 18 standard deviations. Berendsen (2011) explains that the propagation of errors cannot be lineary accounted for by the standard deviation. This should be done by adding up the variance and covariance. An example from Berendsen (ibid.) is shown below .

Defining. $Z = X - Y$ then the variance of Z is

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 + 2Cov(X, Y) \tag{2.3}$$

, the standard deviation of Z is $\sqrt{\sigma_Z^2} = \sigma_Z$.

The covariance incorporates the correlation between the variables X and Y in the variance of Z. In this study, the hypothesis of independence of errors between the 18 storms is made under the assumptions that the atmospheric and oceanic conditions of one storm has no influence on the other storms as well as the modelling process of one storm has no link to the modelling process of the other storms. Under these assumptions the covariance of errors is equal to 0. Thus the variance is calculated as follows to obtain the standard deviation of errors.

$$\begin{aligned} \sigma_\varepsilon^2 &= \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{n} \left( \frac{\varepsilon_i^j - \bar{\varepsilon}}{n} \right)^2 \\ &= \frac{\sum_{j=1}^{m} \sigma_{\varepsilon^j}^2}{m} \\ \sigma_\varepsilon &= \sqrt{\sigma_\varepsilon^2} \end{aligned} \tag{2.4}$$

, where n is the number of records in the dataset for a given storm and m the number

of storms in the dataset that have more than 1 record in the dataset (i.e $n \geq 2$). Indeed, for the calculation of the variance there needs to be at least two records in the dataset for the concerning storm.

The standard deviation of gust error $\sigma_{\varepsilon_{gust}} = 11.41$ and the standard deviation (SD) of wind error $\sigma_{\varepsilon_{wind}} = 7.58$.

The number of records (i.e data points) per storm n are not large numbers (cf table 2.1) the choice is made to use the unbiased estimator of the variance with n-1 (Berendsen 2011) this also changes the value of the SD of errors.

$$\begin{aligned} \hat{\sigma}_{\varepsilon}^2 &= \frac{1}{m} \sum_{j=1}^{m} \frac{n^2}{(n-1)^2} \sigma_{\varepsilon^j}^2 \\ \hat{\sigma}_{\varepsilon}^2 &= \frac{\sum_{j=1}^{m} \hat{\sigma}_{\varepsilon^j}^2}{m} \end{aligned} \tag{2.5}$$

, where n is the number of records in the dataset for a given storm and m the number of storms in the dataset that have more than 1 record in the dataset (i.e $n \geq 2$).

The unbiased estimator of the SD of gust error $\hat{\sigma}_{\varepsilon_{gust}} = 11.42$ and the unbiased estimator of the SD of wind error $\hat{\sigma}_{\varepsilon_{wind}} = 7.59$ are slightly higher than the value obtained previously.

| Name Storm | Mean | Var | SD | NB records |
|---|---|---|---|---|
| ANA | 0,30 | 0,00 | 0,00 | 1 |
| BILL | -3,35 | 51,02 | 7,14 | 32 |
| BONNIE | -1,66 | 31,73 | 5,63 | 5 |
| COLIN | 7,12 | 91,17 | 9,55 | 22 |
| HERMINE | -0,87 | 149,38 | 12,22 | 79 |
| JULIA | -8,29 | 16,14 | 4,02 | 8 |
| MATTHEW | -9,20 | 121,63 | 11,03 | 98 |
| CINDY | -3,24 | 13,17 | 3,63 | 7 |
| EMILY | 1,02 | 90,44 | 9,51 | 6 |
| HARVEY | -4,32 | 202,31 | 14,22 | 78 |
| IRMA | -2,16 | 240,33 | 15,50 | 128 |
| NATE | -1,63 | 110,17 | 10,50 | 40 |
| ALBERTO | -0,81 | 45,04 | 6,71 | 39 |
| FLORENCE | -13,00 | 184,52 | 13,58 | 30 |
| GORDON | -0,24 | 121,44 | 11,02 | 35 |
| MICHAEL | 2,95 | 172,87 | 13,15 | 187 |
| BARRY | 4,30 | 0,00 | 0,00 | 1 |
| DORIAN | -4,61 | 440,72 | 20,99 | 29 |
| Global | -2,09 | 130,45 | 11,42 | 825 |

Table 2.2: Mean, variance, standard deviation, number of records for storms for the gust error data

| Name Storm | Mean | Var | SD | NB records |
|---|---|---|---|---|
| ANA | 9,60 | 0,00 | 0,00 | 1 |
| BILL | -1,69 | 19,98 | 4,47 | 7 |
| BONNIE | 0,00 | 0,00 | 0,00 | 0 |
| COLIN | -6,92 | 24,45 | 4,95 | 5 |
| HERMINE | -3,36 | 69,02 | 8,31 | 53 |
| JULIA | -0,31 | 19,14 | 4,38 | 8 |
| MATTHEW | -5,05 | 54,11 | 7,36 | 91 |
| CINDY | 3,50 | 32,13 | 5,67 | 3 |
| EMILY | 1,33 | 4,80 | 2,19 | 3 |
| HARVEY | -2,79 | 104,66 | 10,23 | 61 |
| IRMA | -0,73 | 172,90 | 13,15 | 126 |
| NATE | -3,77 | 42,89 | 6,55 | 22 |
| ALBERTO | -4,22 | 13,39 | 3,66 | 22 |
| FLORENCE | -9,19 | 92,24 | 9,60 | 29 |
| GORDON | -4,42 | 34,18 | 5,85 | 18 |
| MICHAEL | -4,67 | 64,86 | 8,05 | 82 |
| BARRY | 0,20 | 0,00 | 0,00 | 1 |
| DORIAN | -7,49 | 112,12 | 10,59 | 22 |
| Global | -2,35 | 57,60 | 7,59 | 554 |

Table 2.3: Mean, variance, standard deviation, number of records for storms for the wind error data

Next. the median, min, max, $Q_{25\%}$ and $Q_{75\%}$ are calculated for the error data.

| Data | Min | 1st Qu | Median | Mean | Mode | 3rd Qu | Max |
|---|---|---|---|---|---|---|---|
| Gust | -77.3 | 0 | 0 | -2.1 | 0 | 0 | 40.6 |
| Wind | -50.9 | -0.5 | 0 | -2.4 | 0 | 0 | 41.8 |

Table 2.4: Gust and Wind error data descriptive statistics.

A look at table 2.4 shows that the value of the maximum error in both the gust and wind data are rather close. However, this is not the case for the minimum value, which is much lower for gust error compared to the wind error data.

Now that a general overview of the error data has been done, the focus is going to be on identifying large errors and understanding why the gap between the observed value and estimated value is so important. This is done in the following chapter but before

that the notion of large errors is defined.

# Chapter 3

# Error analysis

This chapter focuses on identifying large errors, analysing and explaining the source of the gap when possible.

## 3.1 Large error (outlier) detection techniques

### 3.1.1 Theory

As discussed in subsection 2.2.2.1 descriptive statistics are inefficient in identifying anomalies in a dataset. The type of anomaly focused on in this study are "large errors" referred to hereinafter as outliers.

Definition 1 : Outlier (Hawkins 1980)

> An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.

Definition 2 : Outlier (Barnett 1994)

> An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.

Hawkins (1980) and Barnett (1994) in their definitions of outlier both convey the same sense of lack of similarity of the observation with the rest of the dataset so much so that the data point seems out of place. The difficulty in identifying outliers comes from the definition itself. Indeed, Hawkins (1980) mentions deviation and Barnett (1994) states

inconsistency as criterias for labelling outliers . However, there is no absolute definition for the notion of deviation and inconsistency from a dataset. Thus the difficulty in identifying outliers, as the results varies considerably depending on how these notions are defined.

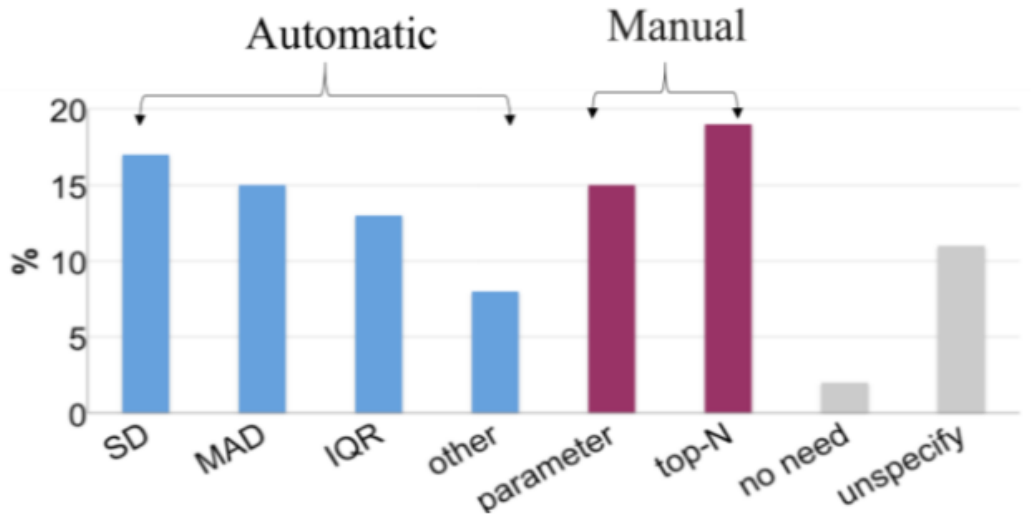



Figure 3.1: "Frequency of Usage of Different Threshold Selection Techniques in Outlier Detection Literature between June 2016 and June 2018" by Yang (2019)

Yang (2019) surveyed different outlier techniques used and mentioned in literature over the period of June 2016 and June 2018, the results are shown in figure 3.1. In this study they divide outlier detection methods into two classes:

- Manual: the user specifies a threshold, any observation beyond that point is flagged as an outlier.

- Automatic: the dataset is used in the outlier detection process.

The choice was made to use an outlier labelling technique that belongs to the automatic class as defined by Yang (ibid.). The automatic class is made up of 4 components : standard deviation (SD), Mean Absolute Deviation (MAD), Interquartile Range (IQR) and other. The other refers to outlying methods that are less rampant but most of them are distance based techniques. All 4 methods from the automatic class were investigated. The standard deviation (SD) technique is presented in the following paragraph as it is the one that is used in this study.

*The "other" technique considered in this study is the Local Outlier Factor (LOF) that is*

*in appendix C.3.*

#### 3.1.1.1 The standard deviation threshold

A widely known and classic outlier detection technique based on the distribution of the variable of interest $X$ is informaly called the "$3\sigma$ rule"(Ng 1997) and Yang (2019) refers to it as the standard deviation (SD) threshold technique. According to Yang (ibid.) this method was first introduced in 1962 by Webb (1962). Under the assumption that the variable of interest X follows a normal distribution, any data point that doesn't belong to the following interval is labelled as an outlier :

$$[\mu_X - 3\sigma_X; \mu_X + 3\sigma_X] \tag{3.1}$$

, where $\mu_X$ is the mean of the variable of interest X and $\sigma_X$ the standard the deviation of the variable of interest X.

The value of 3 was not randomly chosen, it is the 99.87% quantile of the standard normal distribution. Nelson (2011) states that this threshold was chosen under the presumption that only a small number of data points from a dataset are expected to be outliers. Thus, if the variable of interest X follows a normal distribution 0.27% of the data points in the dataset are labeled as outliers with the standard deviation (SD) threshold technique. This approach can be extended to other cases where the distribution of X is not normal, the value of -3 (respectively 3) in equation (3.1) is then replaced with the 0.13% (respectively 99.87%) quantile of the distribution of X.

Yang (2019) regroups 3 flaws of the SD threshold technique:

1. Hampel (1974) and Nelson (2011) argue how the choice of the threshold the 99.87% quantile was based on a normal distribition and it might fall short when handling other distributions especially heavy tailed distributions.
2. The presence of outliers distorts the value of the mean and SD. These values are

used in the "$3\sigma$ rule" thus the outlier detection test is biased which can lead to false positive (Meij 2017).

3. Webb (1962) and Licata (2013) disagree with the hypothesis that the number of outliers is necessarily small which is translated by the choice of the $\pm 3\sigma$. Licata (2013) proposes the value of $\pm 2\sigma$ or $\pm 2.5\sigma$ to account for this as it increases the number of data points that would be considered outliers 4.6% or 1.2% (respectively) from 0.27% with the $3\sigma$ threshold.

An explanation of the reaming 3 methods can be found in appendix C.

Despite the short comings of the standard deviation technique this method is used in this study to identify large errors due to its simplicity and ease of implementation. However, to account for the $3^{rd}$ limitation (Webb 1962; Licata 2013) discussed previously the value of 2.5 was chosen over the value of 3.

### 3.1.2 Application

The standard deviation threshold technique is used to identify large errors in the gust error data and wind error data.

Given that the mean for gust error $\bar{\varepsilon}_{gust} = -2.09$ and the standard deviation for gust error is $\hat{\sigma}_{\varepsilon_{gust}} = 11.42$, it means that any gust error data point that doesn't belong to the interval in equation (3.2) is labelled as a large error in regards to the standard deviation threshold technique for the gust error data.

$$\begin{aligned} &\left[\bar{\varepsilon}_{gust} - 2.5 \times \hat{\sigma}_{\varepsilon_{gust}}; \bar{\varepsilon}_{gust} + 2.5 \times \hat{\sigma}_{\varepsilon_{gust}}\right] \\ &[-2.09 - 2.5 \times 11.42; -2.09 + 2.5 \times 11.42] \\ &[-30.6; 26.5] \end{aligned} \tag{3.2}$$

There are 825 records in the gust error data of which 35 where flagged by the standard deviation threshold technique as large errors. Thus 4% of the gust error data are

considered outliers.

The storms with records where large errors for gust were flagged are: Dorian with 8, Florence with 5, Harvey with 2, Hermine with 1, Irma with 8, Matthew with 2 and Michael with 9.

Given that the mean for wind error $\bar{\varepsilon}_{wind} = -2.35$ and the standard deviation for wind error is $\hat{\sigma}_{\varepsilon_{wind}} = 7.59$, it means that any wind error data point that doesn't belong to the interval in equation (3.3) is labelled as a large error in regards to the standard deviation threshold technique for the wind error data.

$$\begin{aligned} &\left[\bar{\varepsilon}_{wind} - 2.5 \times \hat{\sigma}_{\varepsilon_{wind}}; \bar{\varepsilon}_{wind} + 2.5 \times \hat{\sigma}_{\varepsilon_{wind}}\right] \\ &[-2.35 - 2.5 \times 7.59; -2.35 + 2.5 \times 7.59] \\ &[-21.3; 16.6] \end{aligned} \tag{3.3}$$

There are 554 records in the wind error data of which 31 where flagged by the standard deviation threshold technique as large errors. Thus 6% of the wind error data are considered outliers.

The storms with records where large errors for the wind data were flagged are the same as the one identified for the gust error data. However, the number of large error per storms are different: Dorian with 3, Florence with 4, Harvey with 2, Hermine with 2, Irma with 15, Matthew with 1 and Michael with 4.

In the following paragraphs the large errors that were singled out are examined to find the source of discrepancy.

## 3.2 Storms where large errors where flagged

### 3.2.1 Large errors in the gust error data

*An example of how the large errors were reviewed is shown with storm Florence.*

Nine large errors were detected in the error data for storm Florence: 4 in the wind error data and 5 in the gust error data where 4 of them have the same location as the wind error data points. The 5 points are present in figures 3.2, 3.3 and 3.4.

The maps in figure 3.2 and 3.3 were created to have a general overview of the error in gust intensity (i.e maximum 3-sec wind speed) estimation for storm Florence. A visual representation of errors is thus given by comparing the observed intensity of hurricane Florence in regards to gust (data from the NHC report) with the estimated intensity with the TSR model for hurricane Florence in regards to gust. Comparing the 2 maps shows that the TSR model over-estimates the peak gust wind speed. This is shown as in figure 3.2 (based on the observed gust) there are more orange points which correspond to peak gust wind speed of TSs level, than in figure 3.3 (based on the estimated gust from the TSR model). This point is futhermore proven by the fact that the 5 large errors have a negative value (cf table 2.2) and given the definition of error in equation (2.1) this implies that the peak gust wind speed generated by the TSR model is on average superior to the observed peak gust wind speed for storm Florence. Consequently, storm Florence has the highest mean in absolute value for gust error -13 (rf table 2.2). Similarly, the value of the highest observed gust in-land recorded around 0200 UTC 14 September at Cape Lookout, North Carolina (CLKN7) is 92 kt and the estimated gust by the TSR model is 105.7 the error is -13.1 (values given in table 3.1) which is consistent with the previous findings.

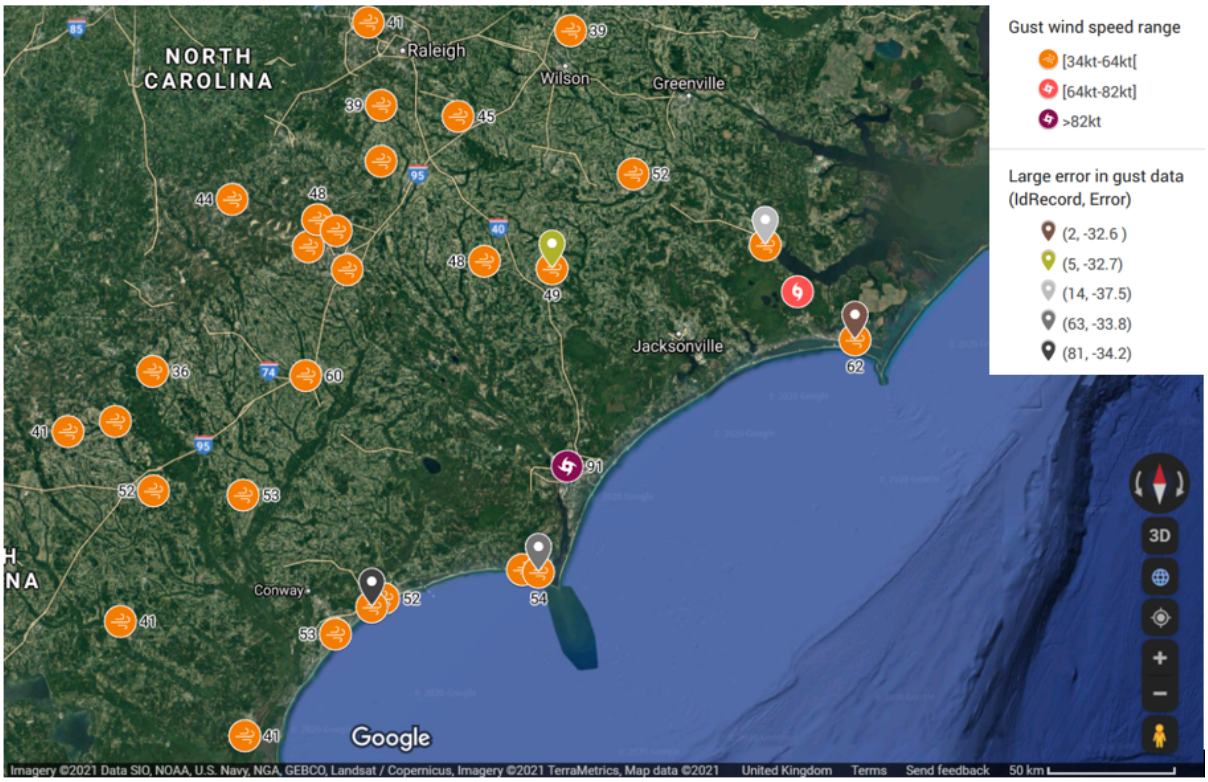



Figure 3.2: Map observed gust for hurricane Florence over land; the locations of the intensities linked to the 5 gust large errors for storm Florence are singled out. When possible the value of the observed gust is given for the locations on the map. IdRecord is the variable that identifies data points here the large error in the gust data records for storm Florence. Error is the value of the error that was flagged as large errors in the gust error database, the 5 points that are singled out. Map created with Google Map.

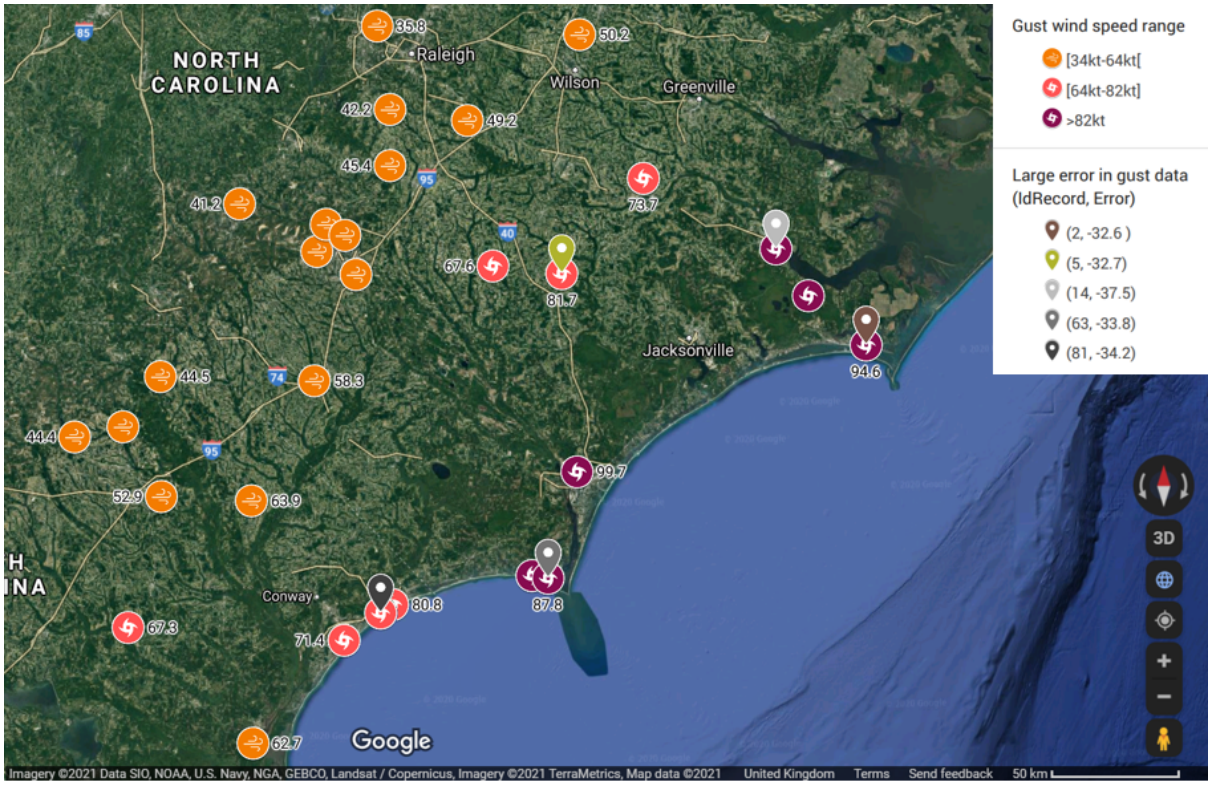



Figure 3.3: Map estimated gust with the TSR model for hurricane Florence over land; the locations of the intensities linked to the 5 gust large errors for storm Florence are singled out. When possible the value of the observed gust is given for the location on the map. IdRecord is the variable that identifies data points here the large error in the gust data records for storm Florence. Error is the value of the error that was flagged as large errors in the gust error database, the 5 points that are singled out. Map created with Google Map.

The following paragraph was extracted from the NHC report: *"The highest land-based sustained wind measured in Florence was a 10-minute average wind of 72 kt recorded around 0200 UTC 14 September at Cape Lookout, North Carolina (CLKN7). Adjusting the 10-minute average wind to a 1-minute average results in an estimated wind speed of 79 kt (Stacy 2019)."*

The 1-min sustained average wind speed produced by the TSR model didn't estimate any values greater than 79 kt in land for storm Florence. Furthermore, the highest land-based estimated sustained wind with the TSR model was 77.1 kt which corresponds to the location where the highest observed land-based sustained wind of 79 kt in Florence was measured. This shows the consistency and relevance of the wind profile generated by the TSR model for storm Florence.


Figure 3.4: Position of the 5 large errors detected in the gust error data for storm Florence, this shows the location, topography of the surrounding area. The legend is made up of the pair IdRecrod to identify the record in storm Florence track and Error the value of the large error that were flagged in gust error data. *This image was created with Google Map and https://www.peko-step.com/en/tool/combine-images.html*

The location where the 5 large error for gust in storm Florence were identified are shown in figure 3.4 and table 3.1 contains the weather station data and error data for gust in storm Florence. Figure 3.4 shows that the 5 large errors all have different layout. However, the topography that stands out the most is the location associated with the value of IdRecord equals 14 (point C). Indeed, due to its position this data point is at

the intersection of 2 types of land surface vegetation and plain ground. This is important as the surface friction (i.e roughness) is higher with vegetation and lower on plain ground. A storm passing through a surface with high roughness will have a lower wind speed associated with it, if the same storm passed through an area with lower surface friction it will then have a higher wind speed. Thus for the data point with IdRecord equals 14 (point C) there is suspicion that the large error is due to spatial uncertainty. The Tropical Storm Risk (TSR) model has a spatial precision of 4 digits meanwhile the observed data obtained from the National Hurricane Centre (NHC) has a precision of 2 digits. Thus it is possible that the TSR model associated the IdRecord equals 14 (point C) to be on the plain ground land surface rather than the vegetation land surface which explain the high gust wind speed of 85.5kt rather than a value close to the observed gust of 48kt. This theory is furthermore justified as the TSR model uses a constant roughness factor overland that captures the surface friction for short grass. Thus, the model does not incorporate the variation in land surface.

For the other large error data points a look at table 3.1 points to the option that the TSR model didn't capture the right rate of decay of the gust wind speed and the TSR model overestimated the value of the 3-sec peak gust wind speed.

Only point C in figure 3.4 is not present as a location associated with large error in the wind dataset for Storm Florence. It was found that the same conclusion for the 4 other data points also stand true in regards to the large errors for the 1-min sustained wind for storm Florence (cf table 3.2).

This example showed the main sources of large errors. While looking at the error data it shows that the large errors are either linked to spatial uncertainty, underestimation or overestimation of the storm intensity by the TSR model.

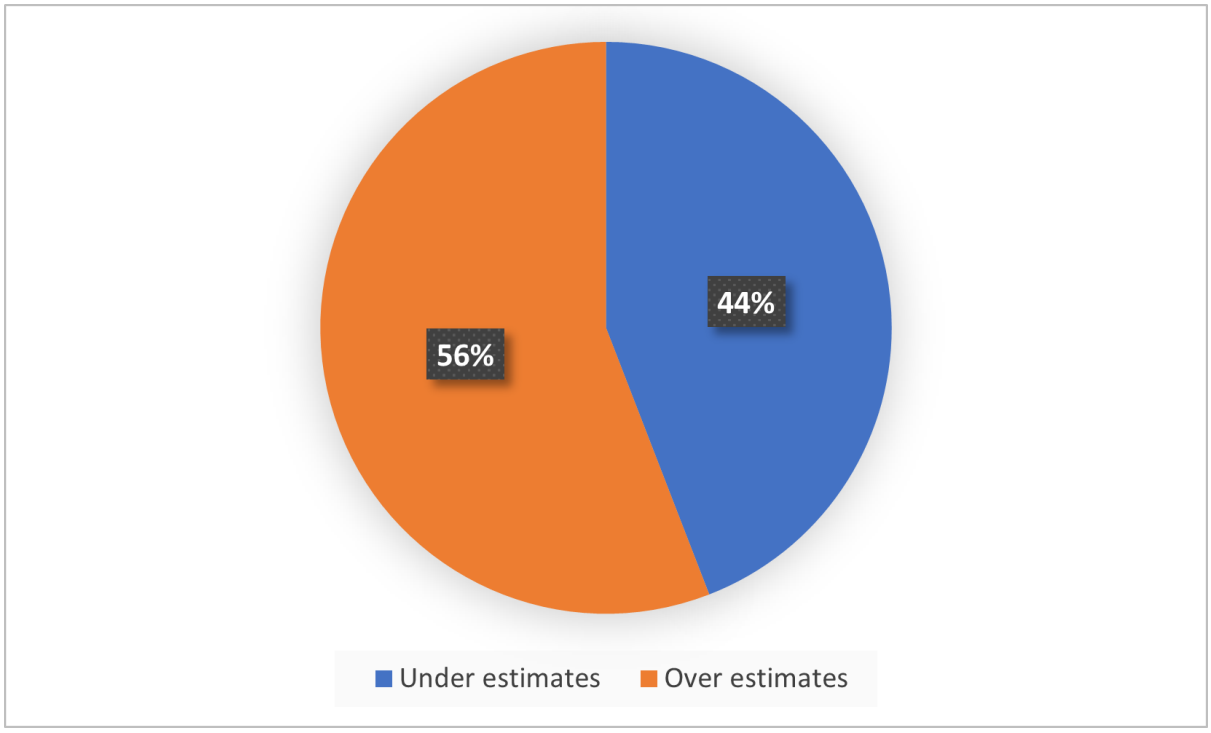



Figure 3.5: Pie chart showing the reparation of the source of large errors for gust

Figure 3.5 shows that the TSR model over estimate 56% of the gust wind speed for large errors. Indeed, all large errors from storm Matthew and Harvey were overestimated by the TSR model, 63% for Irma, 33% for Michael which the remain set of large errors where underestimated including that from storm Hermine. For storm Dorian the TSR model underestimated 50% of the gust wind intensity in the large error dataset. For more details on the large error data for gust see table 3.1.

The same kind of analysis is done for the wind data for large errors. Only the results are discussed in the following section.

| Name | Year | IdRecord | Long | Lat | TSR | Obs | Error | Spatial? | overEstimates? |
|---|---|---|---|---|---|---|---|---|---|
| HERMINE | 2016 | 119 | -75.9 | 35.67 | 33 | 64 | 31 | 0 | 0 |
| MATTHEW | 2016 | 2 | -78.99 | 26.7 | 140.3 | 91 | -49.3 | 0 | 1 |
| MATTHEW | 2016 | 122 | -79.32 | 33.31 | 73.7 | 41 | -32.7 | 0 | 1 |

Continued on next page

**Table 3.1 – continued from previous page**

| Name | Year | IdRecord | Long | Lat | TSR | Obs | Error | Spatial? | overEstimates? |
|---|---|---|---|---|---|---|---|---|---|
| HARVEY | 2017 | 38 | -97.04 | 28.09 | 137.3 | 60 | -77.3 | 0 | 1 |
| HARVEY | 2017 | 45 | -96.68 | 28.65 | 73.8 | 43 | -30.8 | 0 | 1 |
| IRMA | 2017 | 14 | -78.75 | 22.15 | 97.4 | 138 | 40.6 | 0 | 0 |
| IRMA | 2017 | 17 | -78.45 | 22.52 | 165.6 | 105 | -60.6 | 0 | 1 |
| IRMA | 2017 | 20 | -78.12 | 21.85 | 90.8 | 130 | 39.2 | 0 | 0 |
| IRMA | 2017 | 63 | -81.77 | 26.15 | 103.1 | 71 | -32.1 | 0 | 1 |
| IRMA | 2017 | 131 | -81.6 | 26.03 | 125.1 | 86 | -39.1 | 0 | 1 |
| IRMA | 2017 | 181 | -81.3 | 26.17 | 104 | 65 | -39 | 0 | 1 |
| IRMA | 2017 | 218 | -80.73 | 25.92 | 84.7 | 112 | 27.3 | 0 | 0 |
| IRMA | 2017 | 227 | -81.31 | 26.94 | 92.1 | 59 | -33.1 | 0 | 1 |
| FLORENCE | 2018 | 2 | -76.65 | 34.72 | 94.6 | 62 | -32.6 | 0 | 1 |
| FLORENCE | 2018 | 5 | -77.98 | 34.99 | 81.7 | 49 | -32.7 | 0 | 1 |
| FLORENCE | 2018 | 14 | -77.04 | 35.07 | 85.5 | 48 | -37.5 | 1 | 1 |
| FLORENCE | 2018 | 63 | -78.05 | 33.9 | 87.8 | 54 | -33.8 | 0 | 1 |
| FLORENCE | 2018 | 81 | -78.77 | 33.78 | 80.2 | 46 | -34.2 | 0 | 1 |
| MICHAEL | 2018 | 92 | -85.17 | 30.85 | 107 | 68 | -39 | 0 | 1 |
| MICHAEL | 2018 | 93 | -84.6 | 30.55 | 82.4 | 48 | -34.4 | 0 | 1 |

Continued on next page

**Table 3.1 – continued from previous page**

| Name | Year | IdRecord | Long | Lat | TSR | Obs | Error | Spatial? | overEstimates? |
|---|---|---|---|---|---|---|---|---|---|
| MICHAEL | 2018 | 103 | -84.63 | 30.97 | 103 | 43 | -60 | 0 | 1 |
| MICHAEL | 2018 | 258 | -76.65 | 34.72 | 27.1 | 54 | 26.9 | 0 | 0 |
| MICHAEL | 2018 | 261 | -76.9 | 34.9 | 28 | 58 | 30 | 0 | 0 |
| MICHAEL | 2018 | 332 | -76.35 | 37.08 | 25.5 | 54 | 28.5 | 0 | 0 |
| MICHAEL | 2018 | 336 | -76.3 | 36.93 | 25.6 | 65 | 39.4 | 0 | 0 |
| MICHAEL | 2018 | 337 | -76.03 | 36.82 | 25.2 | 55 | 29.8 | 0 | 0 |
| MICHAEL | 2018 | 364 | -76.32 | 37.11 | 28.3 | 57 | 28.7 | 0 | 0 |
| DORIAN | 2019 | 56 | -79.32 | 33.31 | 79.8 | 48 | -31.8 | 0 | 1 |
| DORIAN | 2019 | 83 | -77.43 | 34.71 | 87.1 | 54 | -33.1 | 1 | 1 |
| DORIAN | 2019 | 85 | -77.92 | 34.28 | 81.1 | 49 | -32.1 | 0 | 1 |
| DORIAN | 2019 | 106 | -77.97 | 33.85 | 85.4 | 49 | -36.4 | 0 | 1 |
| DORIAN | 2019 | 130 | -63.55 | 44.59 | 36.3 | 65 | 28.7 | 0 | 0 |
| DORIAN | 2019 | 135 | -63.84 | 46.44 | 25.2 | 63 | 37.8 | 0 | 0 |
| DORIAN | 2019 | 137 | -63.12 | 46.29 | 25.7 | 55 | 29.3 | 0 | 0 |
| DORIAN | 2019 | 138 | -64.68 | 46.11 | 26.8 | 54 | 27.2 | 0 | 0 |

Continued on next page

**Table 3.1 – continued from previous page**

| Name | Year | IdRecord | Long | Lat | TSR | Obs | Error | Spatial? | overEstimates? |
|---|---|---|---|---|---|---|---|---|---|

Table 3.1: Large error in regards to the gust weather station datawhere IdRecord rank of the row as it appeared in the dataset of the concerning storm in Name, Lat is the Latitude, Long the Longitude, TSR estimated value by the TSR model, Obs the observed value extracted from NHC report, Spatial? if the error is due spatial uncertainty, overEstimates? if the TSR model over estimates the intensity at the location of the large error

### 3.2.2 Large errors in the wind error data

After reviewing the large error for wind as it was done in section 3.2.1, the results are presented below.

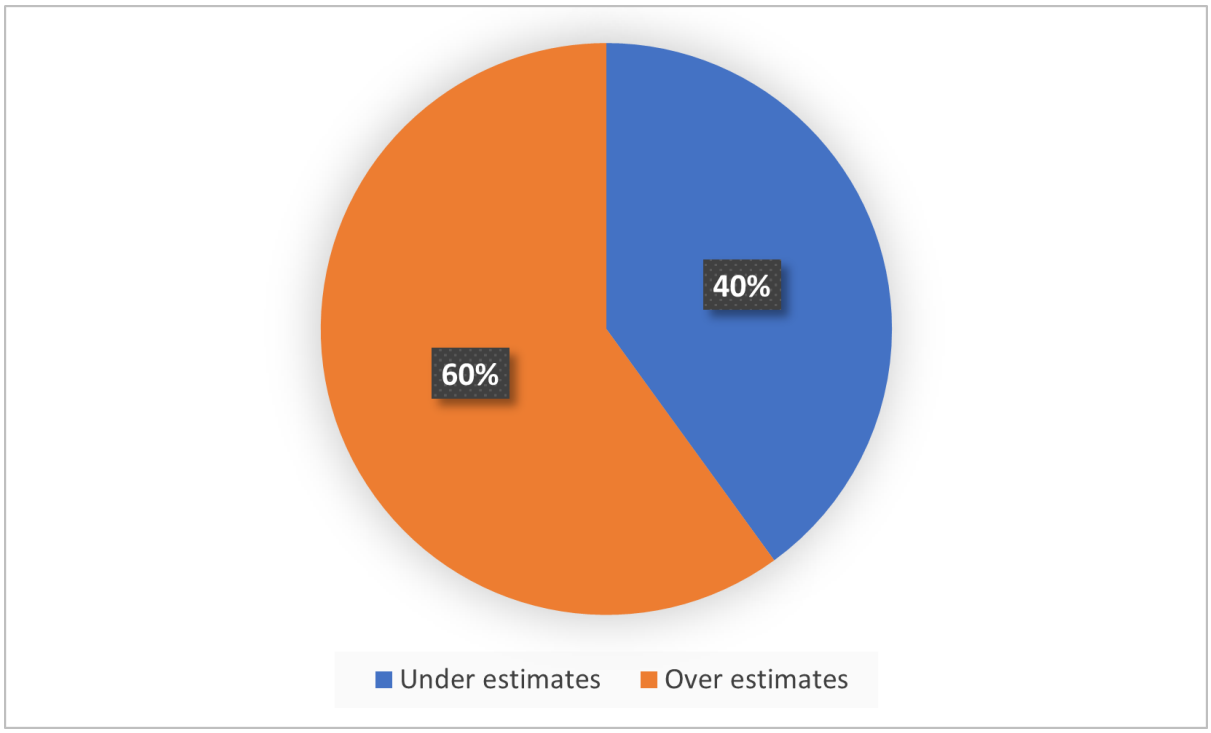



Figure 3.6: Pie chart showing the reparation of the source of large errors for wind

The reparation of the source of large errors for wind is given in figure 3.6 and table 3.2 contain all the large error for wind. 100% of the source of large error for storms Matthew, Harvey, Dorian and Florence comes from the TSR model overestimating the 1-min sustained wind at the location of these large errors. For storm Hermine respectively storm Michael 50% respectively 75% are due to the TSR model overestimating the values at

the location of these large errors and then underestimating for the remaining source of error. The same analysis is done for storm Irma 60% of the large error for wind are due to the TSR model underestimates the value of the 1-min sustained wind at those locations.

| Name | Year | IdRecord | Long | Lat | TSR | Obs | Error | Spatial? | overEstimates? |
|---|---|---|---|---|---|---|---|---|---|
| HERMINE | 2016 | 43 | -82.7 | 27.74 | 27.6 | 46 | 18.4 | 0 | 0 |
| HERMINE | 2016 | 86 | -80.7 | 32.49 | 46.7 | 25 | -21.7 | 0 | 1 |
| MATTHEW | 2016 | 109 | -79.32 | 33.31 | 49.5 | 27 | -22.5 | 0 | 1 |
| HARVEY | 2017 | 21 | -97.04 | 28.09 | 94.9 | 44 | -50.9 | 0 | 1 |
| HARVEY | 2017 | 129 | -97.53 | 29.03 | 43.8 | 21 | -22.8 | 0 | 1 |
| IRMA | 2017 | 2 | -78.79 | 22.03 | 60.6 | 36 | -24.6 | 0 | 1 |
| IRMA | 2017 | 9 | -78.75 | 22.15 | 66.2 | 108 | 41.8 | 0 | 0 |
| IRMA | 2017 | 12 | -78.45 | 22.52 | 115.6 | 83 | -32.6 | 0 | 1 |
| IRMA | 2017 | 22 | -81.52 | 23.02 | 52.6 | 30 | -22.6 | 0 | 1 |
| IRMA | 2017 | 50 | -81.69 | 30.49 | 33.5 | 51 | 17.5 | 0 | 0 |
| IRMA | 2017 | 54 | -81.42 | 30.39 | 33.1 | 59 | 25.9 | 0 | 0 |
| IRMA | 2017 | 112 | -81.48 | 30.39 | 33.2 | 55 | 21.8 | 0 | 0 |
| IRMA | 2017 | 125 | -81.6 | 26.03 | 86 | 62 | -24 | 0 | 1 |
| IRMA | 2017 | 175 | -81.3 | 26.17 | 70.9 | 31 | -39.9 | 0 | 1 |
| IRMA | 2017 | 205 | -81.05 | 29.19 | 38.8 | 61 | 22.2 | 0 | 0 |

Continued on next page

**Table 3.2 – continued from previous page**

| Name | Year | IdRecord | Long | Lat | TSR | Obs | Error | Spatial? | overEstimates? |
|---|---|---|---|---|---|---|---|---|---|
| IRMA | 2017 | 212 | -80.73 | 25.92 | 57.2 | 97 | 39.8 | 0 | 0 |
| IRMA | 2017 | 218 | -80.22 | 26.04 | 48.6 | 70 | 21.4 | 0 | 0 |
| IRMA | 2017 | 221 | -81.31 | 26.94 | 62.4 | 36 | -26.4 | 0 | 1 |
| IRMA | 2017 | 240 | -81.41 | 31.05 | 31.6 | 50 | 18.4 | 1 | 0 |
| IRMA | 2017 | 251 | -80.59 | 32.34 | 29.1 | 49 | 19.9 | 0 | 0 |
| FLORENCE | 2018 | 1 | -76.65 | 34.72 | 64.2 | 35 | -29.2 | 0 | 1 |
| FLORENCE | 2018 | 4 | -77.98 | 34.99 | 55.1 | 33 | -22.1 | 0 | 1 |
| FLORENCE | 2018 | 13 | -77.04 | 35.07 | 57.8 | 32 | -25.8 | 0 | 1 |
| FLORENCE | 2018 | 79 | -78.77 | 33.78 | 54.1 | 25 | -29.1 | 0 | 1 |
| MICHAEL | 2018 | 2 | -84.95 | 21.87 | 66.4 | 84 | 17.6 | 0 | 0 |
| MICHAEL | 2018 | 64 | -85.17 | 30.85 | 73.1 | 49 | -24.1 | 0 | 1 |
| MICHAEL | 2018 | 65 | -84.6 | 30.55 | 55.6 | 33 | -22.6 | 0 | 1 |
| MICHAEL | 2018 | 74 | -84.63 | 30.97 | 70.2 | 30 | -40.2 | 0 | 1 |
| DORIAN | 2019 | 58 | -79.78 | 32.9 | 52.8 | 31 | -21.8 | 0 | 1 |
| DORIAN | 2019 | 83 | -77.43 | 34.71 | 64.4 | 37 | -27.4 | 1 | 1 |
| DORIAN | 2019 | 106 | -77.97 | 33.85 | 57.7 | 36 | -21.7 | 0 | 1 |

Continued on next page

**Table 3.2 – continued from previous page**

| Name | Year | IdRecord | Long | Lat | TSR | Obs | Error | Spatial? | overEstimates? |
|---|---|---|---|---|---|---|---|---|---|

Table 3.2: Large error in regards to the wind weather station data, where IdRecord rank of the row as it appeared in the dataset of the concerning storm in Name, Lat is the Latitude, Long the Longitude, TSR estimated value by the TSR model, Obs the observed value extracted from NHC report, Spatial? if the error is due spatial uncertainty, overEstimates? if the TSR model over estimates the intensity at the location of the large error

In this section and the previous one the source of large errors where identified. However the causes that led to the model overestimating or underestimating the storm intensity was not found. Nevertheless, reviewing the error data showed that the TSR model tends to underestimate the storm's intensity inland while it overestimate it near the coastal regions; comparing the maps 3.7 and 3.8 shows this phenomena. As well spatial uncertainty is an error inherent to the model that leads to either overestimation or underestimation of the storm intensity.

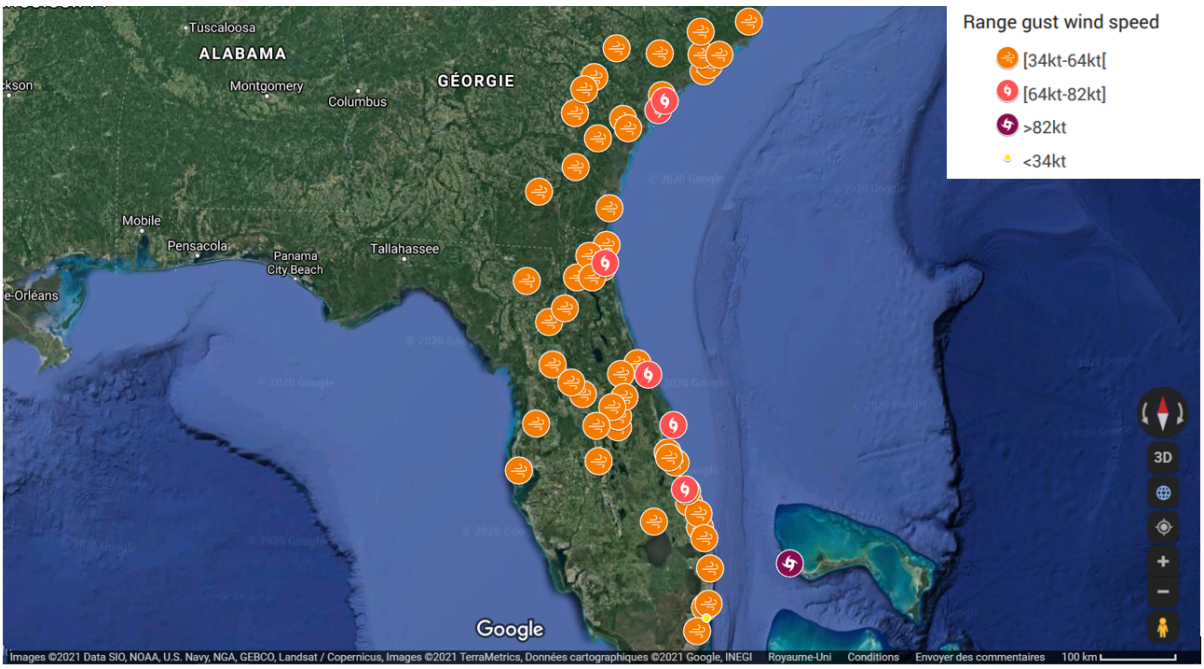



Figure 3.7: Map estimated gust with the TSR model for hurricane Matthew over land. Map created with Google Map.

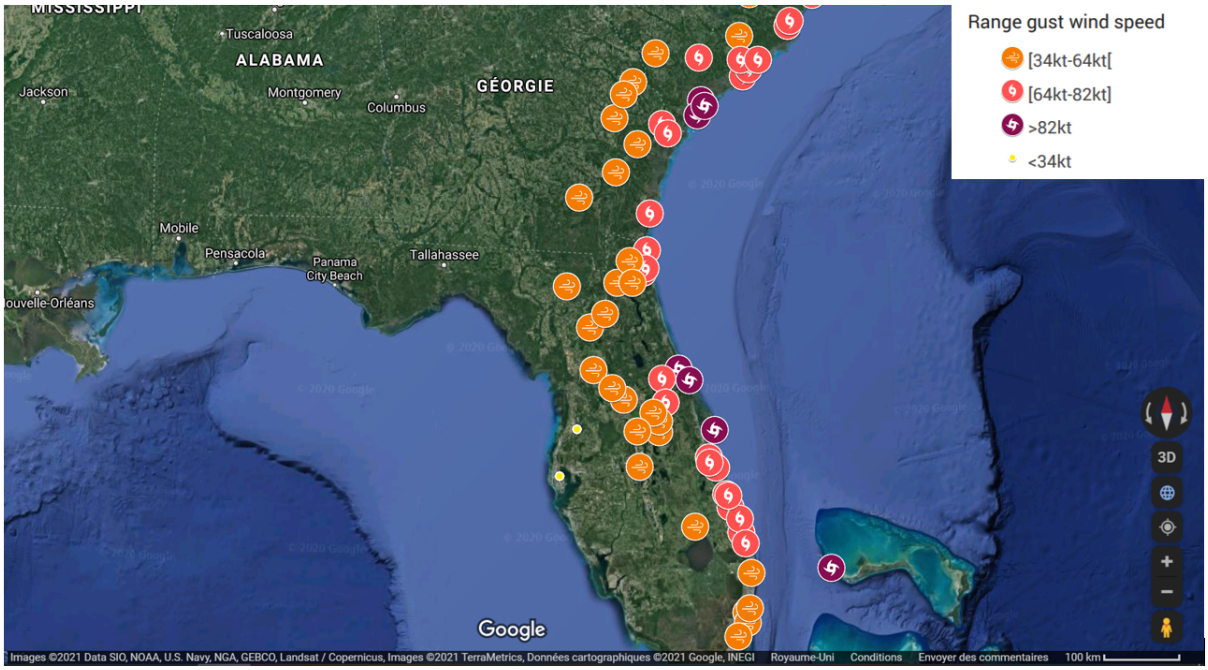



Figure 3.8: Map estimated gust with the TSR model for hurricane Matthew over land. Map created with Google Map.

No large error was deleted before calculating the accuracy of the TSR model done in the following chapter.

# Chapter 4

# Accuracy

The accuracy of the TSR model evaluates how close the data points from the outputs of the model are to the observed values. In this study the outputs of interest from the model are the gust and wind data. Thus, given the same location ( longitude and latitude points), are the values produced by the TSR model, a good estimation of the observed gust and wind data obtained from the NHC reports?

This chapter provides an answer to this question by evaluating the difference between the observed data and the TSR data for storms that made landfall in the US from 2015 to 2019.
Then, the results obtained are compared to the previous TSR Business study of measure of accuracy of the TSR model that was done for storms that made landfall in the US from 2004-2008. As well as with the level of accuracy obtained from NHC intensity error database.

The last part of this chapter is on the importance of understanding the catastrophe model used to correctly evaluate the insured loss obtained.

## 4.1 Accuracy of the TSR model

### 4.1.1 Defining and calculating the accuracy of the TSR model

#### Explaining the notion of accuracy

Under the assumption that the probability distribution of the observations (i.e variable

$X$), is the same as the probability distribution of the values obtained with the model (i.e variable $Y$), the probability distribution of error (i.e variable $Z = X - Y$), should be randomly distributed around 0 such that the mean of error equals 0.
Thanks to the central limit theorem this hypothesis is often made: the distribution of error follows a normal distribution regardless of the distribution of the observations. Given the probability distribution of error $Z$, the range in which the observed error is expected to be in, for a given probability $1 - p$, is then calculated as

$$\left[\mu_Z - \sigma_Z \times q_{\frac{p}{2}}; \mu_Z + \sigma_Z \times q_{\left(1-\frac{p}{2}\right)}\right] \tag{4.1}$$

, where $\mu_Z$ is the mean of Z and $\sigma_Z$ the standard deviation of $Z$, $q_{\frac{p}{2}}$ the quantile of probability $\frac{p}{2}$ of the error distribution. The obtained range is the accuracy of the model for a probability p.

In practice a probability framework is not always used to measure the accuracy of a model. The range of error can also be obtained by using the empirical values of the mean and standard deviation of error. The downside of this approach is that the probability associated with the calculated interval is unknown. There is no notion of confidence level of the obtained range.

Several measure of errors (accuracy) exist and some are presented below (Hsu 1999) :

Standard error:

$$\sqrt{v(Z)} \tag{4.2}$$

Root Mean Square Error (RMSE):

$$\sqrt{m(Z^2)} \tag{4.3}$$

Mean Absolute Error (MAE):

$$m(|Z|) \tag{4.4}$$

$m$ is the function that computes the mean.

$v$ is the function that calculates the variance.

However, in this study the accuracy is measured in term of the Absolute Percentage Error (APE) and is defined shortly afterwards in the following paragraph. The accuracy is obtained by comparing the observations and TSR model outputs for storms that made landfall in the US from 2015 to 2019.

The variable of interest hereinafter is the Absolute Percentage Error (APE) :

$$APE_i = \left| \frac{x_{TSR_i} - x_{obs_i}}{x_{obs_i}} \right| \times 100 \tag{4.5}$$

, where i is the i-ième data point from either the gust or wind dataset.

### Accuracy of the TSR model

The TSR Business, measures the accuracy of the TSR model with the mean APE ($\mu_{APE}$) +/- standard deviation (SD) APE ($\sigma_{APE}$). The higher the value of $\mu_{APE}$ +/- $\sigma_{APE}$ the lower the accuracy and the lower the value of $\mu_{APE}$ +/- $\sigma_{APE}$ the higher the accuracy:

$$\begin{aligned} \nearrow \mu_{APE} \pm \sigma_{APE} &\Longrightarrow \searrow Accuracy \\ \searrow \mu_{APE} \pm \sigma_{APE} &\Longrightarrow \nearrow Accuracy \end{aligned} \tag{4.6}$$

Thus, the same calculation that were done in subsection 2.2.2.1 will be done again, but for the variable APE. The mean and standard deviation of APE are calculated for each storm in the error database. Then the overall mean and standard deviation of APE is calculated for the two datasets in table 4.1.

| Data | Mean APE (%) ± SD APE (%) |
|---|---|
| Gust | 20.4 ± 16.9 |
| Wind | 21.4 ± 21.9 |

Table 4.1: Accuracy TSR model MeanAPE ± SD

### 4.1.2 Accuracy of the TSR model for different range of gust and wind speed

#### 4.1.2.1 Different gust range for accuracy analysis

The characteristics of a storm varies considerably depending on it's wind speed. For example, the Saffir-Simpson Hurricane Wind Scale categorise storms based only on the storm's wind speed. Thus, it is important to have a measure of the accuracy of the TSR model given different range of gust and wind values.

This section presents the results of the accuracy of the TSR model done with the gust and wind error datasets.

The global accuracy of the TSR model for gust : 20.4 ± 16.9% was found in section 4.1.1 and presented in table 4.1. This value is slightly different from the accuracy for gust ≥ 40kts (table 4.2) : 20.5 ± 16.2%. In one case the gust mean APE is higher than the other. Similarly, the standard deviation for gust ≥ 40kts is higher than the other one. This shows a trade off between the mean (bias) and the standard deviation (variance). The gust range with the highest accuracy is for gust > 64kts as it has the lowest mean APE and lowest standard deviation APE. However, the gust range with the lowest accuracy isn't as straight forward but it is between the two gust range [40-50] and [50-64] with accuracy 23.0 ± 17.8% and 22.4 ± 21.8% respectively. On one hand, the mean APE is higher than the other, then this trend is reversed when looking at the SD APE.

| Gust Range (kt) | NB samples | Mean ± SD (%) |
|---|---|---|
| 40-50 | 292 | 23.0 ± 17.8 |
| 50-64 | 142 | 22.4 ± 21.8 |
| >64 | 98 | 19.1 ± 15.2 |
| ≥ 40 | 532 | 20.5 ± 16.2 |

Table 4.2: Accuracy TSR model for different gust range

#### 4.1.2.2 Different wind range for accuracy analysis

The global accuracy of the TSR model for wind 21.4% ± 21.9% was found in section 4.1.1 and presented in table 4.1. Contrary, to what was observed in sub section 4.1.2.1 there is no debate on the improvement of the accuracy of the TSR model for wind when the wind speed range is ≥ 34 mph : 17.6% ± 16.1%. Indeed, for all wind range present in table 4.1.2.1 the mean wind and the standard deviation of the APE are less than the one found for wind in table 4.1.

| Wind Range (mph) | NB samples | Mean (%) ± SD (%) |
|---|---|---|
| 40-50 | 201 | 23.4 ± 19.8 |
| 50-64 | 39 | 16.5 ± 8.9 |
| >64 | 14 | 14.4 ± 8.9 |
| ≥ 34 | 254 | 17.6 ± 16.1 |

Table 4.3: Accuracy TSR model for different wind range

Table 4.3 indicates that the wind speed range with the best accuracy is for wind speed >64 mph : 14.4% ± 8.9%.

Indeed, for wind speed >64 mph the accuracy of the TSR model is higher than the accuracy associated with $[40mph; 50mph]$ : 23.4% ± 19.8% which is also the wind speed range with the lowest accuracy.

A general observation is that the highest accuracy for gust and wind are both associated with the highest intensity range.

In the next section the results obtained in this study will be compared with the previous TSR model accuracy study done by TSR Business.

## 4.2 Reviewing the TSR model's accuracy

After evaluating the accuracy of the TSR model it is of interest to have a relative view of the evolution of the performance of the TSR model over time and in regards to other models used to approximate the intensity of storms. This will be done in the following

paragraphs.

### 4.2.1 TSR model accuracy for storms that made landfall in the US: 2004-2008 vs 2015-2019

#### 4.2.1.1 Gust: TSR model accuracy 2004-2008 vs 2015-2019

The results of the study done by TSR Business for gust is shown in table 4.4 and the results of this current study is shown in table 4.5.

| Gust Range (kt) | NB samples | Mean ± SD (%) |
|---|---|---|
| 34-63 | 606 | 18.2 ± 16.9 |
| 64-82 | 133 | 16.6 ± 14.7 |
| 83-95 | 41 | 15.8 ± 10.9 |
| >95 | 36 | 16.1 ± 10.2 |
| 26-184 | 842 | 17.8 ± 16.0 |

Table 4.4: Accuracy of the TSR Gust History for storms that made landfall in the US : 2004-2008 measured with the absolute percentage error, Source : TSR Business (2020)

Overall, the values obtained in the previous study for gust are not too far off from the values calculated in this study but still out perform the ones obtained in this study. Indeed, for gust ranging from $[26kt; 184kt]$ the study done by the TSR Business estimated an accuracy of $17.8 \pm 16.0\%$ meanwhile in this study the accuracy was estimated at 19.6% ± 14.6%. Here again there is this trade off bias variance.
All means gust APE in table 4.5 from this study are higher than the highest mean in table 4.4 from the TSR Business study expect from the gust range of >95kt at 14.8%.

| Gust Range (kt) | NB samples | Mean ± SD (%) |
|---|---|---|
| 34-63 | 662 | 19.6 ± 15.8 |
| 64-82 | 78 | 22.6 ± 18.3 |
| 83-95 | 13 | 19.6 ± 8.9 |
| >95 | 13 | 14.8 ± 16.1 |
| 26-184 | 818 | 19.6 ± 14.6 |

Table 4.5: Accuracy of the TSR Gust History for storms that made landfall in the US : 2015-2019 measured with the absolute percentage error

#### 4.2.1.2 Wind: TSR model accuracy 2004-2008 vs 2015-2019

The results of the study done by TSR Business for wind is shown in table 4.6 and the results of this study for wind is shown in table 4.7.

| Wind Range (mph) | NB samples | Mean ± SD (%) |
|---|---|---|
| 34-63 | 490 | 23.2 ± 21.6 |
| 64-82 | 58 | 16.3 ± 13.3 |
| >82 | 19 | 18.3 ± 11.2 |
| 34-135 | 567 | 22.3 ± 20.7 |

Table 4.6: Accuracy of the TSR Wind History for storms that made landfall in the US : 2004-2008 measured with the absolute percentage error, Source : TSR Business (2020)

Contrary, to what was discussed in sub section 4.2.1.1, the accuracy estimated in this study out performs that estimated in the TSR Business study for the 1-min sustained wind. The accuracy of the TSR wind model from this study for the wind speed range $[36mph; 135mph]$ is found to be 17.6% ± 14.4% and the value of 22.3 ± 20.7% was obtained in the TSR Business study. This same comparison can be done with all the different wind range and it will give the same result.

| Wind Range (mph) | NB samples | Mean ± SD (%) |
|---|---|---|
| 34-63 | 240 | 17.7 ± 16.3 |
| 64-82 | 5 | 7.7 ± 11.3 |
| >82 | 9 | 17.0 ± 7.7 |
| 34-135 | 254 | 17.6 ± 14.4 |

Table 4.7: Accuracy of the TSR Wind History for storms that made landfall in the US : 2015-2019 measured with the absolute percentage error

A general overview of the points discussed above is that the values calculated in the TSR Business study isn't too far off from those found in this study. Over the years the accuracy of the TSR model for gust seems to have lagged a little yet the accuracy of the TSR model for wind has improved.

Nevertheless, this result is limited as the sample size for the different range of gust and the 1-min sustained wind speed differ from one study to another. For example comparing the accuracy for the wind range [64mph-82mph] of this study with only 5 observations in

the sample (see table 4.7) to that of the previous study with 58 observations in the sample (see table 4.6) isn't statistically sound. Indeed, the more the sample's size increases the more the observed mean APE is closer to the real mean APE likewise when the number of observations in the sample is small it gives unstable results for the observed mean APE. Thus, in the following section the accuracy of the TSR model is compare with that of the NHC's model with the same sample size.

The values that are kept to measure the accuracy of the TSR model obtained in this study in regards to APE is 19.6 ± 14.6 for gust and 17.6 ± 14.4 % for wind (see table 4.8).

| | Range | APE Mean ± SD (%) |
|---|---|---|
| Gust | $[26kt; 184kt]$ | 19.6 ± 14.6 |
| Wind | $[36mph; 135mph]$ | 17.6 ± 14.4 |

Table 4.8: Accuracy of the TSR model obtained in this study

In the following section the accuracy of the TSR model is calculated with other accuracy measure to be able to compare it to the accuracy of other wind model.

### 4.2.2 Comparing accuracy: NHC model vs TSR model

The web site of the NHC contains an error database of storm intensity from 1989-2019. The error data here is the difference between the observed 1-min average sustained wind of storms from the NHC reports which has already been used in this study and the 1-min average sustained wind forecasted by the NHC.

In the following paragraph the error from the NHC is compared to that of the TSR model. In order to do so the rows containing the information of interest needs to be extracted from the NHC error database.

In NHC error database, the name of the storms are not given instead each storm in the database has a unique identifier that makes it possible to find the name of the storm. Indeed the same identifier was also used in the NHC report as well as the the name of

the storm (see NHC reports in the bibliography).

*Storm IRMA 2017 was not present in the NHC error database.*

After finding the name of the storms in NHC error database the rows having the same location as those in the TSR error database are kept because not all the rows in NHC error database will be used. Recall, in chapter 2 several rows were dropped for homogeneity of the data and in the TSR error database only the locations in-land were kept for the accuracy analysis. These are the rows of interest that need to be extracted from the NHC error database to compare to the TSR error database. The software ACCESS was used to join the two database according to the criterias specified above.

In view of how the NHC error database is constructed the choice was made to use the Mean Absolute Error shown in equation (4.4) as measure of accuracy to compare the performance of the model used by the NHC to that of the TSR model. Following on from what was previously explained in section 2.2.2.1 the MAE is not calculated as the global mean of the absolute value of errors but as a weighted average instead. First of all, the individual MAE for each storm present in the database obtained with ACCESS is calculated then the overall MAE is the average of the individual MAE of each storm. The results for the MAE of the NHC model and the TSR model are presented in table 4.9. The results show that the accuracy of the TSR model is as good as the accuracy of the NHC model and even performs slightly better with a MAE of 6.61 compared to 6.65.

| Name | Nb Record | MAE NHC | MAE TSR |
|---|---|---|---|
| BILL | 2 | 2,50 | 8,10 |
| HERMINE | 2 | 7,50 | 7,30 |
| JULIA | 2 | 5,00 | 2,00 |
| MATTHEW | 23 | 4,78 | 8,62 |
| HARVEY | 32 | 4,38 | 8,26 |
| NATE | 3 | 15,00 | 5,23 |
| ALBERTO | 6 | 6,67 | 3,58 |
| FLORENCE | 6 | 7,50 | 12,17 |
| GORDON | 1 | 10,00 | 2,10 |
| MICHAEL | 6 | 7,50 | 7,25 |
| BARRY | 1 | 0,00 | 0,20 |
| DORIAN | 5 | 9,00 | 14,54 |
| Total/Mean | 89 | 6,65 | 6,61 |

Table 4.9: Table MAE of the NHC model and the TSR model

*In the previous sections, different measure of uncertainty for the intensity estimated by the TSR model were given. However, due to no access to data (exposure data, insurance portfolio ) as it is client sensitive information for the insurance/reinsurance companies the impact on the insured loss evaluated with the use of the TSR model is not calculated in this study. Nevertheless, in the following paragraphs, a description on evaluating the output (focus on insured loss) of a catastrophe model is given.*

## 4.3 Evaluating the insured loss from a catastrophe model

Fulcher (2006) states "There is no such thing as an exact modelled loss estimate, as there is always an element of subjectivity involved." For that reason the reinsurer broker company Guy Carpentera proposed a quantitative analysis (LI 2013) to help practitioners in their decision making using a customize view of catastrophe risk.

### 4.3.1 General guidelines

The basis of the quantitative analysis in this context is to get a sense of how well the catastrophe model reproduces known insured loss from the insurance/reinsurance portfolio. Though, events that happened in the past are not a safeguard of the events that will occur in the future, they give valuable information on the performance and adequacy of a model for an insurance/reinsurance portfolio. Thus, it is not uncommon for an insurer/reinsurer company to use catastrophe models from different vendors at the same time.

Indeed, if the insurer/reinsurer portfolio covers an area with various types of surface friction having important discrepancies among them it will not be the best course of action to use the TSR model in the hazard component of a catastrophe model as the TSR model has only one type of surface friction modelled (see error analysis chapter 3 ). This will affect the estimated insured loss and generate unsatisfying results. However, in the case were the insurance/reinsurance portfolio covers an area that is flat and with constant surface friction, the TSR model could be a good match and this can be verified by comparing the insured loss obtained by incorporating the TSR model in a catastro-

phe model to the real insured loss from a portfolio of a known catastrophic event. This example shows that it is also important to know the characteristics of the data used in the input of the catastrophe model for the insurance/reinsurance portfolio covered.

It is important to asses before any calculation is done the scope of the analysis: What is covered? What structures are included/excluded? Where is the area covered? What are the geographic specificities of the area covered? These questions help in selecting the right catastrophe model for the situation. For example, the TSR was calibrated for hurricanes in the North Atlantic basin so it will be inappropriate to use the TSR model for an insurance/reinsurance portfolio with the area covered in the Pacific basin.

Furthermore, quality checking the data helps to prevent these type of mishaps, as it brings about the right questions to be asked and a better understanding of the data, which in turn gives the set of criterias that the catastrophe models needs to have in order to model the risk at hand for the underlying insurance/reinsurance portfolio. Likewise, Fulcher (2006) also points out the importance of quality checking the data as a catastrophe model is like all models, the better the quality of the data in the input of a model the better the data in the output of the model will be.

### 4.3.2 Example

The previous sections and chapters provided an in-depth explanation on how to evaluate the output of a model in term of error, the same process can be used to evaluate the insured loss obtained with a catastrophe model. Given the error associated with the insured loss it is then the responsibility of the actuary, risk manager to set thresholds (in regards to their risk profile, risk appetite) on the value of the error obtained to classify the insured loss estimated given a catastrophe model.

The following paragraph is an example of evaluating the insured loss obtained from a catastrophe model by using a classification of error system.

Basis for the example: assuming that the insured loss of a catastrophic event is known and a catastrophe model has estimated the insured loss for the known event. To evaluate the result of the model, the following is done in this case:

1. Set a range of acceptability that is in accordance with the objectives (or risk profile). Determine what range the values of the estimates should be in to be considered acceptable or not. Here the estimated insured losses shouldn't be more than or less than twice the value of the observed losses.

2. Calculate the absolute value of the difference between the estimated insured loss and the observed insured loss.

3. If the value of the absolute value of the difference is greater than the observed insured loss the value of 1 is associated with the corresponding observed loss. This means that the estimated insured loss is either greater than two times the observed loss or less than 2 times the observed loss. The value of 0 is associated otherwise this means that the absolute value of the difference belongs to the range of acceptability set in step 1.

4. The values of ones and zeros are summed up and divided by the number of estimates. The obtained value is multiplied by 100 to have the value in terms of percentage.

5. The following classification can then be done:

   - If the value calculated in step 4 is less than 25%, the insured losses estimated are considered good results.
   - If the value is between 25% and 50%, the results obtained are considered average.
   - If the value is more than 50% the estimates are considered bad. Indeed, this translates that on average one out of two insured loss estimated deviates completely from the expected results in regards to the range of acceptability determined in step 1.

This example can be extended to rank the insured loss obtained from different catastrophe models in terms of accuracy and reliability of the results such as to choose the

adequate model for the given insurance/reinsurance portfolio and risk factors.

It is important to note that this process of classifying, ranking and comparing is just a guide to reduce the level of subjectivity in the decision of the catastrophe model to be used for a given insurance/reinsurance portfolio. The final decision falls on the actuary, risk manager. Thus, the importance of understanding and knowing the different steps and components of the catastrophe models used.

# Conclusion

## Summary

The aim of this study is to calculate the accuracy of the TSR model in-land for storms that made landfall in the US between 2015-2019 and give a general understanding of wind hazard. This process led to an openness and transparency of the mechanism of the TSR wind model, which is in line with the recommendation of the General Insurance Research Organising Committee (Fulcher 2006).

Before tackling the issue of accuracy of the TSR parametric wind model different parametric wind models were presented: NWS-23, Holland 1980 and Willoughby and others 2006. Then the weather station data obtained from the National Hurricane Centre (NHC) reports for storms that made landfall in the US between 2015-2019 are used to verify the accuracy of the TSR model. This weather station data was homogenised such that 60% of the weather station data was dropped. Following on, outlier techniques were reviewed for large error detection and the standard deviation threshold method was chosen to label large errors from the gust error and wind error data. Thus 4% of the gust error data was identified as outliers as well 6% of the wind error data was flagged as large errors.

An error analysis was then performed to identify the source of these large errors. After which no large error was deleted. Finally, different accuracy measures were used to evaluate the accuracy of the TSR model. For the Absolute Percentage Error (APE) the accuracy of the TSR model is given in terms of mean $\pm$ standard deviation such that the accuracy of the model for gust is 19.6% $\pm$ 14.6% this was compared to the accuracy of the TSR model in terms of gust for storms that made landfall in the US between 2004 and 2008 17.8% $\pm$ 16%. This shows that over time the accuracy of the model for gust

increased in terms of the variance but decrease in terms of the bias. However, the accuracy of the TSR model in terms of the 1 min sustained wind in this study out performed that from the study done for storms that made landfall in the US between 2004 and 2008: 17.6% ±14.4% vs 22.3% ± 20.7. Furthermore, the Mean Absolute Error (MAE) of the TSR model in terms of the 1-min sustained wind for storms that made landfall in the US between 2015 and 2019 was calculated 6.61; this value was then compared to the value of 6.65 the MAE obtained to measure the accuracy of the wind model used by the NHC for the 1-min sustained wind for the same storms.

Overall, these results show that despite the short comings of the TSR model it can still be used to mitigate the risk due to wind hazard when adequate measures are taken. Indeed, a solid understanding of the model and its limits allows for a better interpretation and integration of its outputs in the risk management process.

## To go further

To improve the precision of the accuracy of the TSR model several steps can be taken.

- In the data chapter any weather station where the anemometer had height different from 10m or with an averaging period greater than 5-min were dropped from the database, this a drastic solution to the problem. It is possible to convert these weather station data to be equivalent to the value from an anemometer of height 10m and averaging period to 1-min.
- The link between gust error and wind error needs to be investigated to find what percentage of error from the 1-min sustained wind is transferred to the error of the 3-sec peak gust wind. Recall in chapter 1, it was explained that the gust values from the TSR model are calculated from the value of the TSR model for the 1-min sustained wind. The propagation of error between wind and gust is thus of interest and can be examined by studying the gust factor.

The accuracy of a model is highly dependent on the scope of the study in which it is done; this can always be improved to render more precise and reliable results.

In regards to gaining more understanding on catastrophe model for wind hazard and the impact on the insured loss the site https://www.sbafla.com/method/Home.aspx is a good resource as it is made up of documentations submitted to the Florida Commission on Hurricane Loss Projection Methodology that verify the adequacy of hurricane models. Every odd year catastrophe model vendors (AIR, RMS and others) submit a complete review of their wind model. The documents are filled with information on the accuracy, reliability and construction of their models and the various test they conducted to prove that their models produce outputs (including insured losses) that are statistically sound and conform to actuarial standards and the reglementation.

# Bibliography


Adam, Podlaha Steve Bowen Siamak Daneshvaran Claire Darbinyan Michal Lörinc (2017). *Hurricane Matthew Event Recap Report*. Aon Benfield, Analytics, Impact Forecasting. URL: http://thoughtleadership.aonbenfield.com/Documents/20170424-ab-if-hurricane-matthew-recap.pdf.

AIR-WorldWide, Verisk (2017). *Modeling Extreme Event Risk*. AIR is a Verisk Analytics business.

Ankerst, Mihael Breunig Markus Kriegel Hans-Peter Jörg Sander (June 1999). "OPTICS: Ordering points to identify the clustering structure". In: *SIGMOD record* 28.2, pp. 49–60.

Barnett V., Lewis T. (1994). *Outliers in Statistical Data*.

Berendsen, Herman J. C. (2011). *A Student's Guide to Data and Error Analysis*. Student's Guides. Cambridge University Press, pp. 18–26. DOI: 10.1017/CBO9780511921247.004.

Berg, Robbie (2015). "TROPICAL STORM BILL (AL022015)". In: *NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT*.

— (2017). "HURRICANE HERMINE (AL092016)". In: *NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT*.

— (2018a). "TROPICAL STORM ALBERTO (AL012018)". In: *NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT*.

— (2018b). "TROPICAL STORM CINDY (AL032017)". In: *NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT*.

— (2019). "TROPICAL STORM GORDON (AL072018)". In: *NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT*.

Beven, John Robbie Berg (2018). "HURRICANE NATE (AL162017)". In: *NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT*.

Bhattacharya, Adam Podlaha Steve Bowen Michal Lörinc Anwesha (May 2018a). "Global Catastrophe Recap". In: *Aon Benfield*. URL: http://thoughtleadership.aonbenfield.com/Documents/20180607-ab-analytics-if-may-global-recap.pdf.

— (Sept. 2018b). "Global Catastrophe Recap". In: *Aon Benfield*. URL: http://thoughtleadership.aonbenfield.com/Documents/20181009-ab-analytics-if-sept-global-recap.pdf.

Bhattacharya, Adam Podlaha Steve Bowen Michal Lörinc Anwesha (2018c). "Weather, Climate and Catastrophe Insight". In: *Aon Benfield*. URL: http://thoughtleadership.aonbenfield.com/Documents/20190122-ab-if-annual-weather-climate-report-2018.pdf.

Blake, Eric S. (2017). "TROPICAL STORM JULIA (AL112016)". In: *NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT*.

Bogg, Alexandra Lewis Andrew Wragg David (Jan. 2018). "Costliest year on record for weather disasters with USD344 billion global economic loss in 2017 - Aon catastrophe report". In: *Aon Benfield*.

Bowen, Matthew Cappucci Steve (Nov. 2020). "The record-shattering 2020 Atlantic hurricane season is over, but the scars it left remain". In: *The Washington Post*.

Brennan, Michael J. (2016). "TROPICAL STORM BONNIE(AL022016)". In: *NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT*.

Cangialosi, Richard J. Pasch Andrew S. Latto John P. (2019). "TROPICAL STORM EMILY (AL062017)". In: *NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT*.

Cappucci, Matthew (Nov. 2020). "The record-shattering 2020 Atlantic hurricane season is over, but the scars it left remain". In: *The Washington Post*.

Chen, Edwin (Mar. 2011). "A Kernel Density Approach to Outlier Detection". In:

Claire, Lörinc Podlaha Bowen (Nov. 2017). "Global Catastrophe Recap". In: *Aon Benfield*. URL: http://thoughtleadership.aonbenfield.com/Documents/20171207-ab-analytics-if-november-global-recap.pdf.

Cohn, Carolyn (May 2020). "Lloyd's of London expects COVID-19 claims to match 9/11". In: URL: https://uk.reuters.com/article/us-health-coronavirus-lloyd-s-of-london/lloyds-of-london-expects-covid-19-claims-to-match-9-11-idUKKBN22Q0MT.

Coleman, Denise (Dec. 2015). "Box Plot with Minitab". In:

Darbinyan, Adam Podlaha Steve Bowen Claire (2015). "Global Catastrophe Recap". In: *Aon Benfield*. URL: http://thoughtleadership.aonbenfield.com/Documents/20151208-if-november-global-recap.pdf.

— (2016). "Global Catastrophe Recap". In: *Aon Benfield*. URL: http://thoughtleadership.aonbenfield.com/Documents/20160908-ab-analytics-if-august-global-recap.pdf.

Dawson, Robert (Aug. 2012). "How Significant is a Boxplot Outlier?" In: *Journal of Statistics Education* 19.2.

Desflots, Dr. Ioana Dima Dr. Melicie (2010). *Wind profiles in Parametric Hurricane Models*. AIR-Worldwide.

Fulcher, Phil Archer-Lock Rob Caton David Davies Tanya Fick Gillian James Hanna Kam Paul Kershaw Laura Masi Stephen Postlewhite Justin Skinner David Wong Graham (2006). "Catastrophe Modelling

Working Party". In: *GIRO:General Insurance Research Organising Committee*. URL: https://www.actuaries.org.uk/system/files/documents/pdf/catreport2006.pdf.

Ginger, John Bruce Harper Jeffrey Kepert (2008). *WIND SPEED TIME AVERAGING CONVERSIONS FOR TROPICAL CYCLONE CONDITIONS*. AMS 28th Conference on Hurricanes and Tropical Meteorology.

— (2010). *GUIDELINES FOR CONVERTING BETWEEN VARIOUS WIND AVERAGING PERIODS IN TROPICAL CYCLONE CONDITIONS*. World Meteorological Organization.

Haby, Jeff (2020). *Wind speed increasing with height*. URL: https://www.theweatherprediction.com/habyhints3/749/.

Hagen, Lixion A. Avila Stacy R. Stewart Robbie Berg Andrew B. (2020). "HURRICANE DORIAN (AL052019)". In: *NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT*.

Hampel, F R (1974). "The influence curve and its role in robust estimation". In: *Journal of the American Statistical Association* 69.346, pp. 383–393.

Hans-Peter, Erich Schubert Arthur Zimek (2014). *Generalized Outlier Detection with Flexible Kernel Density Estimates*.

Hase, Ifan Hughes Thomas (2010). *Measurements and Their Uncertainties : A Practical Guide to Modern Error Analysis*. Oxford University Press.

Hawkins, D. (1980). *Identification of Outliers*. Chapman and Hall.

Holland, Dr. Greg (1980). "An analytic model of the wind and pressure profiles in hurricanes". In: pp. 1212–1218.

Hsu, D. Y. (1999). *Spatial Error Analysis: A Unified Application-Oriented Treatment*, pp. 23–33.

J.Beven, Robbie Berg Andrew Hagen (2019). "HURRICANE MICHAEL (AL142018)". In: *NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT*.

Johansen, Martin Berg Christensen Peter Astrup (2012). "A simple transformation independent method for outlier definition". In: *Clinical chemistry and laboratory medicine* 56.9, pp. 1524–1532.

John P. Cangialosi Andrew S. Latto, Robbie Berg (2018). "HURRICANE IRMA (AL112017)". In: *NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT*.

Latchman, Shane (2009). "Modelling catastrophes". In: *+plus magazine*. URL: https://plus.maths.org/content/comment/reply/2373.

Lea, Dr Adam (2020). *Lecture: Tropical Cyclone Wind, Storm Surge and Rainfall Modelling*. MSc IN GEOPHYSICAL HAZARDS.

Lexico (2020). "Meaning of storm in English". In: *Oxford dictionary*. URL: https://www.lexico.com/definition/storm.

LI, XINRONG (2013). "Catastrophe Model Suitability Analysis: Quantitative Scoring". In:

Licata, C Leys C Ley O Klein P Bernard L (2013). "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median". In: *J Exp Soc Psychol* 49.4, pp. 764–766.

Macfee, Garry M. Pearson Allan W. (2006). "Development and Testing of Tropical Cyclone Parametric Wind Models Tailored for Midlatitude Application—Preliminary Results". In: *JOURNAL OF APPLIED METEOROLOGY AND CLIMATOLOGY* 45, pp. 1244–1260.

Manikandan, S. (2011). "Measures of central tendency: Median and mode". In: *J Pharmacol Pharmacother* 2.3, pp. 214–215. DOI: 10.4103/0976-500X.83300.

Meij, T V Pollet L van der (2017). "To remove or not to remove: the impact of outlier handling on significance testing in testosterone data". In: *Adaptive Human Behavior and Physiology* 3.1, pp. 43–60.

Mohandes, S Rehman S.M. Rahman M (2011). *Estimation of Wind Speed Profile Using Adaptive Neuro-fuzzy Inference System (ANFIS)*. Applied Energy, pp. 4024–4032.

Nelson, J P Simmons U Simonsohn L D (2011). "False positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant". In: *Psychological Science* 22.11, pp. 1359–1366.

Ng, Edwin M. Knorr Raymond T. (1997). "A unified notion of outliers: Properties and computation". In: *Proc. KDD*.

NHC, Alex Maynard (2018). *The Saffir-Simpson Hurricane Wind Scale*. Alabamawx weather blog. URL: https://www.alabamawx.com/?p=174787.

NOAA (1979). "Meteorological criteria for standard project hurricane and probable maximum hurricane windfields, Gulf and East coasts of the United States". In: *Technical Report, Coastal Services Center*.

— (2020). "Hurricane Database (HURDAT 2)". In: *National Oceanic and Atmospheric Administration*. URL: https://www.aoml.noaa.gov/hrd/hurdat/Data_Storm.html.

Penny, Richard J. Pasch Andrew B. (2017). "TROPICAL STORM COLIN (AL032016)". In: *NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT*.

Phadke, Kwok Fai Cheung Samuel H. Houston Christopher D. Martino (2003). "Modeling of tropical cyclone winds and waves for emergency management". In: *Ocean Engineering* 30, pp. 553–578.

Podlaha, Bowen Darbinyan Lörinc (June 2017a). "Global Catastrophe Recap". In: *Aon Benfield*. URL: http://thoughtleadership.aonbenfield.com/Documents/20170706-ab-analytics-if-june-global-recap.pdf.

— (July 2017b). "Global Catastrophe Recap". In: *Aon Benfield*. URL: http://thoughtleadership.aonbenfield.com/Documents/20170808-ab-analytics-if-july-global-recap.pdf.

R.Berg, J.Cangialosi A.Hagen (2019). "HURRICANE BARRY (AL022019)". In: *NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT*.

Regler, P.V. Bork H. Grote D. Notz M. (1993). *Data Analysis Techniques in High Energy Physics Experiments*. Cambridge University Press.

S.Bowen, M.Lörinc A.Podlaha (2019). "Weather, Climate and Catastrophe Insight". In: *Aon Benfield*. URL: http://thoughtleadership.aon.com/Documents/20200122-if-natcat2020.pdf.

Sander, M. M. Breunig H.-P. Kriegel R. Ng J. (2000). "LOF: Identifying density-based local outliers". In: *Proc. SIGMOD*, pp. 93–104.

Stacy, Robbie Berg Stewart (2019). "HURRICANE FLORENCE (AL062018)". In: *NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT*.

Stewart, Stacy R. (2015). "TROPICAL STORM ANA (AL012015)". In: *NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT*.

— (2017). *HURRICANE MATTHEW*. NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT.

Takezawa, Kunio (2012). "Guidebook to R Graphics Using Microsoft Windows". In:

Taylor, John R. (1982). *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. University Science Books.

TSR Business (2020). *Tropical Storm Risk Business*. University College London. URL: http://tropicalstormrisk.com/business.

Tucakov, Knorr Edwin Raymond T. Ng Vladimir (2000). "Distance-based outliers: algorithms and applications". In: *The VLDB Journal* 8, pp. 237–253.

Tukey, DC Hoaglin B Iglewicz JW (1986). *Performance of some resistant rules for outlier labeling*.

Webb, H E Hawkes J S (1962). *Geochemistry in mineral exploration*. New York: Harper & Row.

Wieringa, J. (1973). "Gust factors over open water and built-up country". In: *Boundary-Layer Meteorology* 3.4, pp. 424–441.

Willoughby, H. E. R. W. R. Darling M. E. (2006). "Parametric representation of the primary Hurricane vortex. Part II : A new family of sectionnally continuous profiles". In: pp. 1102–1120.

Yang, Rahardja Susanto Jiawei (Dec. 2019). "Outlier detection: how to threshold outlier scores?" In: *ResearchGate*.

Yeung, Robert Ireland Chuen Albert (2020). *A Dictionary of Dentistry*. 2nd ed. Oxford University Press.

Zelinsky, Eric S. Blake David A. (2018). "HURRICANE HARVEY (AL092017)". In: *NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT*.

# Appendices

# Appendix A

# Table of the Saffir-Simpson Hurricane Wind Scale

| Category | Sustained Wind (miles per hour) | Expected Damage Due To Winds |
|---|---|---|
| 1 | 74-95 mph | • Homes may have damage to roofs, shingles, vinyl siding and gutters.<br>• Small unstable trees may be uprooted and branches snapped.<br>• Downed power lines could result in power outages lasting a few to several days. |
| 2 | 96-110 mph | • Homes could endure major roof and siding damage.<br>• Many weak trees will be snapped or uprooted, possibly blocking roadways.<br>• Almost complete power loss can be expected with outages lasting from several days to weeks. |
| 3 | 111-129 mph | • Homes may undergo major damage or removal of roof and decking.<br>• Many trees will be snapped or uprooted, blocking numerous roads.<br>• Electricity and water will be unavailable for several days to weeks after the storm passes. |
| 4 | 130-156 mph | • Well-structured homes can endure severe damages and total loss of roof structure and some exterior walls.<br>• Most trees will be snapped and uproot along with power lines and poles downed.<br>• Fallen trees and downed electric poles will block roads, while power outages will last weeks to possibly months.<br>• Most of the area will be blocked, isolated and uninhabitable for up to weeks or months. |
| 5 | 157 mph or higher | • A high percentage of framed homes will be destroyed, with total roof failure and wall collapse.<br>• Many fallen trees and electric poles will block roads as a high percentage of power outages will last for weeks to months.<br>• Most of the area will be uninhabitable and isolated for up to a month or more. |

Figure A.1: A detailed explanation of the Saffir-Simpson Hurricane Wind Scale (SSHWS), Source: NHC (2018)

# Appendix B

# Detailed legend of figure of all storms figure 2.2

Legend figure 2.2

1. ALBERTO: NASA TERRA/MODIS VISIBLE IMAGE OF ALBERTO AT 1621 UTC 27 MAY 2018 WHILE IT WAS A SUBTROPICAL STORM (Berg 2018a).

2. ANA: NOAA GOES-15 VISIBLE SATELLITE IMAGE SHOWING ANA AT ITS PEAK INTENSITY ON 9 MAY 2015 AS IT MOVED TOWARD THE COAST OF THE SOUTHEASTERN UNITED STATES (Stewart 2015).

3. BARRY: GOES-16 VISIBLE SATELLITE IMAGE AT 1520 UTC 13 JULY OF HURRICANE BARRY AROUND THE TIME IT MADE LANDFALL IN LOUISIANA (R.Berg 2019).

4. BILL: GOES-EAST VISIBLE SATELLITE IMAGE OF TROPICAL STORM BILL AT 1645 UTC 16 JUNE 2015 AT THE TIME OF ITS LANDFALL ALONG THE TEXAS COAST ON MATAGORDA ISLAND (Berg 2015).

5. BONNIE: GOES-13 VISIBLE IMAGE OF BONNIE AT PEAK INTENSITY (1900 UTC 28 MAY 2016). IMAGE COURTESY OF THE U.S. NAVAL RESEARCH LABORATORY TC WEBPAGE (Brennan 2016).

6. CINDY: GOES-16 GEOCOLOR VISIBLE SATELLITE IMAGE OF TROPICAL STORM CINDYAT 2115UTC 20JUNE2017. IMAGE COURTESY OF CIRA AND RAMMB (Berg 2018b).

7. COLIN: MODIS SATELLITE IMAGE OF TROPICAL STORM COLIN OVER THE SOUTHERN GULF OF MEXICO AT 1845 UTC 5 JUNE 2016. IMAGE COURTESY OF NASA (Penny 2017).

8. DORIAN: VIIRS-S-NPP SATELLITE IMAGE OF HURRICANE DORIAN AT 0703 UTC 2 SEPTEMBER 2019 SHOWING THE WELL-DEFINED EYE OF THE CATEGORY 5 HURRICANE OVER EASTERN GRAND BAHAMA ISLAND (Hagen 2020).

9. EMILY: NASA TERRA SATELLITE IMAGE OF TROPICAL STORM EMILY AT 1555 UTC 31 JULY, SHORTLY AFTER LANDFALL (Cangialosi 2019).

10. FLORENCE: PANORAMIC VIEW OF FLORENCE WHEN THE HURRICANE WAS AT CATEGORY 4 STRENGTH ON 10 SEPTEMBER 2018. IMAGE FROM INTERNATIONAL SPACE STATION (ISS) COURTESY OF ESA/NASA ISS ASTRONAUT ALEXANDER GERST (Stacy 2019).

11. GORDON: GOES-16 GEOCOLOR IMAGE OF TROPICAL STORM GORDON AT 2200 UTC 5 SEPTEMBER 2018. IMAGE COURTESY OF THE COOPERATIVE INSTITUTE FOR RESEARCH IN THE ATMOSPHERE (CIRA) (Berg 2019).

12. HARVEY: NASA TERRA MODIS INFRARED IMAGE OF HARVEY AT 0419 UTC 26 AUGUST 2017 JUST AFTER LANDFALL AS A CATEGORY 4 HURRICANE IN TEXAS. IMAGE COURTESY OF UW/CIMSS (Zelinsky 2018).

13. HERMINE: NASA TERRA MODIS IMAGE OF HERMINE AT 1625 UTC 1 SEPTEMBER 2016 OVER THE NORTHEASTERN GULF OF MEXICO, JUST BEFORE IT BECAME A HURRICANE (Berg 2017).

14. IRMA: VIIRS SATELLITE IMAGE OF HURRICANE IRMA WHEN IT WAS AT ITS PEAK INTENSITY AND MADE LANDFALL ON BARBUDA AT 0535 UTC 6 SEPTEMBER (John P. Cangialosi 2018).

15. JULIA: NASA-MODIS VISIBLE IMAGE OF JULIA NEAR PEAK INTENSITY OVER FLORIDA AT 1825 UTC 13 SEPTEMBER 2016 (Blake 2017).

16. MICHAEL: GOES-16 PSEUDO-NATURAL COLOR IMAGE OF HURRICANE MICHAEL AT 1730 UTC 10 OCTOBER 2018. IMAGE COURTESY OF NOAA/NESDIS (J.Beven 2019).

17. MATTHEW: MIMIC MICROWAVE SATELLITE IMAGE OF MATTHEW AS THE EYE OF THE POWERFUL HURRICANE WAS MAKING LANDFALL NEAR LES ANGLAIS, HAITI, AT 1100 UTC 4 OCTOBER 2016. IMAGE COURTESY OF UW-CIMSS WISCONSIN (Stewart 2017).

18. NATE: COMBINED NIGHTTIME VISIBLE/INFRARED IMAGERY OF NATE AT 0712 UTC 8 OCTOBER FROM THE VIIRS INSTRUMENT ON THE SUOMI-NPP SATELLITE. IMAGE COURTESY OF NRL MONTEREY (Beven 2018).

# Appendix C

# Other outlier detection methods: IQR, MAD and LOF

## C.1 MAD

The Mean Absolute Deviation (MAD) was introduced by Hampel (1974) to account for the sensitive of the mean to outliers, $2^{nde}$ limit of the standard deviation threshold exposed in subsection 3.1.1.1 . The MAD gives an interval lower bond $T_{inf}$ and an upper bound $T_{sup}$ where any value outside of the interval is an outlier:

$$\begin{cases} T_{inf} = Q_{50\%} - a \times MAD \\ \\ T_{sup} = Q_{50\%} + a \times MAD \end{cases} \tag{C.1}$$

,where $MAD = b \times Q_{50\%}\,(x - Q_{50\%})$, $x$ is a data point from the dataset of interest, $Q_{50\%}$ the median of the dataset of interest, $a$ is commonly set to 3 Nelson (2011), (Yang 2019) suggest the value of 1.4826 for $b$. The median is seen to be less sensitive to outliers than the mean.

However, the median is not insensitive to outliers. The range in which a data point is not an outlier can be distorted due to the presence of outliers in the calculation of the median. This method resembles and shares the same limitation as the Interquartile Range (IQR) that will be discussed in the next subsection.

## C.2 IQR

The Tukey's schematic boxplot commonly known as boxplot is a graphical representation of several characteristics of a dataset (Yeung 2020). This tool is useful for outlier detection as it needs no prior knowledge of the underlying distribution of the dataset. For Tukey (1986) this is an exploratory data analysis technique that can be described as a "resistant rule" in labeling potential outliers from a univariate dataset.

A boxplot is composed of (Takezawa 2012) ::

- Min value of the dataset
- Potential outliers on the left of the distribution
- Lower whisker
- First quartile: $Q_{25\%}$
- Seconde quartile: $Q_{50\%}$ (median)
- Third quartile: $Q_{75\%}$
- Upper whisker
- Potential outliers on the right of the distribution
- Max value of the dataset

The Interquartile Range (IQR) is $Q_{75\%} - Q_{25\%}$. The whisker is a range of value caracterised by the IQR.

$$\text{The lower whisker is all the values in } [Q_{25\%} - 1.5IQR, Q_{25\%}].$$
$$\text{The upper whisker is all the values in } [Q_{75\%} + 1.5IQR, Q_{75\%}].$$

In figure C.1 representing a boxplot, the lower whisker is the $\dashv$ on the left and the $\vdash$ on the right. Johansen (2012) calls the whiskers, Tukey's fences as any value outside of the "fences" is considered a potential outlier. More formally with the boxplot technique a value in a dataset is labeled as a potential outlier if it doesn't belong to the following interval :

$$[Q_{25\%} - 1.5 \times IQR, 1.5 \times IQR + Q_{75\%}] \tag{C.2}$$

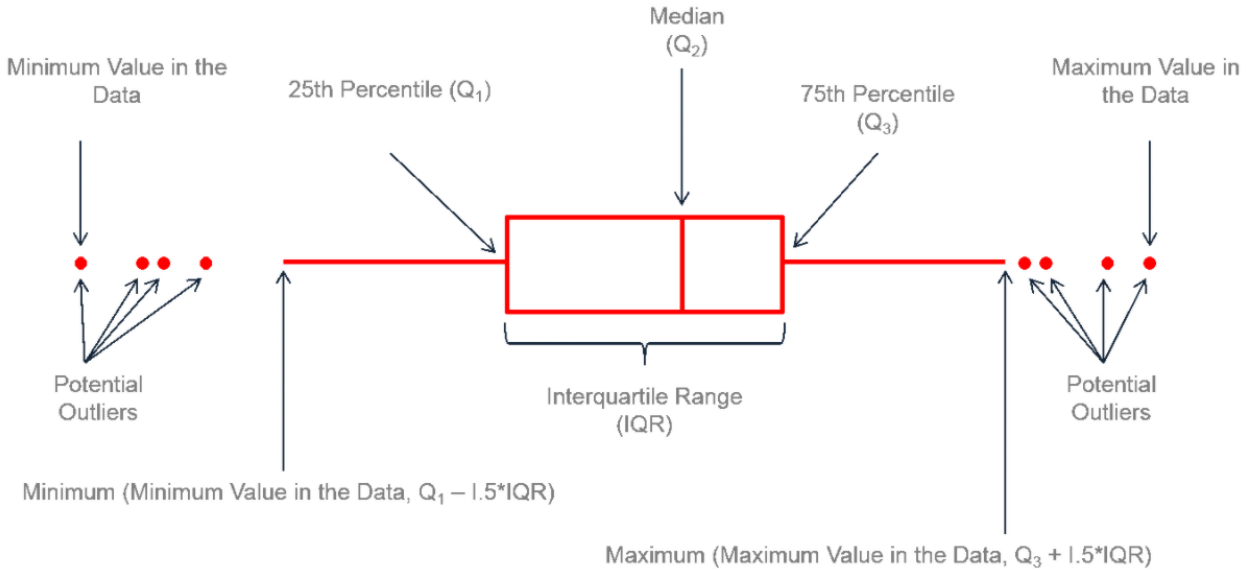



Figure C.1: Boxplot anatomy, author: Coleman (2015)

Dawson (2012), refers to this method as a useful but informal tool as it is highly sensitive to the size of the dataset. Indeed larger dataset imply a much larger IQR thus some data points that could be outliers may not be flagged as one. Johansen (2012) doesn't undermine the effectiveness of this technique but points out its limitations when used for asymmetric distributions.

The inadequacy of the Tukey's fences methodology to identify outliers in large datasets comes mainly from the inability of this technique to account for the level of dispersion in the data points in the dataset in the outlier detection process (Chen 2011). Thus, some authors Ng (1997), Sander (2000), and Tucakov (2000) advice the use of distance based algorithms in labeling potential outliers.

## C.3 Local Outlier Factor (LOF)

*This section is entirely based on the paper LOF: Identifying Density-Based Local Outliers, (Sander 2000) as such, the same definitions and annotations are used.*

A key difference the local outlier factor technique has with the three previous methodologies (see subsections 3.1.1.1, C.1 and C.2 ) is it doesn't consider an object being an

outlier as a binary property but rather gives the degree at which an object is outlying by attributing a score to the data point. For Sander (2000), a data point from a database is called an object, so this term will be used interchangeably hereinafter.

The Local Outlier Factor (LOF), is a distance based approach that assigns a factor to each data point in the database. Hans-Peter (2014) and Ng (1997) give a definition of the concept of a distance based approach for outlier detection. For a database D, an object p in the database D ($p \in D$) is labeled as an outlier if all the objects in the database D without the object p $\left(\forall o \in D \setminus \{p\}\right)$ are at a distance from p larger than a given threshold. The LOF algorithm compares the estimated density of every object o in the database $D \setminus \{p\}$ with the average density of the k-nearest neighbour of object p to attribute a score to p. The k-nearest neighbours of an object p $N_{k-distance(p)}$ refers to a subset of $D \setminus \{p\}$ with the minimum distance from p such that at least k objects belong to this subset; this notion is used in clustering algorithms, a more formally definition will be given latter on.

The LOF algorithm can be divided into two major phases (Hans-Peter 2014) :

1. Density estimation of each object
2. Attributing a local outlier factor to each object by comparing densities

When the value of the LOF is approximately 1 for an object p from a database D, that means the object p is close to its k-distance neighbours (i.e in a neighborhood of objects with similar characteristics), in the database D. The more the LOF value of the object p is greater than 1 the more the object p is identified as an outlier.

The different steps for the LOF algorithm are presented below:

1. Choice of a distance metric

$$d(p, o) = \sqrt{(p - o)^2} \tag{C.3}$$

2. k-distance neighbours

$$N_{k-distance(p)}(p) = \{q \in D \setminus \{p\} \,|d(p,q) \leq k - distance(p)\} \tag{C.4}$$

3. Reachability distance of an object p with regards to object o

$$reach - dist_k(p,o) = max\{k - distance(o), d(p,o)\} \tag{C.5}$$

This step redefines the distance that is used in in the LOF algorithm by using the distance metric introduced by Ankerst (1999) for clustering. Sander (2000) and Hans-Peter (2014) justify this choice as they advocate it smoothing the results by reducing the fluctuations of the distance between objects belonging to the same subset of k-distance neighbours of an object p. This point is illustrated in figure C.2.

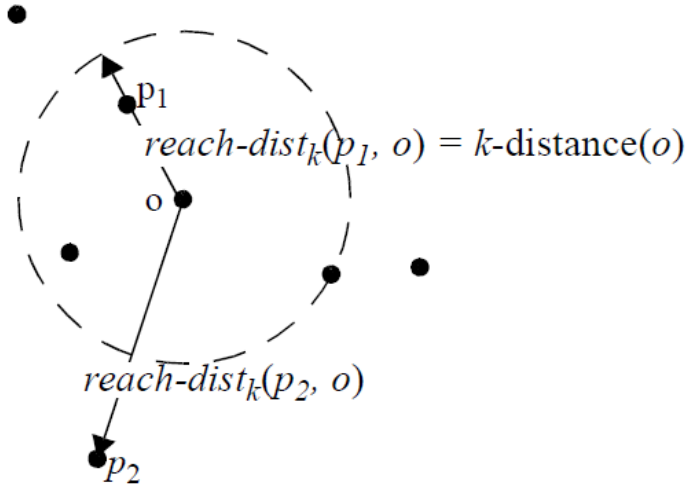



Figure C.2: Example of smoothing quality of the reachability distance function: $reach - dist(p_1, o)$, $reach - dist(p_2, o)$ for k = 3, Source : (Sander 2000). Each object : $o$, $p_1$, $p_2$ has at least 3 objects as their 3-distance neighbours. The object $o$ is in the subset of 3-distance neighbours of $p_1$ and $p_2$. The dashed circle is the 3-distance of $o$ such that at least 3 objects are within the dashed circle. Object $p_1$ belongs to the subset of 3-distance neighbours of $o$. Thus, $reach - dist_3(p_1, o) = 3 - distance(o)$ and $reach - dist_3(p_2, o) = d(p_2, o)$ as $p_2$ doesn't belong to the of 3-distance neighbours of $o$.

4. Local reachability density of an object p

A new parameter is introduce: MinPts the minimum number of objects desired in the subset of nearest neighbours of object p.

$$lrd_{MinPts}(p) = 1/\left(\frac{\sum_{o \in N_{MinPts}(p)} reach - dist_{MinPts}(p,o)}{Card\left(N_{MinPts}(p)\right)}\right) \tag{C.6}$$

, where $Card$ is the cardinal function that gives the number of elements in a set.

5. Local Outlier Factor (LOF)

$$LOF_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{Card\left(N_{MinPts}(p)\right)} \tag{C.7}$$

, where $Card$ is the cardinal function that gives the number of elements in a set.

It is important to note that the Local Outlier Factor (LOF) as defined above isn't based on a "true" density in the mathematical sense, as it doesn't have a unit integral (Hans-Peter 2014).